Mathematics and Statistics 8(1): 42-52, 2020
DOI: 10.13189/ms.2020.080105 http://www.hrpub.org

Gaussian Distribution on Validity Testing to Analyze the Acceptance Tolerance and Significance Level

Arif Rahman*, Oke Oktavianty, Ratih Ardia Sari, Wifqi Azlia, Lavestya Dina Anggreni

Department of Industrial Engineering, Universitas Brawijaya, Malang, Indonesia

Received June 27, 2019; Revised October 1, 2019; Accepted December 23, 2019

Abstract Some researches need data homogeneity. The dispersion of data causes research towards an absurd direction. The outlier makes unrealistic homogeneity. The research can reject the extreme data as outlier to estimate trimmed arithmetic mean. Because of the wide data dispersion, it will fail to identify the outliers. The study will evaluate the confidence interval and compare it with the acceptance tolerance. There are three types of invalidity of data gathering: outliers, too wide dispersion, distracted central tendency.

Keywords Data Collecting, Questionnaire, Validity Testing, Gaussian Distribution, Acceptance Tolerance, Significance Level

1. Introduction

Measurement is a key observation activity to gather numerical expression of evidence dimensions that are required in quantitative study. The measurement results depend on four main elements of measurement: the measured object, the measuring instrument, the measuring operator, and the measuring method. Sampling ensures the measured objects are representative part of the evidence population. Non-static dimension of moving objects is more difficult to be measured than static dimension of motionless objects. The attributes of measuring instruments such as units, scales, limits and technical specification (accuracy, precision, resolution and sensitivity) contribute to rigorous measurement. Thorough measurement needs conscientiousness of measuring operator. Complicated measuring methods sometimes trigger negligence in measurement.

Many studies exploit questionnaires as measuring instrument to gather information. Questionnaire responses depend on selected respondents as representative sample of population [1,2]. There are two types of errors that can occur in the selection of respondents: coverage error and sampling error [3-6] (see Figure 1). Coverage error is an error in making a list of populations in a sampling frame that misses some members of the population (undercoverage) or includes non-members of the population (overcoverage). Sampling error is randomization error in sample selection so any subsets of the population are not represented.

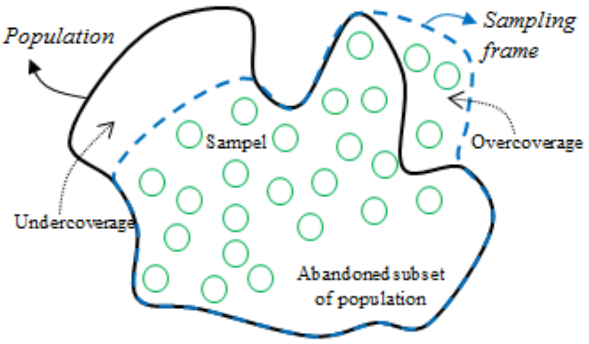


Figure 1. Respondent Selection Errors

Respondents of questionnaires play multiple roles in the measurement. Besides being involved as measuring operators, respondents occasionally act as measuring objects or measuring instruments. Respondent judgment becomes the essence of measurement using questionnaires. Human judgment affects measurement bias, especially on subjective questions. Many mistakes in measurement bias are caused by respondent personality tendencies, such as strictness, leniency and central tendency [7-11] (see Figure 2).

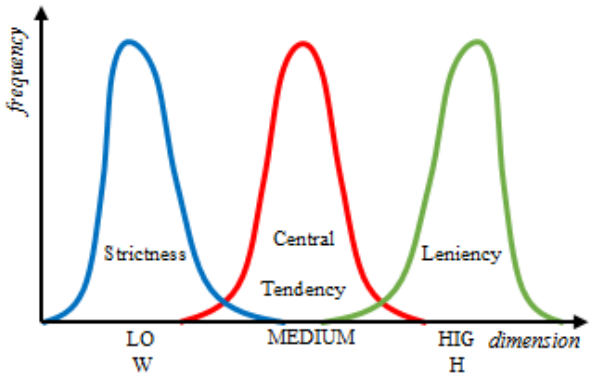


Figure 2. Measurement Bias Due To Respondent Personality Tendency

Strictness indicates severe respondent who tends to measure at lower level than actual performance. Leniency indicates mild respondent who tends to measure at higher level than actual performance. Central tendency indicates indecisive respondent who tends to measure around medium level neglecting actual performance.

Errors during information gathering process raise measurement bias on applying questionnaire as a measuring instrument. There are three types of errors as follows: random error, systematic error and illegitimate error [6,12-14]. Random error arises triggered by random circumstances or other unpredictable factors, for example, the respondent's mental state when answering the questionnaire that causes information distortion. Systematic error arises triggered by traceable factors, for example, the ambiguous questions that cause a wide variety of respondent's answers. Illegitimate error arises triggered by carelessness or impropriety of measuring operation, for example, letting the respondent answer the question even though he does not understand it.

Errors in determining the type, order and content of questions also contribute to measurement bias. The main principle of preparing good questionnaire is ensuring respondent interprets the questions matching researcher's mind. Questionnaire comprehension problems arise because of respondent's background diversity [15]. Different respondents may conceive the same question in different meaning, especially if the questions are translated from other languages that are culturally diverse [16]. Survey guidelines insist the researcher to write the short, simple and clear questions and avoid mistaken questions. Some mistaken questions contain unfamiliar words, ambiguous terms, technical jargons, obscure notion, vague meaning, confusing sentence, etc. [15,17-27].

