Behaviormetrika (2019) 46:389–399
https://doi.org/10.1007/s41237-019-00089-1

SHORT NOTE

Decline of Pearson's *r* with categorization of variables: a large-scale simulation

Takahiro Onoshima[1] · Kenpei Shiina[1] · Takashi Ueda[1] · Saori Kubo[2]

Received: 18 February 2019 / Accepted: 20 July 2019 / Published online: 1 August 2019

Abstract
It is often said that correlation coefficients computed from categorical variables are biased and thus should not be used. However, practitioners often ignore this longstanding caveat from statisticians. Although some studies have examined the bias, the true extent is still unknown. This study is an extensive attempt to determine the range and degree of the biases. In our simulation, continuous variables were categorized according to various thresholds and used to compute Pearson's *r*. The results indicated that there were more serious biases than highlighted in previous studies. The results also revealed that increasing data size did not reduce the biases. Possible ways to cope with the biases are discussed.

Keywords Correlation coefficient · Categorization bias · Number of categories · Likert scale

1 Introduction

It is a common practice in social sciences to compute Pearson's correlation coefficient *r* from ordered categories by assigning integers to the categories, as in a Likert scale. In fact, Karl Pearson, who defined the coefficient, computed *r* from categorized variables but he noticed that *r* is biased when the number of categories is small and, therefore, "broad". He proposed some remedies to address this issue (Pearson 1913). Ritchie-Scott (1918) then proposed the polychoric correlation coefficient and Pearson and Pearson (1922) improved it. However, an executable version of the polychoric correlation coefficient took a long time to appear (Olsson 1979). Despite the longstanding caveat by psychometricians, very few people attempt to use the polychoric correlation coefficient.

Communicated by Kohei Adachi.

✉ Takahiro Onoshima
onoshima.t@gmail.com

[1] Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan

[2] Tokyo Women's Medical University, 8-1, Kawada-cho, Shinjuku-ku, Tokyo 162-8666, Japan

Evidently, researchers do understand the importance of the categorization bias. In marketing science, simulations have been conducted to examine the extent of biases (Morrison 1972; Martin 1973, 1978). According to Martin (1978, p. 307), "the amount of lost information is substantial". In sociology, Bollen and Barb (1981) also conducted simulation studies contrasting the correlation between two original continuous variables and their categorized versions. They concluded that the differences are generally small, but grow when there is high correlation between original continuous variables and the number of categories is small.

These studies seem to have correctly described the global tendency of the biases but have failed to incorporate two important points. First, few studies considered the situation in which the number of categories is different. Shiina et al. (2012) proved that when different numbers of ordered categories are used, Pearson's r cannot be -1 or 1 when: (1) variable X has m (≥ 2) ordered categories and variable Y has n (≥ 2) ordered categories, (2) $n \neq m$, and (3) these categories are used at least once. A simpler new proof is as follows. If all the data are on an oblique line, then $r = 1$ and vice versa. If all the data are on the line, then the number of orthogonal images of the data on X and Y axes should be identical. Therefore, $r = 1$ implies that the number of such images should be identical. From the contrapositive of the proposition, we can conclude that if the numbers of orthogonal images on both axes (the number of categories) are not the same, r cannot be 1. In view of this proof, it is imperative to pay close attention to the situation in which the number of categories is different.

Second, past studies have not extensively examined the effect of the arrangement of thresholds at which original continuous variables are converted into categorized (or integer-valued) variables. This is important because a disorderly arrangement of thresholds can easily destroy the structure of the original continuous distribution.

This paper examines the effects of conversion of continuous variables into categorized ones on the decline of the correlation coefficient, using different numbers of categories and various thresholds. We will first demonstrate how categorized variables with different numbers of categories and disorderly thresholds yield large biases of r. Then, we will run a large-scale simulation and report the full extent of biases of r.

