SPECIAL ISSUE

Bima Journal of Science and Technology, Vol. 7 (2.1) August, 2023 ISSN: 2536-6041

CATEGORIZATION OF COMPOSITE SCORE AND TOTAL SCORE OF LATENT CONSTRUCT

[1]*LAMIDI-SARUMOH ALABA AJIBOLA, [1]ADAMU ISHAKU, [2]ALHARBI NADA MOHAMMEDSAEED, [1]CHAJIRE BUBA PWALAKINO

[1]Department of Mathematics, Faculty of Science, Gombe State University, Gombe, Nigeria

[2]Department of Mathematics, Faculty of Science, Taibah University, Saudi Arabia

Corresponding Author: lalabaajibolasarulam@gsu.edu.ng

ABSTRACT

The categorization of composite scores and total score is an important step to achieve the aim of some quantitative research involving latent construct, but the required understanding about how the categorization should be done aptly based on the features of the data is lacking. The purpose of this study is to guide researchers on how categorization of composite scores and total score are done based on the features of the data. Consideration of the measure of central tendency of the data, the presence of outliers in the composite score and total score are factors that are necessary for consideration in order to validate the categorization. Three different approaches of categorization were deliberated to show that winsorized composite scores yielded an unbiased mean of composite score which can represent the overall response of a target population in research involving latent construct.

Keywords: categorization; composite score; total score; latent construct

INTRODUCTION

Latent constructs are variables that cannot be measured directly or observed directly, these variables can only be observed by the use of instruments which are made up of indicators of the constructs (Curado, Teles, & Marôco, 2014). Categorization is the process of sorting and organizing of ideas, variables into classes or groups which are differentiated into set of basic concepts (Goldstone & Kersten, 2003). The classical theory of categorization is meant to define attributes related to particular group which can simplify complex information for researchers from different fields to easily understand the information or results from another study. Composite score is a single data point that represents combination of information from multiple variables/items which form a latent construct (Hair, Howard, & Nitzl, 2020). In the context of this study, composite score is the weighted average of the scores/scales given as response to the items which represents the indicators of a latent construct. Dividing the total score of responses to indicators of a latent construct by the number of items in the constructs yield the composite score (Lin, Ward, & Fine, 2013). Some latent construct can be assessed with the composite score while some others can be measured with the total score. For instance, latent construct such as speaking skill is measured by addition of scores on grammar, fluency, pronunciation, vocabulary and comprehension (total score) while latent construct such as anxiety of a target population is measured by composite scores. The nature of the instrument, the discipline and the research goals are determinants of the type of scores that will be used for the analysis or assessment of the research.

The advent of behavioural science toward health and diseases (Glanz & Bishop, 2010), social media usage among others important subjects which are affecting the world today have made composite score and total score increasingly used in reporting past and present events. In community health medicine and nursing, composite or total scores are used to verify the effect of interventions to improve the health and wellbeing of a particular populace or place. The importance of composite scores in social sciences, where variables considered are mostly latent constructs cannot be downplayed. Thus, accurate categorization of such vital variable is very important to strengthen scientific results. Aside the important role of composite score and total score, it has been proven that composite scores are better than component score in assessment of latent constructs (Crone et al., 2016; Matsumura et al., 2019; Pilitsis, Fahey, Custozzo, Chakravarthy, & Capobianco, 2021)

The existence of outliers in the composite score and total score may influence the categorization because most categorizations are based on the dispersion of the data (range). The minimum and the maximum scores are used as a base for the computation of most categories. For instance, a researcher may be interested in categorizing a total score into three different level (low, moderate and high) or two categories (good or poor). This categorization requires the minimum composite score to be subtracted from the maximum composite score and the result will be divide into three or two to form the category intervals for each level, respectively. In such a case, the range of the data is the determinant of the category interval. If there is an existence of extreme outliers at both end of the data, the category interval calculated with such data will be biased.

$$\text{Category interval} = \frac{X_{\text{max score}} - X_{\text{min score}}}{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