Uncertainty in the questionnaire result was increasing when it asks questions about the attitudes, opinions or perceptions of respondents [22,28-33]. Table 1 denotes four fundamental levels of measurement scales, their properties and the common applicable type of statistical analysis [23,25]. To measure the attitudes, opinions or perceptions of respondents, many studies use dichotomous-ordinal or polytomous-ordinal [24,25,34-37] with 2 points till 11 points of scales. The attitude measurement constructs scales of questionnaire using the rating scale technique, such as Likert, Guttman, Thurstone, and semantic differential, etc. [19,24,25,27,38-42]. Some attitude scales are illustrated in Figure 3.

Table 1. Levels of Measurement and Properties [23]

No	Level	Properties	Analysis Type
1	Nominal	Distinction	Mode, Crosstabs, χ^2-test
2	Ordinal	Ordering	+ Median, Range, Mann-Whitney U-test, polychoric correlation, Spearman correlation
3	Interval	Ordering with equal interval	+ Mean, Standard deviation, *t*-test, *F*-test, Pearson correlation.
4	Ratio	Ordering with equal interval and absolute zero of origin	+ Linear regression, Non-linear regression

Favorability						
Most Unfavorable	① Dislike Extremely	② Dislike	③ Neither Like nor Dislike	④ Like	⑤ Like Extremely	Most Favorable
Satisfaction						
Most Dissatisfied	① Severe	② Poor	③ Fair	④ Good	⑤ Excellent	Most Satisfied
Agreement						
Most Disagree	① Strongly Disagree	② Disagree	③ Neutral	④ Agree	⑤ Strongly Agree	Most Agree
Importance						
Least Important	① Not at all Important	② Very Unimportant	③ Fairly Important	④ Very Important	⑤ Extremely Importance	Most Important
Attention						
Least Noticed	① Boring	② Uninteresting	③ Indifferent	④ Interesting	⑤ Exciting	Most Noticed

Figure 3. Attitude Scales

Two fundamental criteria of measurement in scientific research are reliability and validity [19, 20, 43-48]. The mentioned sources of errors lead problems of measurement reliability and validity. Reliability refers to the extent to which the repeated measurement provides consistent result. Reliability is consistency of measurement across time, across measured objects, and across measuring operators. Validity refers to the extent to which the measure reflects the variable it is intended to. Validity is conformity of measurement across measuring operators, across measuring instrument and across measuring methods.

The ideal validity test compares the measured evidence to measurement that is conducted by qualified operator using calibrated instrument and standard methods. In measurement using questionnaire, it is absurd to acquire the calibrated questions and to assign the trained respondents. Researcher can enhance the questions, but not the respondents, while respondent judgment is included as part of measuring instrument. And the diversity of respondents becomes another barrier of validity in the measurement using questionnaires.

2. Objectives

This study is concerned with validity of measurement using questionnaires. It aims to apply Gaussian distribution for validity testing. It assumes the valid value according to most respondent opinion. Due to mode, median and mean coinciding in the Gaussian distribution, it assigns each mean of questionnaire variable as reference target.

The Gaussian distribution forms a symmetrical bell-shaped curve. For single specific measuring object, the good measurement shall gather data tending to be around a specific value as center of curve. If data is distorted significantly from the center of curve, then it can be suspected as measurement bias. The study also aims to inspect three types of measurement bias: outliers, too wide dispersion, distracted central tendency.

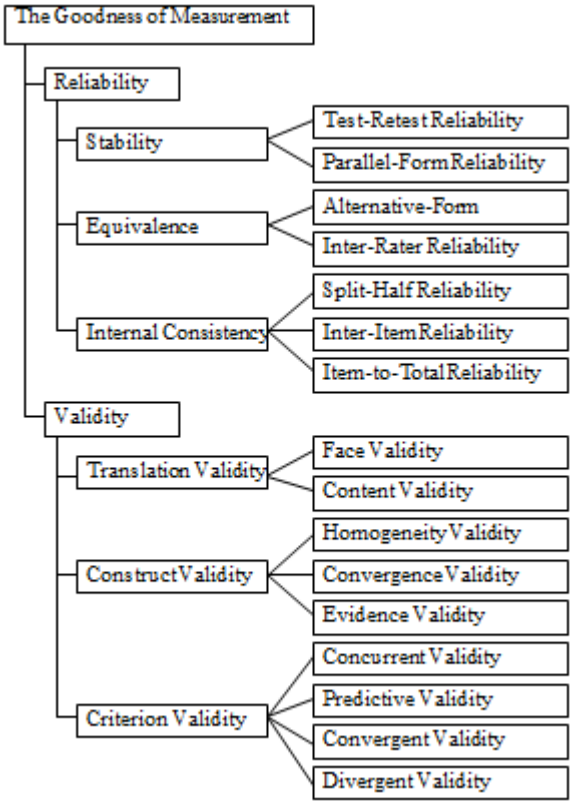


Figure 4. Classification of Reliability and Validity

3. Methods

The goodness of measurement has two essential tools: reliability and validity. Many studies [45-48] have classified several types of reliability and validity (see Fig. 4). Reliability test had three categories: Stability, Equivalence and Internal consistency. Validity test had three categories: Translation validity, Construct validity and Criterion validity.

3.1. Reliability

Reliability refers to the consistency of measurement. It represents the degree to which the repeated measurement gathers consistent data across time and across other various elements.

Stability reliability is the consistency of repeated measurement across time with the same measured objects, the same measuring instruments and the same measuring operators. Stability reliability is assessed through test-retest reliability or parallel-form reliability. For the test-retest reliability, each respondent is asked to answer the same questionnaires twice at different times (significant time apart). For the parallel-form reliability, each respondent is asked to answer two sets of similar questionnaires (different questions order) at once.

Equivalence reliability is the consistency among multiple measuring operators or among alternative forms. Equivalence reliability is assessed through alternative-form reliability or inter-rater reliability. For the alternative-form reliability, each respondent is asked to answer a pair of questionnaires (different versions of questions, but comparable). For inter-rater reliability, it compares the responses of questionnaires between respondents.

