Non-normal Distributions Commonly Used in Health, Education, and Social Sciences: A Systematic Review

Roser Bono[1,2]*, María J. Blanca[3], Jaume Arnau[1] and Juana Gómez-Benito[1,2]

[1] Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Barcelona, Spain, [2] Institute of Neurosciences, University of Barcelona, Barcelona, Spain, [3] Psychobiology and Behavioral Sciences Methodology, Faculty of Psychology, University of Málaga, Málaga, Spain

Statistical analysis is crucial for research and the choice of analytical technique should take into account the specific distribution of data. Although the data obtained from health, educational, and social sciences research are often not normally distributed, there are very few studies detailing which distributions are most likely to represent data in these disciplines. The aim of this systematic review was to determine the frequency of appearance of the most common non-normal distributions in the health, educational, and social sciences. The search was carried out in the Web of Science database, from which we retrieved the abstracts of papers published between 2010 and 2015. The selection was made on the basis of the title and the abstract, and was performed independently by two reviewers. The inter-rater reliability for article selection was high (Cohen's kappa = 0.84), and agreement regarding the type of distribution reached 96.5%. A total of 262 abstracts were included in the final review. The distribution of the response variable was reported in 231 of these abstracts, while in the remaining 31 it was merely stated that the distribution was non-normal. In terms of their frequency of appearance, the most-common non-normal distributions can be ranked in descending order as follows: gamma, negative binomial, multinomial, binomial, lognormal, and exponential. In addition to identifying the distributions most commonly used in empirical studies these results will help researchers to decide which distributions should be included in simulation studies examining statistical procedures.

Keywords: non-normal distributions, gamma distribution, negative binomial distribution, multinomial distribution, binomial distribution, lognormal distribution, exponential distribution, systematic review

INTRODUCTION

The data obtained in many fields of health, education, and the social sciences yield values of skewness and kurtosis that clearly deviate from those of the normal distribution (Micceri, 1989; Lei and Lomax, 2005; Bauer and Sterba, 2011; Blanca et al., 2013). In his imaginatively titled article 'The Unicorn, The Normal Curve, and Other Improbable Creatures,' Micceri (1989) concluded that real data commonly follow non-normal distributions. His analysis of the distributional characteristics of over 440 large-sample achievement and psychometric measures revealed several classes of deviation from the normal distribution, with the highest percentage corresponding to extreme deviation. In a more recent study, Blanca et al. (2013) analyzed the shape of 693 distributions from real psychological data by examining the values of the third and fourth central moments

as a measurement of skewness and kurtosis in small samples. They found that most distributions were non-normal; considering skewness and kurtosis jointly the results indicated that only 5.5% of the distributions were close to expected values under normality. Overall, 74.4% of distributions presented either slight or moderate deviation, while 20% showed more extreme deviation.

Variables with skewed distributions are also commonly used in a variety of psychological and social research. Arnau et al. (2014) listed some of these variables: reaction times or response latency in cognitive studies (Ulrich and Miller, 1993; Van der Linden, 2006; Shang-Wen and Ming-Hua, 2010), survival data from clinical trials (Qazi et al., 2007), clinical assessment indexes in drug abuse research (Deluchi and Bostrom, 2004), physical and verbal violence in couples (Szinovacz and Egley, 1995; Soler et al., 2000), divorced parents' satisfaction with co-parenting relationships in family studies (McKenry et al., 1999), and labor income (Diaz-Serrano, 2005) or health care costs (Zhou et al., 2009) in sociological studies. More recent examples involving non-normal data include neuropsychological data (Donnell et al., 2011; Oosthuizen and Phipps, 2012), data about paranoid ideation (Bebbington et al., 2013), fatigue symptoms of breast cancer patients (Ho et al., 2014), data on violence or sexual aggression (Swartout et al., 2015), and numerous studies on the cost of health care, such as costs among patients with depression or anxiety (Halpern et al., 2013; Vasiliadis et al., 2013), costs following brief cognitive behavioral treatment for insomnia (McCrae et al., 2014), and costs of anorexia nervosa (Stuhldreher et al., 2015), among others. Campitelli et al. (2016) also showed how the gamma distribution fits reaction times better than other well-studied distributions.

Although there is a wide variety of probability distributions, the most frequently used distributions involving real data are much fewer in number. The set of exponential distributions is very common in disciplines associated with the health and social sciences. The exponential family includes the normal, exponential, gamma, beta, and lognormal as continuous distributions, and the binomial, multinomial, and negative binomial as discrete distributions. The lognormal distribution, for example, is frequently found in medicine, social sciences, and economics (Limpert et al., 2001).