2 Assumptions on the data generating process

Let x and y be two continuous latent variables obeying a bivariate normal distribution (BND):

$$\phi\left(x, y|\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\right) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \mathrm{e}^{-\frac{1}{2}\left\{\frac{(x-\mu_x)^2}{\sigma_x^2\left(1-\rho^2\right)} - \frac{2(x-\mu_x)\left(y-\mu_y\right)}{\sigma_x\sigma_y\left(1-\rho^2\right)}\rho + \frac{\left(y-\mu_y\right)^2}{\sigma_y^2\left(1-\rho^2\right)}\right\}}. \tag{1}$$

It is assumed that the original variables are categorized and yield manifest variables X and Y. We should consider the number of categories (m for X and n for Y), as well as the arrangement of thresholds, because how we divide the original continuous latent variables into categories will strongly affect the extent of biases. We can

set $\mu_x = \mu_y = 0$ and $\sigma_x = \sigma_y = 1$ without loss of generality and can define thresholds for x and y as

$$\begin{aligned} -\infty = \theta_0 < \theta_1 < \theta_2 < \cdots < \theta_m = \infty \\ -\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_n = \infty \end{aligned} \tag{2}$$

such that, if $\theta_{i-1} < x < \theta_i$, then $X = i$;if $\tau_{j-1} < y < \tau_j$ then $Y=j$.

The true probability γ_{ij} of each cell in the contingency table (correlation table) corresponds to the rectangular region $[\theta_{i-1}, \theta_1]\times[\tau_{j-1}, \tau_j]$ in the x–y space and is given by

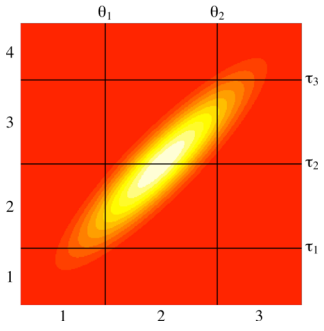
$$\gamma_{ij} = \int_{\theta_{i-1}}^{\theta_i} \int_{\tau_{j-1}}^{\tau_j} \phi\left(x, y|\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\right) \mathrm{d}y\mathrm{d}x \tag{3}$$
$$i = 1, 2, \ldots, m; j = 1, 2, \ldots, n.$$

Figure 1 illustrates the original BND and an example of true probabilities of each cell in the contingency table.

3 Expected *r* when computing from categorized variables

We can compute the expected values of r with categorized variables using true probabilities of each cell. Expected r is given by

$$E(r) = \sum_{X=1}^{m} \sum_{Y=1}^{n} \gamma_{XY} \times \frac{(X - E(X))(Y - E(Y))}{\sqrt{V(X)V(Y)}}. \tag{4}$$



		X		
		1	2	3
Y	4	0.000	0.005	0.061
	3	0.001	0.336	0.096
	2	0.096	0.336	0.001
	1	0.061	0.005	0.000

Computed r from the table

$r = 0.692$

Fig. 1 Left: heatmap of the original continuous distribution ϕ (x, y| 0, 0, 1, 1, 0.9); right: true probabilities of each cell γ_{ij} defined by Eq. (3) for $m=3$, $n=4$, $\theta_1 = -1, \theta_2 = 1, \tau_1 = -1.5, \tau_2 = 0$ and $\tau_3 = 1.5$, and expected r computed from Eq. (4)

