Reliability and Validity of Survey Scales

Louangrath, P. I. ★

About the Author
★Louangrath, P.I. is an Assistant Professor in Business Administration at Bangkok University, Bangkok, Thailand. He could be reached by email at: Lecturepedia@gmail.com

ABSTRACT
In this paper, we answered two questions: What is the reliability of a response scale in a question? What is the validity of a response scale in a question? The purpose of this paper is to present practical tools for measuring the reliability and validity of response scales used in written survey. Reliability measures consistency and validity measures precision. Our objective is to determine the reliability and validity of Likert and non-Likert scales used in research instrument. The data came from the numerical values of each type of scale. The Likert-type of scales include (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10). Non-Likert scale was (0,1,2,3). Reliability was measured by the estimated of λ under system analysis. The response space was proxied as a system to create a range between maximum and minimum values in the scale. Validity was tested by using the Fisher transformation of the estimated Z score of λ series. Empirical evidence shows that non-Likert scale (0,1,2,3) is 92% reliable while the Likert-type of scale had 90, 89, and 88% reliability. Validity test showed that non-Likert scale was 93% reliable, while the Likert-type scale had 89, 61, and 57% precision. Through Monte Carlo simulation and NK landscape method for optimization, the ability of information retention for non-Likert scale was 0.96 and 0.73, 0.75, and 0.77 for Likert scales. We standardize the scale efficacy in a 5.0 system, the non-Likert scale is 4.73 and 2.35, 2.45, and 2.41 for Likert scales.

Keywords: Likert, questionnaire, reliability, scale, survey, validity

JEL Code: C12, C13, C15, C18, C83, C93

CITATION:
Louangrath, P. (2018). "Reliability and Validity of Survey Scales." *Inter. J. Res. Methodol. Soc. Sci.*, Vol., 4, No. 1: pp. 50-62. (Jan. – Mar. 2018); ISSN: 2415-0371.

1.0 INTRODUCTION
Scientific research must meet two requisites: (i) reliability and (ii) validity. Reliability is defined as consistency in results from repeated measurements (Glasser *et al.*, 1990). Validity is defined as the precision of the result; precision is defined as minimal difference between the observed and

expected value (BS 5497-1, 1979; and BS ISO 5725-1, 1994). In social science, these two requirements points to the instrument for data collection as the unit of analysis.

In social science, data are generally collected through the use of written questionnaires or survey. In order for the research findings to achieve empirical and scientific standing, the instrument must be properly calibrated. Instrument calibration means that it must pass reliability and validity tests. These tests answer two fundamental questions: (a) Is the instrument reliable by producing consistent results with repeated measurements? and (b) Is the instrument valid by achieving the precision in its measurement?

There are four possible combinations of outcome in reliability and validity in a given instrument testing. These four possible combinations of instrument reliability and validity may be explained by the Pearson 2x2 table. Firstly, an instrument may be reliable, but not precise. In this scenario, the instrument may produce achieve homogeneity in results when subjected to repeated measurements, but failed to in achieving precision (AD). The failure in precision may be evidence by significant difference between the observed value and the expected value. In a second scenario, the instrument may fail to produce homogenous results and also fails in precision (BD). In this case, the instrument is a complete failure. A third scenario may come from a possible outcome where the instrument does produce homogenous results, but passes precision testing (CB). This is possible if the respondents consistently misunderstand the question and consistently give wrong answers. This type of instrument is also faulty. A four scenario, where the instrument produces homogenous results and precise measuring, is an ideal instrument (AC).

Table 1: Possible outcome of reliability and validity test of an instrument

	YES	NO
Reliability	*A*	*B*
Validity	*C*	*D*

The test statistic for an ideal instrument (AC) may be accomplished by:

$$\chi^2_{df=1,95\%} = \frac{(n-1)(\mathrm{ad}-\mathrm{bc})^2}{(a+b)(a+c)(b+d)(c+d)} \qquad (1)$$

where a, b, c, d are the frequency of items that falls into the YES and NO categories, and n is the number of response in a question in case of per question testing and number of questions in a survey in case of survey testing. The theoretical value is $\chi^2_{df=1} = 3.80$ for 0.95 confidence interval (Kanji, 2006, p. 85; and Yates, 1934).