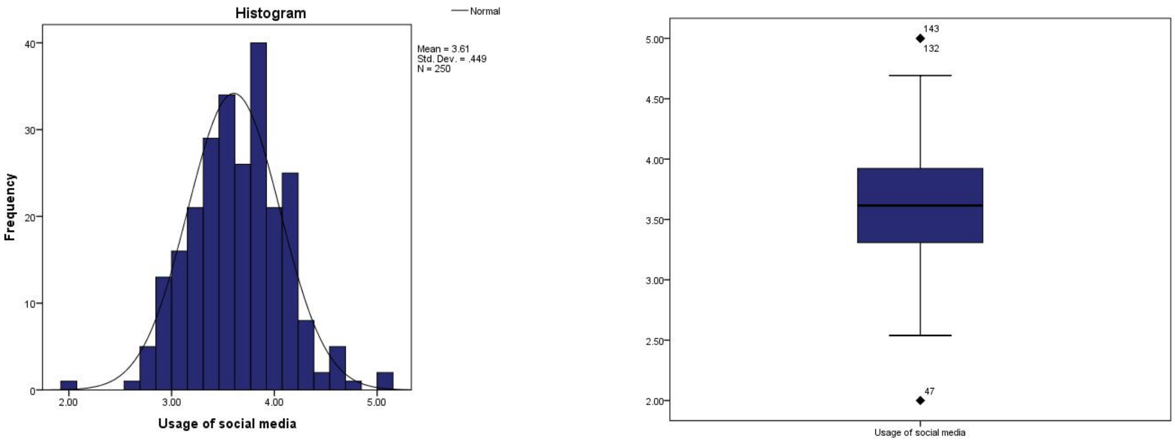
where N = The number of category

$X_{\text{max score}}$ = Maximum number among the composite score/total score

$X_{\text{min score}}$ = Minimum number among the composite score/total score

ILLUSTRATION

Real-life psychometric data collected in a cross-sectional survey to measure strategies which are used to alleviate English language writing anxiety among college students was used to show extreme end outliers. It can be observed from the histogram and boxplot in Figure 1, that the extreme values are outlier from both extreme end of the data. The outliers need to be treated before appropriate categorization of the data.



(a) (b)

Figure 1: (a) Histogram and (b) boxplot depicting the outliers of a data.

In some cases where the latent construct is measured based on 5-point Likert scale (from 1 to 5), it is easy to assume the maximum composite score to be 5 and the minimum score to be 1, However, this assumption may be faulty if the data has legitimate outliers and skewness. Due to the spuriousness of a real-life data, it is important to observe the features of the data before categorization. The objective of this study is to guide researchers on how categorization of composite scores and total score are done based on the features of the data especially in the presence of outliers.

There are various ways of detecting outliers in a data, such as boxplots and histogram, as shown in Figure 1 above. The calculation of mean, standards deviation, inter quartile range, percentile and z-score can also be used to detect the presence of outlier (Smiti, 2020). Among the three most popular central tendency, mean, median and mode, the mean of the data set is mostly affected by outliers (Bickel, 2003). In some quantitative research which involve the use of latent construct, the mean of composite score of the latent construct may be used as a yardstick to measure the overall score of latent construct. Deleting an outlier may lead to total loss of information if the outlier is genuine. Moreover, removal of outliers may influence the projection of future research mostly in comparative and interventional study. There are three major ways to treat outliers in a data set aside removing outlier to avoid total loss of information, that is, reducing the weights of the outlier which require trimming the weights, changing the values of the outliers (winsorization and imputation), and using robust estimation techniques such as *m*-estimation. The degree of the outlier is a factor that needs to be considered in order to choose an appropriate method to deal with it. The degree of the outlier can take two different forms: mild outliers (i.e., beyond an “inner fence” on either side) and extreme outliers (i.e., beyond the outer fence). Apart from categorization, in some research involving questionnaire, composite score and total score are used as independent and dependent variables to achieve the aim of the research (t-test, analysis of variance, regression analysis etc.). After weighing the pros and cons of the methods to deal with the outliers evolving from the composite score and total score, it was discovered that limiting extreme values in the composite score or total score to reduce the effect of outliers is the simplest way, this method is called winsorization. Winsorization reduces the high variability of the data, which in turn reduces the range of data. It also reduces biased coefficient and distorted goodness of fits measure. It has also been proved that the winsorization of dependent variable influence the mean square error in a regression analysis (Taylor, Lien, & Balakrishnan, 2005).

WINSORIZATION

Generally, winsorization is the transformation of a data set by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. There is a possibility that outliers can strongly influence the distribution of many statistical data. A typical strategy is to set all outliers (values beyond a certain threshold) to a specified percentile of the data. For example, a 90% winsorization would see all data below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile. This can simply be attained by replacing the extreme value with a certain percentile from each end. Winsorized estimators are usually more robust to outliers than their more standard forms (Yang, Xie, & Goh, 2011). Winsorization was named after an engineer-turned-Biostatistician, Charles P. Winsor (1895–1951). The effect of winsorization on a data is the same as clipping in signal processing (Leonowicz, Karvanen, & Shishkin, 2005). In order to ascertain the effect of winsorizing the composite score, three methods were considered.

Winsorization of Data using R Codes

Winsorise (x, minval =NULL, maxval = NULL, probs = c(0.05, 0.95), na.rm = FALSE, type =7)

Argument

x	A numerical vector to be winsorized
minval	The lower border, all values being lower than this will be replaced by this value. The default is set the 5%-quantile of x
maxval	The high border, all values being larger than this will be replaced by this value. The default is set to the 95%- quantile of x
probs	numeric vector of probabilities with values in [0,1] as used in quartile
na.rm	NAs should be omitted to calculate the quantiles
type	An integer between 1 and 9 selecting one of the nine quantile algorithm detailed in the quantile to be used.