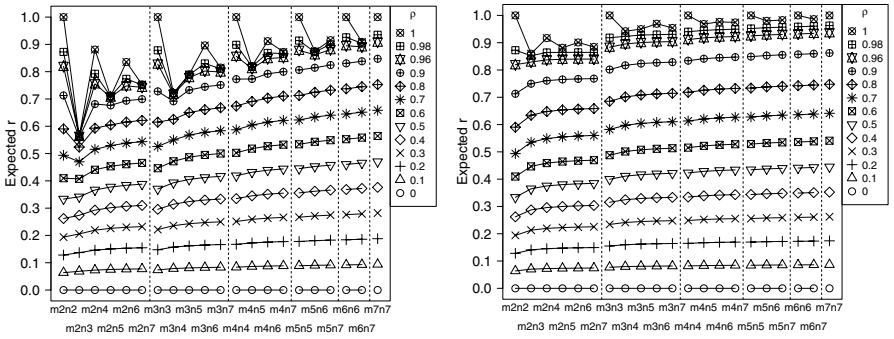


Fig. 2 Expected values of r computed from categorized variables (in the left panel, thresholds are placed in $[-3, 3]$; in the right panel, thresholds are placed in $[-1, 1]$)

Figure 2 depicts expected values of r where there are two to seven categories and thresholds almost equally divide the interval from -3 to 3 and the interval from -1 to 1. More precisely, thresholds are defined as

$$\theta_0 = -\infty, \theta_i = -3 + \frac{6}{m}i,\ i = 1, 2, \ldots, (m-1),\ \theta_m = \infty$$
$$\tau_0 = -\infty, \tau_j = -3 + \frac{6}{n}j,\ j = 1, 2, \ldots, (n-1), \tau_n = \infty$$

in the left panel of Fig. 2, and

$$\theta_0 = -\infty,\ \theta_i = -1 + \frac{2}{m}i,\ i = 1, 2, \ldots, (m-1),\ \theta_m = \infty \tag{5}$$

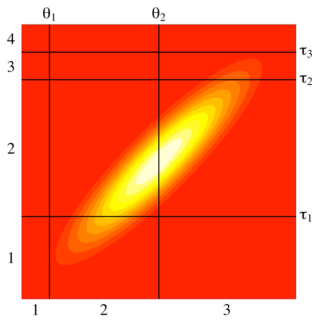
$$\tau_0 = -\infty, \tau_j = -1 + \frac{2}{n}j,\ j = 1, 2, \ldots, (n-1), \tau_n = \infty \tag{6}$$

in the right panel of Fig. 2.

Figure 2 shows two general tendencies. One is that a smaller number of categories increase biases of r and the other is that biases of r become greater as ρ increases.

The computation of expected r in Fig. 2 used equalized categories and we did not fully consider the location of thresholds. While equalized categories are partitioned by "well-organized" thresholds, there is some type of the arrangement of thresholds that destroys the structure of original continuous distribution. Such "ill-organized" thresholds will induce more serious biases and Fig. 3 shows an example that categorizing continuous variables with ill-organized thresholds ($\theta_1 = -2, \theta_2 = 0; \tau_1 = -1, \tau_2 = 1.5, \tau_3 = 2$) generates a severe decline of the correlation coefficient compared to well-organized thresholds shown in Fig. 1.

By "well-organized thresholds," we mean a set of thresholds that keeps the properties of the original BND, which includes symmetric and single-peaked shape, no void region in the center of the distribution, and no-overconcentration. By "ill-organized



		X			
		1	2	3	
Y	4	0.000	0.000	0.023	Computed r from the table
	3	0.000	0.000	0.044	
	2	0.000	0.342	0.433	$r = 0.494$
	1	0.023	0.135	0.000	

Fig. 3 An example of categorizing continuous variables with ill-organized thresholds: true probabilities of each cell γ_{ij} defined by Eq. (3) for $m=3$, $n=4$, $\theta_1 = -2, \theta_2 = 0, \tau_1 = -1, \tau_2 = 1.5$ and $\tau_3 = 2$, and expected r under the same original continuous distribution in Fig. 1 (ϕ (x, y| 0, 0, 1, 1, 0.9))