2.0 LITERATURE REVIEW

2.1 Test for reliability

Reliability is defined as consistency in results after repeated measurements (Taylor, 1999). This consistency is defined by the homogeneity of the results. The argument in homogeneity is that among the k samples, their variances are equal. The decision rule is given by: $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ and $H_A : \sigma_i^2 \neq \sigma_2^2$. The alternative argument is that at least one of the paired variance is not equal. The existence of at least one unequal pair is a proof that the k samples lack homogeneity. There are two tests available for testing the homogeneity of data: (i) Lavene test and (ii) Barlett's Test.

2.1.1 Lavene Test for homogeneity

The first test for homogeneity is the Lavene's test. The Lavene test is given by:

$$W = \frac{N-k}{k-1}\left(\frac{\sum_{i=1}^{k} N_i\left(\bar{Z}_{i\bullet} - \bar{Z}_{\bullet\bullet}\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(N_{ij} - \bar{Z}_i\right)^2}\right) \quad (1)$$

where Z_{ij} can be one of the following:

$Z_{ij} = \left|Y_{ij} - \bar{Y}_i\right|$ where $\bar{Y}$ is the mean of the i^{th} group,

$Z_{ij} = \left|Y_{ij} - \bar{Y}_{i_{med}}\right|$ where $\bar{Y}_{i_{med}}$ is the median of the i^{th} group, or

$Z_{ij} = \left|Y_{ij} - \bar{Y}_i'\right|$ where $\bar{Y}'$ is the 10% trim of the mean.

The decision rule for the test statistic is based of the critical value of the F table where homogeneity does not exists if $W > F_{\alpha,k-1,\mathrm{N}-1}$ at error level α and degrees of freedom $k-1$ and $N-1$.

2.1.2 Barlett Test for homogeneity

The basis for the argument under the Bartlett's test is the same as in Lavene: $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ and $H_A : \sigma_i^2 \neq \sigma_2^2$. The test statistic of the Bartleet's test is given by (Snedecor and Cochran, 1989):

$$T = \frac{(N-k)\ln s_p^2 - \left(\sum_{i=1}^{k}(N_i - 1)\ln s_i^2\right)}{1 + \left(\frac{1}{3(k-1)}\right)\left(\sum_{i=1}^{k}\frac{1}{N_i - 1}\right) - \left(\frac{1}{N-k}\right)} \quad (2)$$

where s_i^2 = variance of the group; N = total sample size; N_i = sample size of the i^{th} group; k = number of groups; and s_p^2 = pooled variance. The pooled variance is obtained through:

$$s_p^2 = \sum_{i=1}^{k}\left(\frac{(N_i - 1)s_i^2}{N-k}\right) \quad (3)$$

The argument is that there is unequal variance if $T > \chi^2_{1-\alpha,k-1}$ where $k-1$ is the degree of freedom. Another expression of the Bartlett's test is written as:

$$M = v\ln S^2 - \left(\sum_{i=1}^{k} v_i \ln S_i^2\right) \quad (4)$$

where $v = \sum_{i=1}^{k} v_i$, $v_i = n_i - 1$, $S^2 = \sum_{i=1}^{k} v_i \div v$, $S_i^2 = \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_{i\bullet}\right)^2 \div \left(n_i - 1\right)$. In this case M is distributed approximately χ^2_{k-1} with a minimum sample size of $n < 5$. However, this M value is a biased estimated which requires a correction C where:

$$C = 1 + \frac{1}{3(k+1)}\left(\sum_{i=1}^{k}\frac{1}{v_i} - \frac{1}{v}\right) \tag{5}$$

Therefore, in evaluation homogeneity of a data set, instead of using M, the corrected Bartlett's test is simply: M / C. Thus, the hypothesis testing becomes: $H_0 : \frac{M}{C} \le \chi^2_{k-1}$ and $H_A : \frac{M}{C} > \chi^2_{k-1}$ (Dixon and Massey, 1969).