Categorization with the limits of Likert scale

Suppose the Likert scale considered for the above data start from 1 to 5 and the research is interested in 3 categories of strategies. Then, from equation (1),

$$\text{Category interval} = \frac{5 - 1}{3} = 1.33$$

Table 1: Categorization based on Likert scale limits

Strategies	Categorization	Frequency	Percentage
Low	1.00 -2.33	1	0.4
Moderate	2.34 – 3.66	135	54.0
High	3.67 – 5.00	114	45.6

Mean of Composite score = 3.80

Categorization with the minimum and maximum values of the data

The minimum and maximum of the above data is 2 and 5, respectively, as shown on the histogram in Figure 1 with inclusion of the outliers, from equation (1),

$$\text{Category interval} = \frac{5 - 2}{3} = 1.00$$

Table 2: Categorization based on minimum and maximum values of the data

Strategies	Categorization	Frequency	Percentage
Low	2.00 – 3.00	26	10.4
Moderate	3.01 – 4.00	181	72.4
High	4.01 – 5.00	43	17.2

Mean of Composite score = 3.80

Categorization after 90% winsorization of the data

After winsorization of the data, the minimum and the maximum values are 2.54 and 4.69, respectively. From equation (1)

$$\text{Category interval} = \frac{4.69 - 2.54}{3} = 0.72$$

Table 3: Categorization after winsorization of the data

Strategies	Categorization	Frequency	Percentage
Low	2.540 – 3.26	57	22.8
Moderate	3.261 – 3.98	141	56.4
High	3.981 – 4.69	52	20.8

Mean of Composite score = 3.50

DISCUSSION

In order to obtain an unbiased mean on the same set of data, three categorizations were considered. It was observed that the limits of the Likert scale accommodates all the outliers, which gives room for generalization of the results (comparison of results from the control group to the experimental group, results from one location to another). However, the presence of outliers invalidates the result if the study is meant for a particular target population and does not involve comparison, such as cross-sectional survey. Furthermore, if the research involves the use of the mean as a benchmark to quantify all the responses of the participants, the result will be biased. The presence of outliers weakened the existence of college students who were categorized as having low strategy to alleviate English language writing anxiety and maximized the number of college students with moderate and high strategy in English language writing anxiety (Table 1).

Categorization with the min-max of the data with inclusion of the outliers yields a result that seems to show balance in the categorization of the data (Table 2). However, the problem that arises with the usage of Likert scale limits will also emerge if the research is interested in the overall mean of the composite score to quantify the latent construct due to the presence of outliers. Categorization with the winsorized data and using the max-min of the data showed an unbiased categorization of the data based on a target population and also yielded a valid mean of the overall composite score which can represent the latent construct (Table 3).

CONCLUSION

The winsorized mean (Table 3) has an advantage over the first two methods (Table 1 and Table 2) because the mean of composite score calculated after winsorizing the data can be used as a benchmark to measure the overall mean of the latent construct representing a target population. Also, the winsorized composite scores can be used in advanced statistical methodologies which can yield precise estimates.

REFERENCES

Bickel, D. R. (2003). Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency. *Journal of Statistical Computation and Simulation*, *73*(12), 899–912.

Crone, L. J., Lang, M. H., Franklin, B. J., Halbrook, A. M., Crone, L. J., Lang, M. H., … Franklin, B. J. (2016). Composite Versus Component Scores : Consistency of School Effectiveness Classification Composite Versus Component Scores : Consistency of School Effectiveness Classification, *7347*(June).

Curado, M. A. S., Teles, J., & Marôco, J. (2014). Analysis of variables that are not directly observable: Influence on decision-making during the research process. *Revista Da Escola de Enfermagem*, *48*(1), 146–152.

Glanz, K., & Bishop, D. B. (2010). The role of behavioral science theory in development and implementation of public health interventions. *Annual Review of Public Health*, *31*, 399–418.

Goldstone, R. L., & Kersten, A. (2003). *Concepts and Categorization*. *Handbook of Psychology*.

Hair, J. F., Howard, M. C., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research*, *109*(November 2019), 101–110.

Leonowicz, Z., Karvanen, J., & Shishkin, S. L. (2005). Trimmed estimators for robust averaging of event-related potentials. *Journal of Neuroscience Methods*, *142*(1), 17–26.

Lin, M. S. F., Ward, S. E., & Fine, J. P. (2013). Composite Variables, *62*(1), 45–49.

Matsumura, K., Kumar, T. P., Guddanti, T., Yan, Y., Blackburn, S. L., & McBride, D. W. (2019). Neurobehavioral Deficits After Subarachnoid Hemorrhage in Mice: Sensitivity Analysis and Development of a New Composite Score. *Journal of the American Heart Association*, *8*(8).

Pilitsis, J. G., Fahey, M., Custozzo, A., Chakravarthy, K., & Capobianco, R. (2021). Composite Score Is a Better Reflection of Patient Response to Chronic Pain Therapy Compared With Pain Intensity Alone. *Neuromodulation*, *24*(1), 68–75.

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306.

Taylor, P., Lien, D., & Balakrishnan, N. (2005). On Regression Analysis with Data Cleaning via Trimming , Winsorization , and Dichotomization On Regression Analysis with Data Cleaning via Trimming , Winsorization , and Dichotomization. *Communications in Statistics - Simulation and Computation*, (34), 839–849.

Yang, J., Xie, M., & Goh, T. N. (2011). Outlier identification and robust parameter estimation in a zero-inflated Poisson model. *Journal of Applied Statistics*, *38*(2), 421–430.