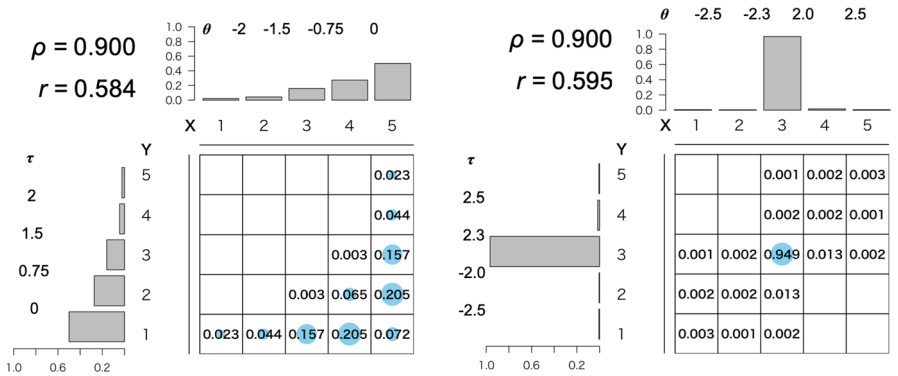


Fig. 4 Some examples of categorical distributions with ill-organized thresholds (the value and the size of dot in each cell show the cell probability)

thresholds," we mean the opposite, that is, a set of thresholds that yields asymmetry, multiple-peaks, voids, concentrations, monotone decreasing, or increasing.

Because these properties are qualitative and vague and thus are difficult to represent numerically, we present some examples for further understanding. Figure 4 illustrates how some sets of ill-organized thresholds destroy the original structure of BND, which results in the decline of r. It is noted that asymmetry of the distributions of two categorized variables causes massive decline of r as in the left panel of Fig. 4 and overconcentration into one cell also decreases r considerably as in Fig. 4 right panel.

4 Simulation

In our simulation, four factors were manipulated: ρ of BND, ϕ (x, y | 0, 0, 1, 1, ρ), data size, the number of categories, and the thresholds.

We generated a pair of random numbers (x and y) from BND. We then categorized the two continuous variables into two integer-valued (Likert) variables X and Y and computed the correlation between the categorized variables. Table 1 shows the factors and the levels in our simulation.

Because one of the aims of our simulation is to examine how various threshold settings affect the bias of correlation, we set up two threshold settings. One setting used thresholds from continuous uniform distribution. This was because uniform distribution is ordinarily used when no reasonable prior information is available. More precisely, random numbers were generated from continuous uniform distribution $U(-1, 1)$ and then they were arranged in ascending order. We call this "the uniform setting" for short. According to a standard result from order statistics (David and Nagaraja 2003), kth threshold is beta-distributed with

$$\text{mean} = -1 + \frac{2k}{l+1}, \quad \text{variance} = \frac{4k(l-k+1)}{(l+1)^2(l+2)},$$

where l denotes the number of thresholds. Therefore, the locations of thresholds tend to be systematic while the thresholds in middle position will have a large variance.

The other setting used thresholds with small "noise," which is a random threshold version of the situation in the right panel of Fig. 2, which will be more familiar to psychometricians (the Law of Categorical Judgment, Torgerson 1958). We call this "the equal setting" for short. To avoid possible crossovers of the fluctuated thresholds, truncated normal (TN) was used. The general form of TN probability density is given by

$$f(z;\mu,\sigma^2,a,b) = \frac{\phi\left(\frac{z-\mu}{\sigma}\right)}{\sigma\left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right)},$$

where $\phi(\cdot)$ is the standard normal distribution, $\Phi(\cdot)$ is the cumulative standard normal distribution, and $[a, b]$ is the domain. In the current situation, we set $\mu = -1 + \frac{2}{m}i$ using (5), and $\sigma^2 = (0.05)^2$, $a = \mu - \frac{1}{m}$, and $b = \mu + \frac{1}{m}$ to represent the probability distribution of threshold θ_i. Therefore, we have

$$\theta_i \sim \text{TN}\left(-1 + \frac{2}{m}i, (0.05)^2, -1 + \frac{2i}{m} - \frac{1}{m}, -1 + \frac{2i}{m} + \frac{1}{m}\right),$$

which means that the mean, mode, median of the probabilistic variable θ_i are the same due to the positioning of μ, a, b such that $b - \mu = \mu - a = 1/m$. The restriction on thresholds:

$$-\infty = \theta_0 < \theta_1 < \theta_2 < \cdots < \theta_m = \infty,$$

is always conserved in this sampling scheme. The probabilistic variable τ_j was determined in a similar manner.