It is a common practice in social science to report Cronbach's alpha as the indicator for the reliability of the survey. The Cronbach's alpha is given by:

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum v_i}{v_T}\right) \tag{6}$$

where $n =$ number of items, $v_i =$ variance of the score for the individual question, and v_T = variance of the entire score in the survey. (Cronbach, 1951). The function of the Cronbach's alpha is to verify the consistency of answers among respondents. It does not verify whether the instrument itself is reliable. The Cronbach alpha measures the consistency of responses, not the reliability of the instrument. The consistency of responses in a sample tells the similarity or closeness of response values in a group where the "group" is defined as the sample. A defective instrument may produce consistent responses among respondents and, thus, giving high Cronbach's alpha. In such a case, the reading of Cronbach's alpha is misleading. Although consistent, such responses may still be unreliable due to the defect of the instrument.

Cronbach's alpha has been misused by researchers who synomously confused internal consistency with homogeneity (Cortina, 1993; Nunally and Bernstein, 1994; Schmitt, 1996; Sitjsma, 2009; and Streiner, 2003). Internal consistency is not sufficient condition for homogeneity (Green, *et al.*, 1977). High alpha value may mean low validity due to redundancy of questions asking for similar responses at the expense of construct coverage (Boyle, 1985, 1991; and Kline, 1979). It is clear that alpha value is not indicator for reliability. As for its purported use for reliability, the standard reading has been that for basic research, the value should be $0.70 < \alpha < 0.80$ and for issues requiring high standard, the range should be $0.90 < \alpha < 0.95$. Nevertheless, these recommendations and use of Cronbach's alpha evaluated the entire survey does not assess the reliability of each question. This inadequacy was best expressed by Cronbach himself who wrote that:

> "I no longer regard the alpha formula as the most appropriate way to examine most data. Over the years, my associate and I developed the complex generalizability (G) theory." (Cronbach *et al.* (1963); Cronbach *et. al*. (1973); *see also* Brennan (2001); Shavelson and Webb (1991), which can be simplified to deal specifically with a simple two way matrix and produce coefficient alpha (Cronbach (2004), p. 403). *Cited in* N.M. Webb, R.J. Shavelson and E.H. Haertel (2006). "Reliability Coefficients and Generalizibility Theory." Handbook of Statistics, Vol. 26, p. 2. ISSN 0169-7161.

The testing of the scale used in the survey is a different issue than what Cronbach's alpha intended to serve. Cronbach's alpha tests the consistency of responses which are external to the instrument. The reliability or consistency of responses is not the same as the reliability of the instrument. Cronbach's alpha is not a tool for instrument calibration. This paper introduces the tool for testing the reliability and validity of the instrument by examining the response space in the individual survey question.

Reliability is defined as consistency in results. Validity is the precision of the measurement. We argue that if the instrument is reliable, the response would also be reliable even if not consistent. Respondents may give different responses to the same question and give rise to low Cronbach's alpha. Such scenario does not vitiate the reliability of the instrument. In case of validity testing, the precision of the instrument becomes the unit of analysis. Within a response space, there is an expected value that could be used as a threshold value to gauge the scale's precision. A survey scale whose observed value differs significantly from the expected value is an imprecise instrument. Such an instrument fails validity test. This paper presents two research questions: (i) what is the reliability of the survey scale? and (ii) what is the validity of the survey scale?

2.2 Test for validity

As a test for precision, the test for validity requires a threshold value against which the observed value from the survey scale may be assessed. We depend on the normal distribution curve as the basis for evaluating validity. Using the percentage probability, the mid-point of the percentage probability of 50% or a critical value of $Z = 0$ is used as the threshold value with $\pm 5\%$ error. The test statistic proposed for validity test is given by:

$$\hat{V} = \frac{F(Z) - 0.50}{\sqrt{\frac{n}{12}}} \tag{7}$$

where $F(Z)$ is the percentage probability of the observed value of the scale, and n is the number of choices contained in the scale. For instance, a scale in a form of (0,1,2,3) has $n = 4$ and (1,2,3,4,5) has $n = 5$. Formula (7) is based on the optimization equation in NK landscape simulation.

3.0 DATA CONFINED TO ELEMENTS OF THE SCALES

The data used in this paper comes directly from the content of each scale. No outside data is necessary. There are four scales selected for the study. These scales are categorized into two types: Likert and non-Likert. Likert scales include (1,2,3,4,5) (Likert, 1932), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10). Non-Likert scale is (0,1,2,3).

Despite disagreement in the literature over the question of whther Likert scale is quantitative or qualitative (Jamieson, 2004; and Norman, 2010), we treat Likert and non-Likert as quantitative data. Likert scales do not contain zero. Thus, they could only be subjected to continuous distribution testing (Abramowitz and Stegun, 1972). Non-Likert scale, on the other hand, contains zero.