Internal consistency reliability is the consistency among two or more measuring instruments, which measure the same measured objects. Internal consistency reliability is assessed through split-half reliability, inter-item reliability or item-to-total reliability. For split-half reliability, each pair of similar questions (different versions of questions) is set at the first-half and second-half of the questionnaires or at even and odd sequence. For inter-item reliability, several questions are spread in the questionnaires, which actually measure the same variable. For item-to-total reliability, a set of formative or reflective variables are compared to their sum or average.

3.2. Validity

Validity refers to the exactness of measurement. It represents the degree to which the measurement gathers the unbiased data.

Translation validity is the exactness of measuring instruments to be adequately translated into the intended measure. Translation validity is assessed through face validity or content validity. For face validity, it uses human judgment to assess glance appearance of measurement result. For content validity, it examines content domain to ensure the question covers the intended measure.

Construct validity is the exactness of measurement resulting to induce inferences related to the relevant theories or logical hypothesis. Construct validity is assessed through homogeneity validity, convergence validity or evidence validity. For homogeneity validity, the gathered data congregate around a specific score. For convergence validity, it compares the data measured by two or more measuring instruments (sometimes requiring scale conversion). For evidence validity, it compares the gathered data to theoretical propositions.

Criterion validity is the exactness of measurement to differentiate measured objects on the referents or criterions. Criterion validity is assessed through concurrent validity, predictive validity, convergent validity or divergent validity. For concurrent validity, the scale distinguishes measured objects which are known to be different. For predictive validity, the measurement enables to interpolate or extrapolate referring to predictive criterion. For convergent validity, the gathered data of two or more variables denote highly correlated, the same as the referent. For discriminant or divergent validity, the gathered data of two or more variables denote uncorrelated, the same as the referent.

3.3. Evolution of Validity Standard

The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have developed standards for validity testing. Since 1974, they have collaborated to revise the standards which were pioneered by APA in 1952 [49-58].

Table 2. Evolution of Validity Standard [51]

Publication	Validity Classification
Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal [52]	*Categories*: predictive, status, content, congruent
Technical Recommendations for Psychological Tests and Diagnostic Techniques [53]	*Types*: construct, concurrent, predictive, content
Standards for Educational and Psychological Tests and Manuals [54]	*Types*: criterion-related, construct-related, content-related
Standards for Educational and Psychological Tests [55]	*Aspects*: criterion-related, construct-related, content-related
Standards for Educational and Psychological Testing [56]	*Categories*: criterion-related, construct-related, content-related
Standards for Educational and Psychological Testing [57]	*Sources of evidence*: content, response processes, internal structure, relations to other variables, consequences of testing
Standards for Educational and Psychological Tests [58]	*Sources of evidence*: content, response processes, internal structure, relations to other variables, consequences of testing

Table 2 shows the evolution of validity standard that was published by The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). Since 1999, they refer to types of validity evidence, rather than distinct types of validity. They emphasize that reliability and validity are functions of the interpretations of measurement results for the intended uses and not of the measurement itself.

3.4. Conceptual Thinking

Gaussian distribution or normal distribution is a very common and well-known continuous probability distribution. In 1809, Karl Friedrich Gauss developed the well-fitted formula for the distribution of measurement errors in scientific study. Gaussian distribution is also related to central limit theorem.

The Gaussian distribution with probability density function $f(x)$ and parameters, mean μ and standard deviation σ, has following properties:

- It is symmetric about its mean, median and mode which coincide at the point $x=\mu$.
- It is unimodal.
- Its density is log-concave and infinitely differentiable.
- It has smooth and symmetrical bell-shaped curve.

Because of the central limit theorem, Gaussian distribution is widely used in scientific study. Most scientific studies, which require statistical method to examine differences between means, apply Gaussian distribution. In 1924, Walter Andrew Shewhart developed the statistical quality control chart based on principles of Gaussian distribution.

This study applies the Gaussian distribution to test the validity of measurement using questionnaires. Considering a given acceptance tolerance, it verifies the measurement whether there is or not measurement bias. Using the principles of applying the Gaussian distribution to the Shewhart Process Control, it presumes acceptance tolerances like specification limits and significance levels of confidence interval like control limits.

Figure 5 shows the conceptual framework of this study. The validity testing is applying Gaussian to examine the sources of evidences of measurement bias. It requires determining an acceptance tolerance compared with evidence distribution of measurement data based on Gaussian distribution. It will check three types of measurement errors, i.e. outliers, too wide dispersion, and distracted central tendency.