The uniform and the equal settings are completely different sampling methods that generate different multidimensional distributions of thresholds. A uniform setting can produce more pathologically ill-organized thresholds than an equal

Table 1 Factors and levels in the simulation

Factors	Levels	Note
ρ	0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.96, 0.98,1.0	
Data size	64, 256, 1024	
Number of categories	m2n2, m2n3, m2n4, m2n5, m2n6, m2n7, m3n3, m3n4, m3n5, m3n6, m3n7, m4n4, m4n5, m4n6, m4n7, m5n5, m5n6, m5n7, m6n6, m6n7, m7n7	m for X, n for Y
Thresholds setting	The uniform setting 10,000 pairs of random threshold sets were determined as follows: $(m-1)$ and $(n-1)$ random numbers were generated from continuous uniform distribution $U(-1, 1)$. They were then arranged in ascending order. Pairs of these arranged sets were used for categorizing x and y into X and Y	
	The equal setting 10,000 pairs of random threshold sets were determined as follows: First, $(m-1)$ and $(n-1)$ points are defined as in (5) or (6). Then, to each point, a random number from doubly truncated normal distribution TN $(0, 0.05^2, a, b)$, where $[a, b]$ is the domain, $a = -1/m$ and $b = 1/m$ for X and $a = -1/n$ and $b = 1/n$ for Y, was added to represent a fluctuating threshold	

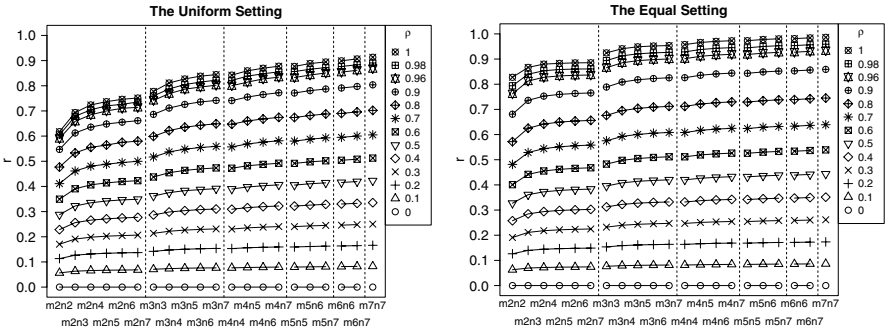


Fig. 5 Average values of r (left panel: the uniform setting; right panel: the equal setting; data size: 1024)

setting. For example, a set of thresholds $(-\infty, 0.81, 0, 82, 0.83, \infty)$, when $m = 4$, is possible only for the uniform setting. All the sets of thresholds generated from equal setting can be generated (with different probability) from uniform setting but not vice versa.

In both threshold settings, we used a range $[-1, 1]$ for generating a set of thresholds for the following reasons. First, since overconcentration causes considerable decline of r as in Fig. 4, it is fruitless to set too large of a range, $[-20, 20]$ for example, that likely induces an overconcentration and voids in both sampling schema. Moreover, it might produce zero variance, which will induce division by zero in (4). Second, this study is a first attempt to examine the threshold effect on the decline of r; therefore, it is somewhat arbitrary because we should start from somewhere.