Discrete and continuous distribution may be used for testing non-Likert scale or scale containing zero value. The ability of the data set to allow the type of probability analysis is not a small issue for purposes of hypothesis testing. A data set that allows both discrete and continuous distribution testing affords researchers the flexibility and varied tools for hypothesis testing. The information obtained from such data set is more extensive compared to the inflexible Likert data set allowing only continuous distribution testing. The distinguishing characteristics of the Likert and non-Likert scales come from two factors: (i) the presence of zero in non-Likert scale, and (ii) the number of choices or the size of the subspace in each scale. These two characteristics are relevant in evaluating the efficacy of the scale.

Table 1: Common scales used in survey questionnaires

Scale Type	Scale Description	Data Distribution	Data Flexibility
Type 1: Non-Likert	(0,1,2,3)	Discrete & continuous	Yes
Type 2: Likert	(1,2,3,4,5)	Continuous	No
Type 3: Likert	(1,2,3,4,5,6,7)	Continuous	No
Type 4: Likert	(1,2,3,4,5,6,7,8,9,10)	Continuous	No

4.0 METHODOLOGY

4.1 Linearize the scale by QQ plot

The scale data is subjected to QQ plotting in order to obtain a linear equation: $Y = a + bX$. The linear equation will be used to obtain the expected value for each element in the cale. The QQ plot starts with calculating the time function:

$$F(t) = \frac{i - 0.30}{n + 0.40} \tag{8}$$

where i = item of sequential listing of the scale elements, and n = number of items in the scale. For example, (0,1,2,3) has $n = 4$, (1,2,3,4,5) has $n = 5$, (1,2,3,4,5,6,7) has $n = 7$, and (1,2,3,4,5,6,7,8,9,10) has $n = 10$. Since these items occur in sequence, the time function $F(t)$ is used as a starting point to obtain the predictive model for the scale. Dependent (Y) and independent (X) variables in the QQ plot, to create the linear equation, are obtained by:

$$X_{qq} = \ln\left(\ln\left(\frac{1}{1 - F(t)}\right)\right) \tag{9}$$

$$Y_{qq} = \ln(X_{scale}) \tag{10}$$

In order to construct a linear equation in a form of $Y = a + bX$, the following basic statements are required:

$$\begin{aligned} I &= n\sum XY - \sum X \sum Y \\ II &= n\sum X^2 - \left(\sum X\right)^2 \end{aligned} \tag{11}$$

The slope of the line is obtained by $b = I \div II$ and the intercept is obtained by $a = \overline{Y} - b\overline{X}$. With known linear equation, the expected value for each element of the scale subspace is estimated to be $\hat{Y}_j$.

With known expected value series of $\hat{Y}_j$, the CDF and PDF of each element of the scale is determined. The CDF is given by the Z score equation:

$$Z = \frac{\hat{Y}_j - \dot{Y}}{S_{\hat{y}}} \tag{12}$$

where $\hat{Y}_j$ = each expected value in the *j* group or response subspace, $\dot{Y}$ = mean value of the *j* group, and $S_{\hat{y}_j}$ = standard deviation of the *j* group. The Z score represents the critical value for which the percentage probability may be read from the Z table. The corresponding percentage probability from the Z table is called cumulative distribution probability (CDF or $\phi(z)$). This series of percentage probability is used to obtained *lambda* or the reliability of the scale.

4.2 Reliability testing of scales used in survey

The reliability of the scale is obtained through the calculation of lambda. In order to obtained lambda, it is necessary to know the value of CDF and PDF or probability distribution function. The PDF may be obtained by:

$$PDF = \frac{\Phi(z_2) - \Phi(z_1)}{\hat{y}_2 - \hat{y}_1} \tag{13}$$

Now, lambda may be determined by:

$$\lambda = \frac{CDF}{1 - PDF} \tag{14}$$

The value of λ will come in series of *k* elements in the scale. For instance, if the scale is (0,1,2,3), *k* element is equal to $k = 4$. The reliability of the scale is equated to the stability of the λ which is estimated to be: $\bar{\lambda} + s_\lambda$ or the mean of λ plus its standard deviation.