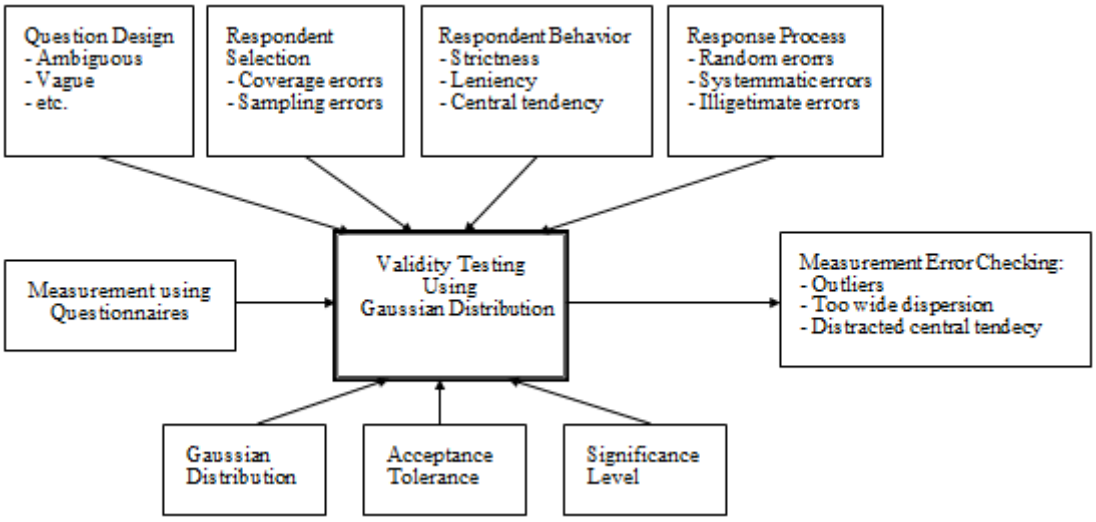


Figure 5. Conceptual Framework of Validity Testing

The variables in this study were defined by the following notations:

$x_{i,j}$: j-th respondent answer for i-th question,
μ_i : average of data for i-th question,
at_i : acceptable tolerance for i-th question,
AT_i : acceptance tolerance for i-th question,
m : the number of questions,
n : the number of respondents,
α : acceptable significance level,
sig_i : probability of i-th rejection region,

where i represents the question number of m questions in the questionnaires, and j represents the respondent number of n respondents. The acceptance tolerance is defined by adding and subtracting the given acceptable tolerance to the mean as follows:

$$AT_i = \mu_i \pm at_i \quad (1)$$

$$\mu_i = \frac{\Sigma x_{i,j}}{n} \quad (2)$$

$$sig_i = P\{x < \mu_i - at_i\} + P\{x > \mu_i + at_i\} \quad (3)$$

Figure 6 shows the conceptual logic of applying Gaussian distribution to validity testing. Every data that lies out of acceptance tolerance is outlier. Platykurtic distribution indicates error of too wide dispersion, due to high standard deviation that is widening the empirical range and enlarging the rejection region greater than the acceptable significance level. Bimodal triggers error of distracted central tendency, since significantly separated mean and mode raise high standard deviation.

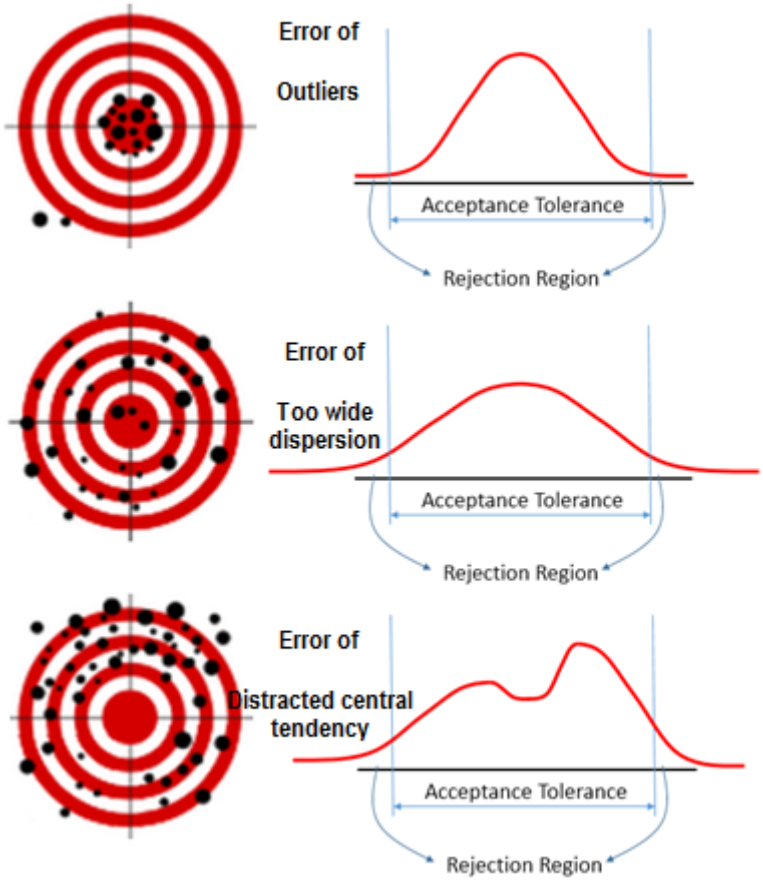


Figure 6. Conceptual Logics of Measurement Error Checking

4. Results

We distributed questionnaires to gather respondent opinion about a service provider performance in order to demonstrate a numerical example of validity test implementation. The Questionnaires contain 12 questions with a 5-points of Likert scales to 30 respondents. We recommend using acceptable tolerance about 1.3 to 1.7 for 5-points of Likert scales. The recapitulation of data gathering result is shown in Table 3. The content of each table cell indicates the number of respondents whose answers are corresponding to the column label for each question.

Table 3. Questionnaires Data Gathering

Question Number	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
1	10	11	7	0	2
2	1	3	8	13	5
3	7	9	2	2	10
4	0	0	2	12	16
5	0	3	2	15	10
6	9	9	6	2	4
7	14	12	4	0	0
8	8	11	5	0	6
9	0	5	23	2	0
10	9	7	2	4	8
11	0	2	3	14	11
12	13	11	6	0	0

We recalculate the measured data of questions which have outliers after removing the outliers (see Table 5). The other questions are not valid and they have to be revised, because of mistakes on question design.

The measured data of each question are calculated to obtain the average, the standard deviation, the acceptance tolerance and the probability of rejection region. Table 4 shows the calculation result. We use 1.5 for acceptable tolerance and 0.10 for acceptable significance level. According to the probability of rejection region, there are 6 questions of 12 questions that are valid ($sig_i < \alpha$), but two of which have outliers.