In each combination of four factors (ρ, data size, pairs of the number of categories, threshold settings), we computed Pearson's correlation coefficient 1000 times. In this way, we utilized a total number of 16.38 billion ($= 13\ \rho s \times 3$ data size $\times 21$ pairs of the number of categories $\times 2$ threshold settings $\times 10{,}000$ threshold sets $\times 1000$ times) correlation coefficients between two categorized variables.

5 Results

The average values of r between categorized variables are depicted in Fig. 5. Regardless of data size, category size, and threshold location, average value of r showed robust underestimation of ρ, but the pushdown bias decreased as the number of categories increased. Not surprisingly, the extent of bias differed between two thresholds settings.

Comparing two threshold settings, the uniform setting caused more serious decline of r. For example, in the case where $m = 3$, $n = 4$, $\rho = 0.9$, and data size is 1024, the average value of r in the uniform setting was 0.726 while the value was 0.818 in the equal setting.

Compared with expected values of r with well-organized thresholds in the right panel of Fig. 2, while no marked discrepancies were observed in the equal setting except for special cases where $\rho = 1.0$, there were substantial declines

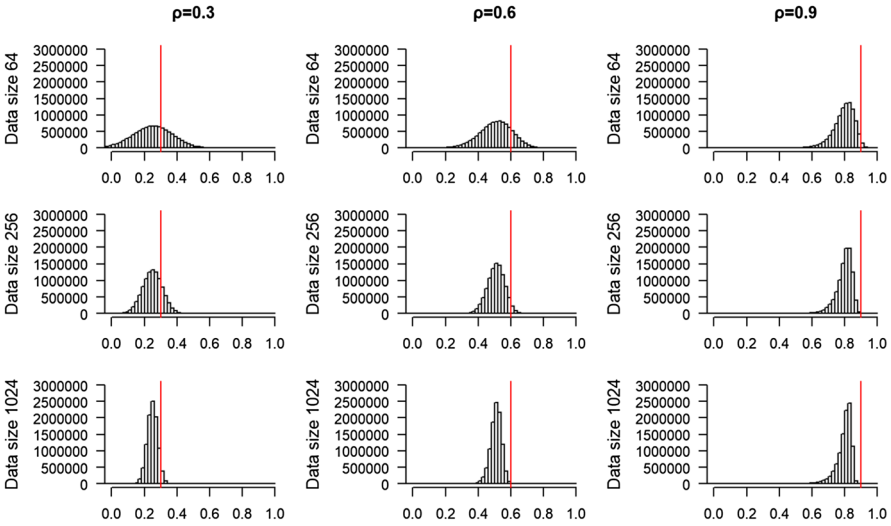


Fig. 6 Distributions of r where $m=6$ and $n=7$ (vertical line depicts ρ)

in the average values of r in the uniform setting. For example, in the case where $m=4$, $n=5$, and $\rho=0.8$, the expected r with well-organized thresholds ($\theta_1=-0.5, \theta_2=0, \theta_3=0.5; \tau_1=-0.6, \tau_2=-0.2, \tau_3=0.2, \tau_4=0.6$) is 0.726, while the average value of r in the uniform setting, where data size is 1024, is 0.660. To provide another example, in the case where $m=5$, $n=7$, and $\rho=0.8$, the expected r with well-organized thresholds ($\theta_1=-0.6, \theta_2=-0.2, \theta_3=0.2, \theta_4=0.6$; $\tau_1=-0.71, \tau_2=-0.43, \tau_3=-0.14, \tau_4=0.14, \tau_5=0.43, \tau_6=0.71$) is 0.688, whereas in the uniform setting, where data size is 1024, it is 0.635.

The cause of greater bias in the uniform setting could be that the simulation results include both well-organized and ill-organized thresholds. The uniform setting allows a set of thresholds to be ill-organized. For example, when the distance of thresholds is very close, a resulting contingency table tends to include an empty or almost empty category. In addition, when all the values of thresholds approach to upper or lower limits, a resulting table tends to be asymmetric. It is reasonable that such transformation of the original distribution causes considerable decline of r as indicated in Fig. 4, though the uniform setting also allows well-organized thresholds. On the other hand, such destructive transformation of the distribution is not possible in the equal setting. Therefore, it is plausible that the difference between two settings is derived from whether the setting tends to allow ill-organized thresholds.