While $\bar{\lambda} + s_\lambda$ represents the reliability of the scale with *k* elements, the value of R may be converted to critical value in the Z table in order to obtain the level of confidence in the reliability of $\bar{\lambda} + s_\lambda$. Thus, with known R in the range of $0 < R < 1$, trace the value in the Z table for the corresponding critical value. The critical value of Z is subjected to Fisher transformation in order to calculate the upper and lower limits of the error for the estimation. The Fisher transform is obtained through:

$$Z = 0.50 \ln\left(\frac{1+R}{1-R}\right) \tag{15}$$

This transformed Z is an estimate; thus, it must carry a range of error. The upper and lower error limits of the transformed Z value is obtained through:

$$\xi_u = Z + \hat{Z}\sqrt{\frac{1}{n+3}} \tag{16}$$

$$\xi_l = Z - \hat{Z}\sqrt{\frac{1}{n+3}} \tag{17}$$

where Z = observed value and $\hat{Z}$ = threshold value at a specified percentage confidence interval. The range of the reliability R of the scale becomes $R + \xi_u$ for the upper value and $R - \xi_u$ for the lower value. This range will later provide the maxima and minima for Monte Carlo simulation.

4.3 Validity testing of scales used in survey

In order to determine the validity of the scale, the Fisher transformed Z and the critical Z value read from the table, corresponding to R, is used. This level of confidence of R is obtained in two steps, firstly:

$$Z^* = \frac{Z - \bar{Z}}{S} \tag{18}$$

where Z^* = critical value read from the Z table corresponding to R, $\bar{Z}$ = Fisher transformed Z in equation (9), and $S = \sqrt{1/(n-3)}$.

The value of Z^* is in a form of critical value which is compared to the theoretical value of $Z_\theta = 0.50$ or the mid-point of the distribution curve. The critical point at 0.50 is the mid-point of the distribution curve; we use this point as the threshold value for precision measurement. The difference $Z^* - Z_\theta$ represents the error or pValue of the estimated validity and $1-(Z^* - Z_\theta)$ represents the validity of the scale. If scale is precise, the value of $Z^* - Z_\theta$ must be zero or close to zero within an acceptable level of error, i.e. located "bulls eye" at the center of the distribution curve. Thus, $\Pr[Z^* - Z_\theta] \le \Phi(z)$ where $\Phi(z) \le 0.05$.

4.4 Monte Carlo simulation and NK landscape optimization

In order to increase the accuracy of the estimated results of reliability and validity tests for the survey scales, Monte Carlo simulation was used to obtain an optimum score. The optimum score was determined by NK landscape optimization method. Monte Carlo simulation began with the determination of the number of repetitions for the measurement, thus:

$$N = \left(\frac{3\sigma}{E}\right)^2 \tag{19}$$

where σ = the estimated standard deviation of the three components of Monte Carlo numbers: $x_1 = \max$, $x_1 = \min$, and $x_1 = (\max - \min)/2$, and $E = ((\max - \min)/2) \div 50$. The value series came from the *k* group or $\lambda_1, ..., \lambda_k$ in reliability and $\hat{Z}_1, ..., \hat{Z}_k$ for validity precision testing.

Since the Monte Carlo simulation provides a good estimate, the estimated values of reliability and validity are then used to find the optimum points. Optimization was accomplished by using NK landscape method:

$$Z_{opt} = \frac{F(\mathrm{x}) - 0.50}{\sqrt{\frac{n}{12}}} \tag{20}$$

where $F(x) = 1 - error$ and $error = U - 0.50$. The value for U is described in (22) and n = number of answer choice in each survey question.

The observed optimum value is compared to the theoretical value. The theoretical value is obtained by:

$$Z^* = \frac{(\mu + \sigma) - 0.50}{\sqrt{\frac{N}{12}}} \quad (21)$$

where μ = expected mean within the scale per survey question, and σ = expected standard deviation within the scale.

4.5 Prospect theory for validity level

Since each scale consists of quantitative elements in sequence: (0,1,2,3), (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10), the weight and probability of each may be calculated. The weight is determined by $1/n$ in each scale. Each response choice is accorded with equal weight. The probability is given by the CDF of each element of the scale. With known weight, probability and value of each element in the scale, the prospect of each value may be calculated. The Kahnman-Tvertsky prospect theory is used for this purpose:

$$U = \sum w_i p_i x_i \quad (22)$$

where w = weight of each element in the scale; p = percentage probability of each element; and x = observed value of the scale element.