After removing the outliers, the probability of rejection region decreases becoming smaller than acceptable significance level, $sig_i < \alpha$. This also occurs on two questions, 5th and 11th, which are previously valid but having outliers.

Table 4. Acceptance Tolerance Calculation

Question Number	Mean (μ_i)	Standard Deviation	Acceptance Tolerance (AT_i)	Probability of Rejection (sig_i)	Note
1	2.10	1.09	0.60 – 3. 60	0.170	2 outliers
2	3.60	1.00	2.10 – 5.10	0.135	Too wide dispersion
3	2.97	1.65	1.47 – 4.47	0.363	Distracted cent. tend.
4	4.47	0.63	2.97 – 5.97	0.017	Valid
5	4.07	0.91	2.57 – 5.57	0.098	Valid (3 outliers)
6	2.43	1.36	0..93 – 3.93	0.269	Too wide dispersion
7	1.67	0.71	0.17 – 3.17	0.034	Valid
8	2.50	1.43	1.00 – 4.00	0.295	6 outliers
9	2.90	0.48	1.40 – 4.40	0.002	Valid
10	2.83	1.64	1.33 – 4.33	0.361	Distracted cent. tend.
11	4.13	0.86	2.63 – 5.63	0.081	Valid (2 outliers)
12	1.77	0.77	0.27 – 3.27	0.053	Valid

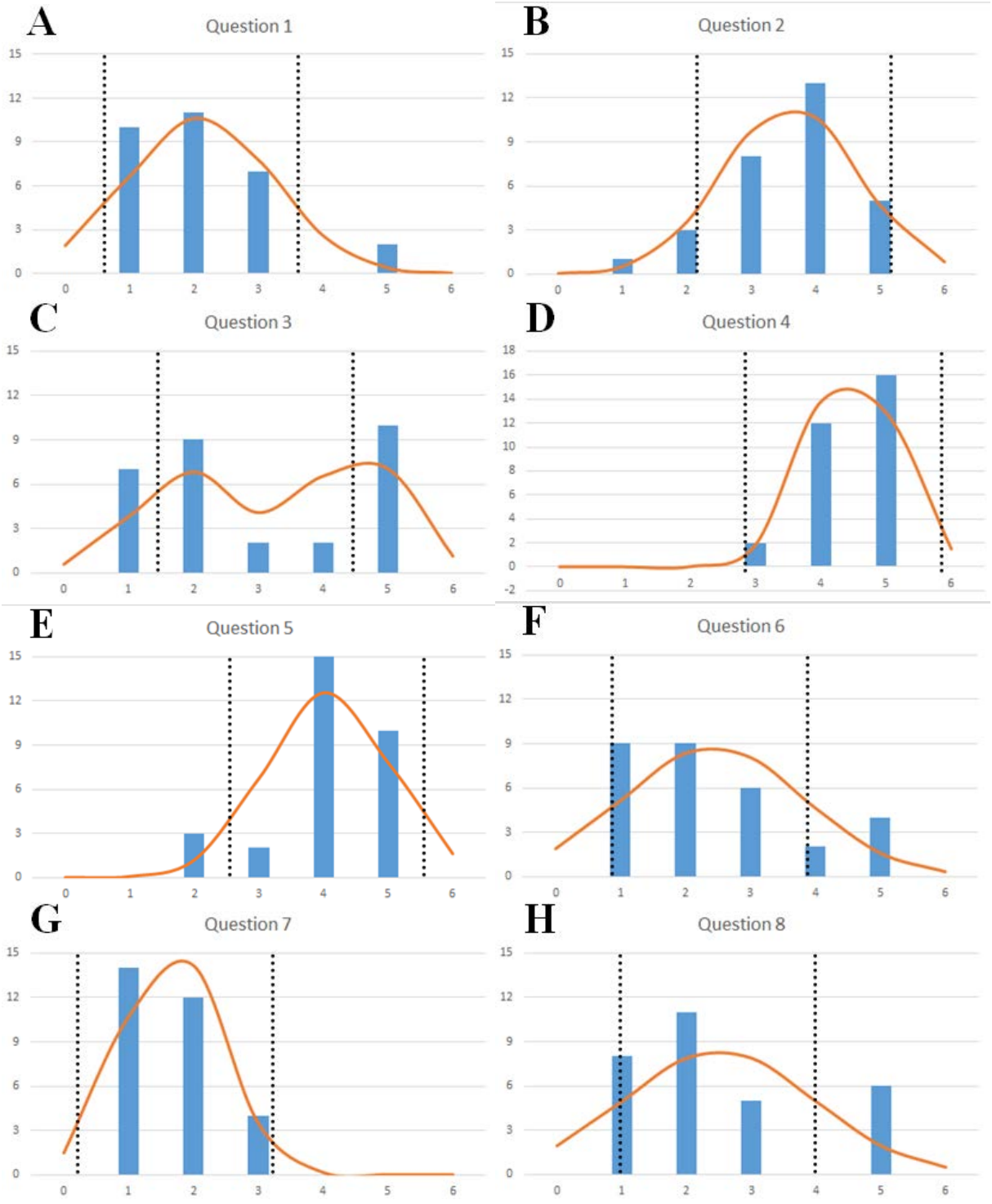
Table 5. Recalculation Result

Question Number	Mean (μ_i)	Standard Deviation	Acceptance Tolerance (AT_i)	Probability of Rejection (sig_i)	Note
1	1.89	0.79	0.39 – 3. 39	0.056	Valid
5	4.30	0.61	2.80 – 5.80	0.014	Valid
8	1.88	0.74	0.38 – 3.38	0.043	Valid
11	4.29	0.66	2.79 – 5.79	0.023	Valid

5. Discussion

The validity testing results of the numerical example are illustrated in Figure 7.