Figure 6 shows the effect of data size on variations (sampling distribution) of r where $m=6$ and $n=7$ in the uniform setting. It is revealed that larger data size decreases variations of r but does not shift the central position of distribution. This means that increasing data size does not reduce systematic biases of r, but it only accurately estimates biased r. This demonstrates a simple fact that, because a categorization is a non-linear transformation, as soon as we transform an original

continuous distribution into a categorized distribution, we have two different distributions and parameters estimated from different distributions are not generally the same.

It is very difficult to know the true locations of thresholds in real research situations. At the same time, it seems very reasonable to postulate that the locations of thresholds are different from person to person or from situation to situation. Therefore, an implication of the simulation results is that Pearson's correlation coefficient between categorized variables will decrease more if we consider a variety of data acquisition procedures and variety of threshold locations.

6 Conclusion

This study ran a large-scale simulation regarding biases of r when using categorized variables, carefully manipulating threshold locations. The results have shown that more serious biases of r occurred when thresholds are ill-organized. The findings suggest that previous simulation studies may have underestimated biases of r, and users of Likert-scales in social science should take the biases caused by categorized variables more seriously. Otherwise, biased values of r would result in incorrect interpretations of obtained data.

One of the possible ways to cope with the biases is the use of polychoric correlation. Estimation procedures of polychoric correlation were proposed by Olsson (1979) using maximum likelihood procedures and by Shiina et al. (2018) using the EM algorithm, although the use of polychoric correlation is not common in psychology.

There may be some limitations in this study. First, we have paid attention only to Pearson's r, not to other kinds of correlations (such as polychoric correlation or Spearman's rank correlation). Therefore, further studies are needed to examine the extent of the biases of different types of correlations. Second, our simulation has not completely examined the possible arrangement of thresholds. Although we set upper and lower limits $[-1, 1]$ for generating a set of thresholds, other upper and lower limits should also be considered. Such considerations will provide insights into the nature of the bias.

Funding This work was supported by JSPS KAKENHI Grant Numbers: 16H02050 and 18K03048.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Bollen KA, Barb KH (1981) Pearson's *r* and coarsely categorized measures. Am Sociol Rev 46:232–239
David HA, Nagaraja HN (2003) Order statistics, 3rd edn. Wiley, New Jersey
Martin WS (1973) The effects of scaling on the correlation coefficient: a test of validity. J Mark Res 10(3):316–318
Martin WS (1978) Effects of scaling on the correlation coefficient: additional considerations. J Mark Res 15(2):304–308
Morrison DG (1972) Regressions with discrete dependent variables: the effect on R2. J Mark Res 9(3):338–340
Olsson U (1979) Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika 44(4):443–460
Pearson K (1913) On the measurement of the influence of "Broad Categories" on correlation. Biometrika 9:116–139
Pearson K, Pearson ES (1922) On polychoric coefficients of correlation. Biometrika 14:127–156
Ritchie-Scott A (1918) The correlation coefficient of a polychoric table. Biometrika 12:93–133
Shiina K, Ouchi Y, Kubo S, Ueda T (2012) A dreadful secret of Pearson's *r*. The 76th annual convention of Japanese Psychological Association. https://psych.or.jp/meeting/proceedings/76/contents/pdf/1EVA14.pdf. Accessed 13 Feb 2019
Shiina K, Ueda T, Kubo S (2018) Polychoric correlations for ordered categories using the EM algorithm. In: Wiberg M et al (eds) Quantitative psychology. Springer Proceedings in Mathematics & Statistics, pp 247–259
Torgerson WS (1958) Theory and methods of scaling. Wiley, New York