The prospect U serves as the basis for validity testing. Recall that validity is the test for precision or level of accuracy (Metz, 1978). The threshold for precision in a distribution is located at the center of the distribution curve or $Z = 0.50$. The U value must be contrasted with the mid-point of the distribution curve in order to verify the precision level of the survey scale, i.e. $U - 0.50 = error$. This *error* is then used to obtain the confidence level $1 - error$. The confidence level is the precision level. This precision level defined validity.

Lastly, we ask another question: how much does the scale retain information? What is the probable information retention rate? Conversely, what is the probable information loss for each type of scale? The answers are obtained through the use of NK landscape optimization method (14). This last information was used to score the survey scale. Each survey scale was subjected to a 5-points scoring system of precision where 5 = highest, 4 = high, 3 = medium, 2 = low and 1 = very low. The rational of this scoring system is to convert the validity level into a 5-points nominal data for easy understanding.

5.0 FINDINGS AND DISCUSSION

The findings are presented in two parts. Part 1 is the raw estimate of the reliability (5.1) and validity (5.2) for the scales. Part 2 is the theoretical values of reliability and validity obtained through Monte Carlo simulation and NK landscape optimization method (5.4). In section 5.5, we carried the discussion from a single question testing to the entire instrument (global environment) based on the findings in per question subspace (local environment).

5.1 Non-Likert scale is more reliable

Non-Likert scale (0,1,2,3) is more reliable because its λ series is more stable. This stability indicates consistency of the system. The system is defined by elements of the response sequence in the scale. Since λ is an estimate, it is subjected to error. Thus, we use $\lambda + s$ as the indicator for reliability. Under this method, non-Likert scale (0,1,2,3) scored highest followed by (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10), respectively.

Table 2. Reliability of various scales

Item k	Scale 1 Non-Likert Type	Scale 2 Likert Type	Scale 3 Likert Type	Scale4 Likert Type

elements	(0,1,2,3)	(1,2,3,4,5)	(1,2,3,4,5,6,7)	(1,2,3,4,5,6,7,8,9,10)
1	0	1	1	1
2	1	2	2	2
3	2	3	3	3
4	3	4	4	4
5	-	5	5	5
6	-	-	6	6
7	-	-	7	7
8	-	-	-	8
9	-	-	-	9
10	-	-	-	10
R	0.92	0.90	0.89	0.88
$1-R$	0.08	0.10	0.11	0.12
pValue	0.07	0.11	0.39	0.43
Conclude	Pass	Fail	Fail	Fail

5.2 Validity of survey scale

The pValue reported in Table 2 measures the error level. The error is the indication of how far $Z*$ is located from the center of the distribution curve or $Z_\theta = 0$; this is the midpoint of the symmetrical normal curve. We found that non-Likert scale is 0.07 points further from the center while other Likert-type scales were 0.11, 0.39 and 0.43 further away from the center. Since $Z_\theta = 0$ is the threshold for precision, the closer the critical value of the scale's reliability score is to $Z_\theta = 0$, the more precise or valid is the scale. In this case, a non-Likert scale in a form of (0,1,2,3) is more precise than all three forms of Likert scales: (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10).

5.3 Simulation confirm reliability and validity tests

The reliability series consists of the reliability indicator for the four types of scale: 0.92, 0.90, 0.89, and 0.88. The validity series consists of the confidence level determined by $1-error$ in Table 4. Each series was subjected to Monte Carlo simulation in order to obtain the expected value. The purpose is to verify the significance of the Z values under (14) and (15) with the following decision rule: $H_0 : Z_{opt} \le Z*$ and $H_0 : Z_{opt} > Z*$

The results of the simulations for reliability and validity show that the observed optimized values were larger than the estimated. In both cases, the difference $\left(Z_{opt} - Z*\right)$ is equal to 0.01. For reliability test simulation, the Monte Carlo repetitions were 635.96 and 28.15 repetitions for validity test. The errors between the observed and estimated values were 0.01 in both cases. Since this error is less than 0.05, we found no significant difference between the simulated and observe values obtained through the method proposed in this paper.