Fig. 7D, 7G and 7I show that the Gaussian distribution of questions (4th, 7th and 9th) has leptokurtic distribution. The probability of rejection region of these questions is respectively 0.017, 0.034 and 0.002 (see Table 4). All respondent opinions are concentrated in the average.


A
Question 1
B
Question 2
C
Question 3
D
Question 4
E
Question 5
F
Question 6
G
Question 7
H
Question 8

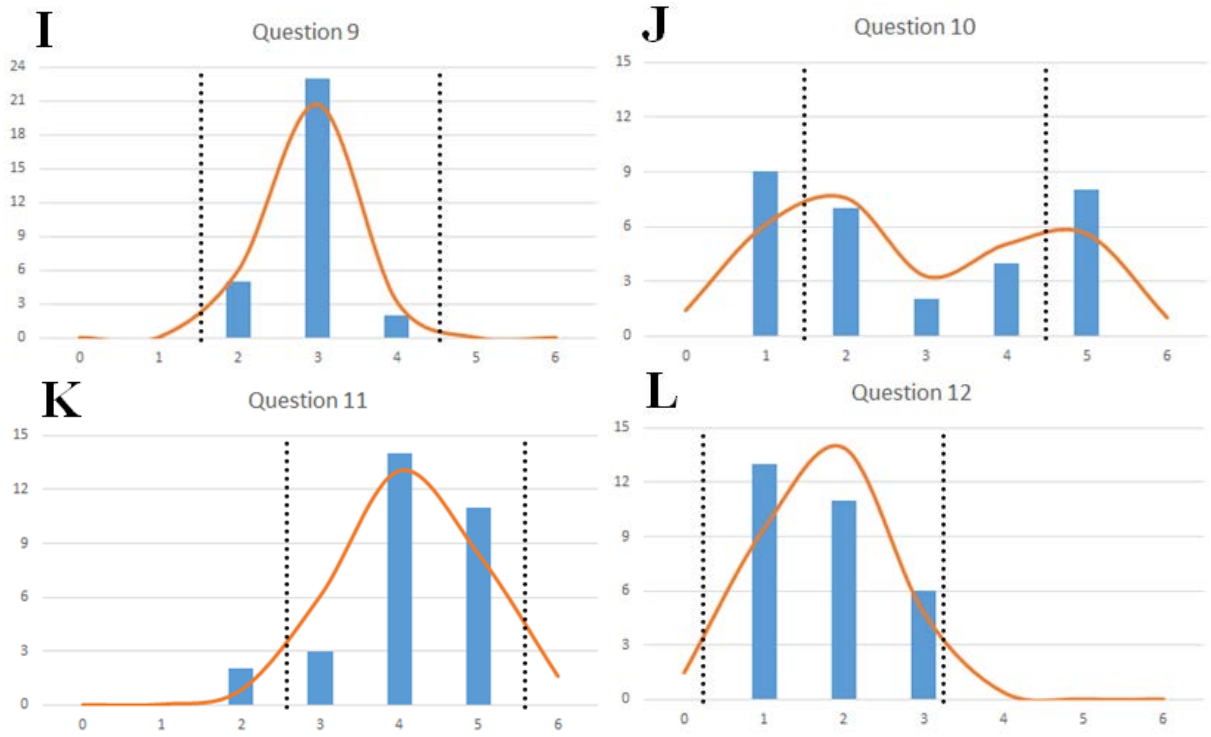


Figure 7. Validity Testing on 5-points of Likert Scale

Fig. 7E, 7K and 7L show that the Gaussian distribution of questions (5th, 11th and 12th) has mesokurtic distribution. The probability of rejection region is respectively 0.098, 0.081 and 0.053. Most of the respondent opinions lie around the average, but there are outliers in two questions.

Fig. 7A and 7B also show that the Gaussian distribution of questions (1st and 2nd) has mesokurtic distribution but with lower peak. The probability of rejection region is respectively 0.170 and 0.135. Respondent opinions are almost equally spread over 3 choices in Fig 7A. There are outliers in 1st question that contribute to higher probability of rejection region. Respondent opinions spread on all choices in Fig 7B. It is an error of too wide dispersion. It makes an unreasonable conclusion because it combines significantly contradictory answers.

Fig. 7C, 7F, 7H and 7J show that the Gaussian distribution of questions (3rd, 6th, 8th and 10th) has platykurtic distribution. The probability of rejection region is respectively 0.363, 0.269, 0.295 and 0.361. The empirical histogram briefly shows bimodal presence. It can be influenced by the presence of bimodal or the existence of outliers. The presence of bimodal is clearer in Fig. 7C and 7J, it indicates that 3rd and 10th questions have error of distracted central tendency. The platykurtic distribution in Fig. 7F and 7H is more influenced by the existence of outliers in 6th and 8th questions. Fig 7F also shows that 6th question has error of too wide dispersion.

Outlier errors occur due to several factors, especially human error of respondent. They can be triggered by respondent failing to understand the question or otherwise the questions beyond the capability of respondent. The questions that contain jargon, slang, foreign language, uncommon terms or abbreviation will lead respondent misunderstanding. Outlier errors can be induced by respondent bias, such as strictness, leniency, recency effect or memorable experience effect. Outlier errors can be caused by errors in respondent selection or response process.

Errors of too wide dispersion occur due to the unclear scale and the content of the question. They can be affected by some mistakes in question design, such as biased question, emotional question, and vague question.

Errors of distracted central tendency occur due to mistakes in question design, such as confusing question, ambiguous question, double barrel question, double negativequestion, and hypothetical question.

The application of the principle of Gaussian distribution in validity testing helps to detect the measurement bias of the questionnaires' responses with result that is caused by mistakes in question design, respondent selection, respondent behavior and response process. It diagnoses the mistakes by checking error of outliers, error of too wide dispersion, and error of distracted central tendency.

6. Conclusions

In the questionnaire, the sources of evidence of measurement bias consist of question design, respondent selection, respondent behavior and response process. Gaussian distribution principles are applied for validity

testing to examine the evidence of bias. It notices to the empirical bell-shaped curve and compares it relatively to the given acceptance tolerance. It finds out the probability of rejection region. The measurements are valid if the probability of rejection region is less than acceptable significance level.

REFERENCES

[1] M.H. Hansen, W.N. Hurwitz and W.G. Madow. Survey Methods and Theory, John Wiley and Sons, New York, 1953.

[2] W.G. Cochran. Sampling Techniques, 3rd Ed. John Wiley and Sons, New York, 1977

[3] J.P. Banda. Nonsampling Errors in Surveys. UN Secretariat, Statistics Division, 2003. Online available from https://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_7.pdf

[4] International Labour Office. Consumer Price Index Manual: Theory and Practice. International Labour Office, Geneva, 2004. Online available from https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf

[5] United Nation, Designing Household Survey Samples: Practical Guidelines. Studies in Methods Series F No.98. Department of Economic and Social Affairs, Statistics Division, United Nations, 2008. Online available from https://unstats.un.org/unsd/demographic/sources/surveys/Series_F98en.pdf

[6] S. Scholtus. Editing and Estimation of Measurement Errors in Administrative and Survey Data. Doctor Dissertation, Faculteit der Sociale Wetenschappen, Vrije Universiteit, Amsterdam, 2018

[7] M. Deblieux. Performance Appraisal Source Book: A Collection of Practical Samples. Society for Human Resource Management, Alexandria-Virginia, 2003.

[8] F.C. Lunenburg. Performance Appraisal: Methods and Rating Errors. International Journal of Scholarly Academic Intellectual Diversity, Vol.14 No.1, 1-9, 2012.

[9] V. Lukovac, D. Pamucar and V. Jovanovic. Assessor Distributional Errors in Evaluating Employee's Work Quality – Identification and Decrease. Proceedings of The XIII International Symposium SYMORG, Zlatibor, 848-853, 2012.

[10] G. Dessler. Human Resources Management, 13th Ed. Pearson Education, New Jersey, 2013.

[11] J.N. Obidinnu, V.E. Ejiofor and B. Ekechukwu. Distributional Errors Normalisation Model (DENoM) for Improving The Variability of Supervisors Appraisal's Ratings. African Journal of Computing & ICT, Vol.7 No.1, 43-48, 2014.

[12] T. Lauwagie, H. Sol and W. Heylen. Handling Uncertainties in Mixed Numerical-Experimental Techniques for Vibration Based Material Identification. Journal of Sound and Vibration, Vol.291, 723-739, 2006.

[13] H. Alkhatib, I. Neumann and H. Kutterer. Uncertainty Modeling of Random and Systematic Errors by Means of Monte Carlo and Fuzzy Techniques. Journal of Applied Geodesy, Vol.3, 67-79, 2009.

[14] C. Blattman, J. Jamison, T. Koroknay-Palicz, K. Rodrigues and M. Sheridan. Measuring The Measurement Error: A Method to Qualitatively Validate Survey Data. Journal of Development Economics, Vol.120, 99-112, 2016

[15] N. Schwarz. Questionnaire Design: The Rocky Road from Concepts to Questions. In L. Lyberg, P. Biemer, M. Collins, E. Deleeuw, C. Dippo and N. Schwarz (editors), Survey Measuremen and Process Quality. John Wiley and Sons, Chichester, 1997.

[16] R.W. Brislin. The Wording and Translation of Research Instruments. In W.J. Lonner and J.W. Berry (editors), Field Methods in Cross-Cultural Research. Sage, California, 1986.

[17] W.A. Belson. The Design and Understanding of Survey Questions. Gower, Aldershot, 1981.

[18] P. Gendall. A Framework for Questionnaire Design: Labaw Revisited. Marketing Bulletin, Vol.9, 28-39, 1998.

[19] K.N. Ross. Quantitative Research Methods in Educational Planning. International Institute for Educational Planning, UNESCO, Paris, 2005.

[20] Y.K. Singh. Fundamental of Research Methodology and Statistics. New Age International, New Delhi, 2006.

[21] E.D. de Leeuw, J.J. Hox and D.A. Dillman. International Handbook of Survey Methodology. Routledge, London, 2008

[22] P. Lietz. Questionnaire Design in Attitude and Opinion Research: Current State of an Art. Priorisierung in der Medizin, FOR 655 Nr.13, 1-20, 2008. Online available from https://pages.stolaf.edu/wp-content/uploads/sites/432/2013/08/Questionnaire-Design-in-Attitude-and-Opinion-Research_Current-State-of-Art-by-Lietz-2008.pdf

[23] W.P. Brinkman. Design of A Questionnaire Instrument. In S. Love (editor). Handbook of Mobile Technology Research Methods. Nova Science Publisher, New York, 2009.

[24] J.A. Krosnick and S. Presser. Question and Questionnaire Design. In P.V. Marsden and J.D. Wright. Handbook of Survey Research 2nd ed. Emerald Group Publishing, Bingley, 2010.

[25] Y. Song, Y.J. Son and D. Oh. Methodological Issues in Questionnaire Design. Journal of Korean Academy of Nursing, Vol.45 No.3m 323-328, 2015.

[26] D.F. Alwin and B.A. Beattle. The KISS Principle in Survey Design: Question Length and Data Quality. Sociological Methodology, Vol.46 No.1, 121-152, 2016.