Table 3: Optimum reliability and validity under simulation

Test Type	Series	Monte Carlo N	Z_{opt} Observed	$Z*$ Estimate	$\left(Z_{opt} - Z*\right)$
Reliability	λ_k	635.96	0.06	0.05	0.01
Validity	R_j	28.15	0.27	0.26	0.01

The validity test is based on the Kahnman-Tvertsky prospect U where each U is contrasted with 0.50 threshold value. The value 0.50 is used as the threshold because it represents the mid-point of the distribution curve. The difference between the observed U and 0.50 or $U - 0.50 = error$ represents the inaccuracy of the scale and $1 - error$ represents the precision level or validity. The result of this calculation is reported in Table 4.

Table 4: Prospect of information retention and information loss by scale type

Survey Scale Type	Prospect U	Error $U - 0.50$	Confidence $1 - error$	OPT Informatics	Information Loss
Type 1: $n = 4$ *	0.54	0.04	0.96	0.79	0.21
Type 2: $n = 5$	0.77	0.27	0.73	0.36	0.64
Type 3: $n = 7$	0.75	0.25	0.75	0.33	0.67
Type 4: $n = 10$	0.73	0.23	0.77	0.30	0.70

*Scale types: Type 1 = (0,1,2,3); Type 2 = (1,2,3,4,5); Type 3 = (1,2,3,4,5,6,7); and Type 4 = (1,2,3,4,5,6,7,8,9,10).

Under NK landscape optimization simulation method, scale Type 1 (0,1,2,3) score highest in the prospect of information retention and lowest in the prospect of information lost. This testing is outside of the scope of reliability and validity. Whereas reliability test optimum reliability and validity in Table 3 shows the intrinsic property of the scale, the optimum prospect simulation gives an indication for the amount of information obtained through the survey scale. This latter piece of information may be another indicator for scale quality evaluation. If reliability test tells us how much does the scale produce consistency in response and validity test provides the level of precision within the scale subspace, then the utility prospect measurement provides the amount of information obtained within the scale response subspace. The efficacy of the scale under the prospect U is further converted into a 5-points score system by:

$$SCALE_{score} = OPT\left(\ln\left(N\alpha^2\right)\right) \tag{23}$$

where $OPT = \left(F(Z) - 0.50\right) \div \sqrt{n/12}$, N = Monte Carlo repetition counts, and α = significance level. The scale score is given in a range of $0 < SCALE_{score} < 5$ The results of the calculation show that scale (0,1,2,3) has the highest scale score of 4.73 followed by (1,2,3,4,5,6,7) at 2.45, (1,2,3,4,5,6,7,8,9,10) at 2.41 and (1,2,3,4,5) at 2.35.

Table 5: Efficacy score for survey scale

Survey Scale Type	Monte Carlo N repetition	Log Monte Carlo $k = \ln\left(N\alpha^2\right)$	OPT Info. Retention	Scale score On 5-points System
Type 1: $n = 4$	158,548.76	0.79	5.98	4.73*
Type 2: $n = 5$	281,864.46	0.36	6.56	2.35
Type 3: $n = 7$	634,195.04	0.33	7.37	2.45
Type 4: $n = 10$	1,426,938.84	0.30	8.18	2.41

*The 5-points are defined as: 5 = highest; 4 = high; 3 = medium; 2 = low; and 1 = lowest.

5.4 Evaluating entire survey under Prospect Theory

Thus far, we have presented the evaluation of the scale in a survey question. With known per question reliability, validity and optimized prospect, we could estimate the same measurements for the entire survey with m number of questions. The estimate survey reliability may be given by the DeMoivre-Laplace theorem (Walker, 1985):

$$Z = \frac{X - np}{\sqrt{npq}} \tag{24}$$

where $n = m$ or number of questions in a survey, p = probability of success read from R for reliability and λ for validity, and $q = 1 - p$. The test number of questions are 10, 20, …, 100 in one survey. The answer obtained (24) answers the questions of "what is the maximum number of questions in a survey in order to maintain the level of R and λ at 0.95 confidence level?" The results of the test are presented in Table 6.