[27] S. Mourougan an K. Sethuraman. Enhancing Questionnaire Design, Development and Testing through Standardized Approach, IOSR Journal of Business and Management, Vol.19 No.5, 1-8, 2017.

[28] A.N. Oppenheim. Questionnaire Design and Attitude Measurement. Heinemann, London, 1966

[29] G.F. Bishop, R.W. Oldendick, A.J. Tuchfarber and S.E.

Bennett. Effects of Opinion Filtering and Opinion Floating: Evidence from A Secondary Analysis. Political Methodology, Vol.6 No.3, 293-309, 1979.

[30] E.E. Dwyer. Attitude Scale Construction: A Review of The Literature. Retrieved from ERIC database (ED359201). 1993. Online available from https://files.eric.ed.gov/fulltext/ED359201.pdf

[31] S. Presser and J. Blair. Survey Pretesting: Do Different Methods Produce Different Results?. Sociological Methodology, Vol.24, 73-104, 1994.

[32] H. Schuman and S. Presser. Questions & Answers in Attitude Survey. Sage Publications, London, 1996.

[33] B.A. Bergstrom and M.E. Lunz. Rating Scale Analysis: Gauging The Impact of Positively and Negatively Worded Items. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, 1998. Retrieved from ERIC database (ED423289). Online available from https://files.eric.ed.gov/fulltext/ED423289.pdf

[34] P. Darbyshire and H. Mcdonald. Choosing Response Scale Labels and Length: Guidance for Researchers and Clients. Australasian Journal of Market Research, Vol.12 No.2, 17-26, 2004

[35] M.A. Revilla, W.E. Saris and J.A. Krosnick. Choosing The Number of Categories in Agree-Disagree Scales. Sociological Methods & Research, Vol.43 No.1, 73-97, 2014.

[36] J.C. Martin, C. Roman and C. Gonzaga. How Different N-Point Likert Scales Affect The Measurement of Satisfaction in Academic Conferences. International Journal of Quality Research, Vol. 12 No.2, 421-440, 2018.

[37] A. DeCastellarnau. A Classification of Response Scale Characteristics that Affect Data Quality: A Literature Review. Quality & Quantity, Vol.52 No.4, 1523-1559, 2018.

[38] B.A. Babbitt and C.O. Nystrom. Questionnaire Construction Manual Annex Questionnaires: Literature Survey and Bibliography. ARI Research Product 89-21. US Army Research Institute for The Behavioral and Social Sciences. Virginia, 1989. Online available from https://apps.dtic.mil/dtic/tr/fulltext/u2/a213255.pdf

[39] D. Krebs and Y.G. Bachner. Effects of Rating Scale Direction Under The Condition of Different Reading Direction. Methods, Data, Analyses, Vol.12 No.1, 105-126, 2018.

[40] G. Albaum. The Likert Scale Revisited: An Alternate Version. (Product Preference Testing). Journal of The Market Research Society, Vol.39 No.2, 331-342, 1997.

[41] A. Williams. How to ... Write and Analyse A Questionnaire. Journal of Orthodonties, Vol.30, 245-252, 2003.

[42] T.W. Smith. Methods for Assessing and Calibrating Response Scales Across Countries and Languages. Paper presented at the Sheth Foundation/Sudman Symposium on Cross-National Survey Research, Urbana, 2004

[43] R. Pierce. Research Methods in Politics: A Practical Guide. Sage, London, 2008.

[44] N. Golafshani. Understanding Reliability and Validity in Qualitative Research. The Qualitative Report, Vol.8 No.4, 597-607, 2003.

[45] E.A. Drost. Validity and Reliability in Social Science Research. Education Research and Perspectives, Vol.38 No.1, 105-124, 2011.

[46] S. Bajpai and R. Bajpai. Goodness of Measurement: Reliability and Validity. International Journal of Medical Science and Public Health, Vol.3 No.2, 112-115, 2014.

[47] R. Heale and A. Twycross. Validity and Reliability in Quantitative Research. Evidence-Based Nursing, Vol.18 No.3, 66-67, 2015.

[48] H. Mohajan. Two Criteria for Good Measurements in Research: Validity and Reliability. Annals of Spiru Haret University, Vol.17 No.3, 58-82, 2017.

[49] L.D. Goodwin and N.L. Leech. The Meaning of Validity in The New Standards for Educational and Psychological Testing: Implications for Measurement Courses. Measurement and Evaluationin Counseling and Development, Vol.36, 181-191, 2003

[50] S. Shaw and V. Crisp. Tracing The Evolution of Validity in Educational Measurement: Past Issues and Contemporary Challenges. Research Matters, Vol.11, 14-17, 2011.

[51] S. Sireci and J.L. Padilla. Validating Assessments: Introduction to The Special Section. Psicothema, Vol.26 No.1, 97-99, 2014.

[52] American Psychological Association. Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal. American Psychologist, Vol.7, 461-465, 1952

[53] American Psychological Association. Technical Recommendations for Psychological Tests and Diagnostic Techniques. Psychological Bulletin, Vol.51, 1954.

[54] American Psychological Association. Standards for Educational and Psychological Tests and Manuals. American Psychological Association, Washington, 1966.

[55] American Psychological Association, American Educational Research Association and National Council on Measurement in Education. Standards for Educational and Psychological Tests. American Psychological Association, Washington, 1974.

[56] American Educational Research Association, American Psychological Association and National Council on Measurement in Education. Standards for Educational and Psychological Testing. American Psychological Association, Washington, 1985.

[57] American Educational Research Association, American Psychological Association and National Council on Measurement in Education. Standards for Educational and Psychological Testing. American Educational Research Association, Washington, 1999.

[58] American Educational research Association, American Psychological Association and National Council on Measurement in Education. Standards for Educational and Psychological Testing. American Educational Research Association, Washington, 2014.