Table 6: Maximum number of questions in a survey

Scale Type	Reliability R	Max Questions without Distorting Reliability	Validity λ	Max Questions without Distorting Validity
(0,1,2,3)	0.92	20	0.96	30
(1,2,3,4,5)	0.90	20	0.73	10
(1,2,3,4,5,6,7)	0.89	40	0.75	20
(1,2,3,4,5,6,7,8,9,10)	0.88	50	0.77	30

Note that the answers provided in table 6 do not answer the question: "what is the reliability of the survey with *m* number of questions?" Instead, Table 6 answer the question of "what is the maximum number of question could a survey have in order to maintain 95% confidence levels of its reliability and validity?" The answer to this question is reserved for Part 2 of this paper. In Part 1 of the research, we limit the scope of the paper to per question assessment.

6.0 CONCLUSION

Non-Likert type of scale in a form of (0,1,2,3) has the highest level of reliability and validity. Likert-type scales: (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10), fail both reliability and validity tests. Reliability and validity are the cornerstones of scientific research. The reliability and validity testing tools presented in this paper are valuable for calibrating research instrument in social science. An unreliable instrument produces unreliable data. Unreliable data leads to faulty analysis and conclusion. Imprecise instrument could not bring the researcher close to the answer of the research question. Thus, it is important for researchers to use instruments that are reliable and valid. We reject Cronbach's alpha as inapplicable for instrument calibration. This paper provides practical tools for researchers to test the reliability and validity of the survey. Such tools are valuable contribution to research methodology in social science.

REFERENCES

Abramowitz, M. and Stegun, I. A. (Eds.) (1972). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, pp. 927 and 930.

Boyle, G. J. (1985). "Self-report measures of depression: Some psychometric considerations." *The British Journal of Clinical Psychology*, **24**(1), 45-59. doi:10.1111/j.2044-8260.1985.tb01312.x

Boyle, G. J. (1991). "Does item homogeneity indicate internal consistency or item redundancy in psychometric scales?" *Personality and Individual Differences*, **12**(3), 291-294. doi:10.1016/0191-8869(91)90115-R

BS 5497-1 (1979). "Precision of test methods. Guide for the determination of repeatability and reproducibility for a standard test method."

BS ISO 5725-1 (1994). "Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions." p.1.

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

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

Dixon, W. J. and Massey, F.J. (1969). Introduction to Statistical Analysis, McGraw-Hill, New York. https://archive.org/details/IntroductionToStatisticalAnalysis

Glasser, Mark; Mathews, Rob; Acken, John M. (June 1990). "1990 Workshop on Logic-Level Modelling for ASICS." *SIGDA Newsletter*, **20** (1).

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 37, 827-838. doi:10.1177/001316447703700403

Jamieson, Susan (2004). "Likert Scales: How to (Ab)use Them." *Medical Education*, **38**(12), pp.1217-1218.

Kanji, Gopal K. (2006). *100 Statistical Tests*; p. 85, 3rd ed., SAGE.

Kline, P. (1979). *Psychometrics and psychology*. London, United Kingdom: Academic Press.

Levene, H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin *et al.* eds., Stanford University Press, pp. 278-292.

Likert, Rensis (1932). "A Technique for the Measurement of Attitudes." *Archives of Psychology*. **140**: 1–55.

Metz, C.E. (October 1978). "Basic principles of ROC analysis." *Sem. in Nucl. Med.*, **8** (4): 283–98.

Norman, Geoff (2010). "Likert scales, levels of measurement and the "laws" of statistics". *Advances in Health Science Education*. **15**(5) pp. 625-632.

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

Nunnally, J. C., & Bernstein, I. H. (1994) Psychometric theory (3rd ed.). New York, NY: McGraw-Hill, Inc.

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

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

Snedecor, George W. and Cochran, William G. (1989), Statistical Methods, Eighth Edition, Iowa State University Press.

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

Taylor, John Robert (1999). An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books. pp. 128–129. ISBN 0-935702-75-X.

Yates, F. (1934). "Contingency table involving small numbers and the $\chi 2$ test." *Supplement to the Journal of the Royal Statistical Society*, **1**(2): 217–235. JSTOR 2983604. https://www.jstor.org/stable/2983604

Walker, Helen M (1985). "De Moivre on the law of normal probability." *In* Smith, David Eugene. A source book in mathematics. Dover. p. 78. ISBN 0-486-64690-4. https://www.york.ac.uk/depts/maths/histstat/demoivre.pdf