The normal distribution is the most well-known distribution and the most frequently used in statistical theory and applications. In fact, normality is one of the underlying assumptions of parametric statistical analysis. In practice, however, data can be drawn from other types of distribution, and in order to obtain accurate results researchers have to decide which statistical technique is best suited to the specific distribution of data. Monte Carlo simulation studies are commonly used to identify the robustness of statistical techniques under violation of underlying assumptions. In relation to continuous distributions, numerous simulation studies have analyzed the lognormal distribution (Algina and Keselman, 1998; Keselman et al., 2000; Kowalchuk et al., 2004; Arnau et al., 2012; Oberfeld and Franque, 2013; Bono et al., 2016, among others), and also the exponential distribution (Lix et al., 2003; Arnau et al., 2012). Among discrete distributions, simulation studies

have been conducted with binomial (Wu and Wu, 2007; Fang and Louchin, 2013) and multinomial distributions (Kuo-Chin, 2010; Bauer and Sterba, 2011; Jiang and Oleson, 2011). If the results of simulation studies are to be truly useful they need to include the distributions most commonly used in empirical contexts. However, there are very few studies detailing which distributions are most likely to represent data in different disciplines.

The aim of the present study was to determine the frequency of appearance of the most common non-normal distributions used in the health, educational, and social sciences. To this end, we conducted a systematic review of papers published between 2010 and 2015, coding two variables: shape of the distribution and field of study.

METHODS

Selection of Studies for Inclusion in the Review

The search was carried out in the Web of Science (WOS) database and used the following terms: '*nonnormal distribution*' OR '*non-normal distribution*' OR '*nonnormal data*' OR '*non-normal data*' OR '*ordinal data*' OR '*categorical data*' OR '*multinomial data*' OR '*binary data*' OR '*binomial data*' OR '*gamma distribution*' OR '*beta distribution*' OR '*lognormal distribution*' OR '*log-normal distribution*' OR '*log normal distribution*' OR '*exponential distribution*' OR '*binary distribution*' OR '*binomial distribution*' OR '*multinomial distribution*' OR '*nonnormal distributions*' OR '*non-normal distributions*' OR '*gamma distributions*' OR '*beta distributions*' OR '*lognormal distributions*' OR '*log-normal distributions*' OR '*log normal distributions*' OR '*exponential distributions*' OR '*binary distributions*' OR '*binomial distributions*' OR '*multinomial distributions*.' The use of these terms was agreed by two reviewers (first and third author), such that the search strategy employed general descriptors of non-normal distributions, descriptors for ordinal or categorical data, and specific descriptors of the most common non-normal distributions. The term '*negative binomial distribution*' was not included as it was encapsulated by the term '*binomial distribution.*' No restriction on the language of publication was made. The terms included were refined to the following WOS research areas: *Psychology*, *Health Care Sciences Services*, *Education and Educational Research*, *Social Sciences Other Topics*, *Psychiatry*, *Social Issues*, *Behavioral Sciences*, and *Biomedical Social Sciences*.

The selection of studies, based on title and abstract, was performed independently by two reviewers (first and second author). The following kinds of study were excluded from the review: theoretical studies of a statistical test, new procedures, mathematical development, comparison of models, simulation studies, tutorials, reviews of other authors' work, comments on other articles, systematic reviews, meta-analyses, studies about the teaching/learning of distributions, software, and studies carried out in areas other than health, education, or social sciences. We also excluded conference abstracts and proceedings, and book reviews. Any articles that did not specify the type of distribution or which referred to the normal distribution

were likewise excluded. The inter-rater reliability for selection of articles was assessed with Cohen's kappa (Cohen, 1968). The weighted kappa was 0.84, which can be interpreted as almost perfect agreement (Landis and Koch, 1977). Disagreements were resolved by discussion.

Data Extraction

Information about the type of distribution and the field of study was extracted from the content of the abstract and title of the included articles. In the event that more than one distribution was mentioned in an abstract, they were all

recorded. Data were extracted independently by two reviewers (first and second author). The inter-rater reliability regarding the type of distribution was 96.5%. Discrepancies were resolved by consensus after reviewing again the abstracts in question; in the event that a consensus could not be reached, the final decision was taken by a third reviewer (fourth author).

RESULTS

Of the 984 articles that were initially retrieved we eliminated, in stage 1, three duplicate records, three articles from

FIGURE 1 | Flow chart of the study selection process.
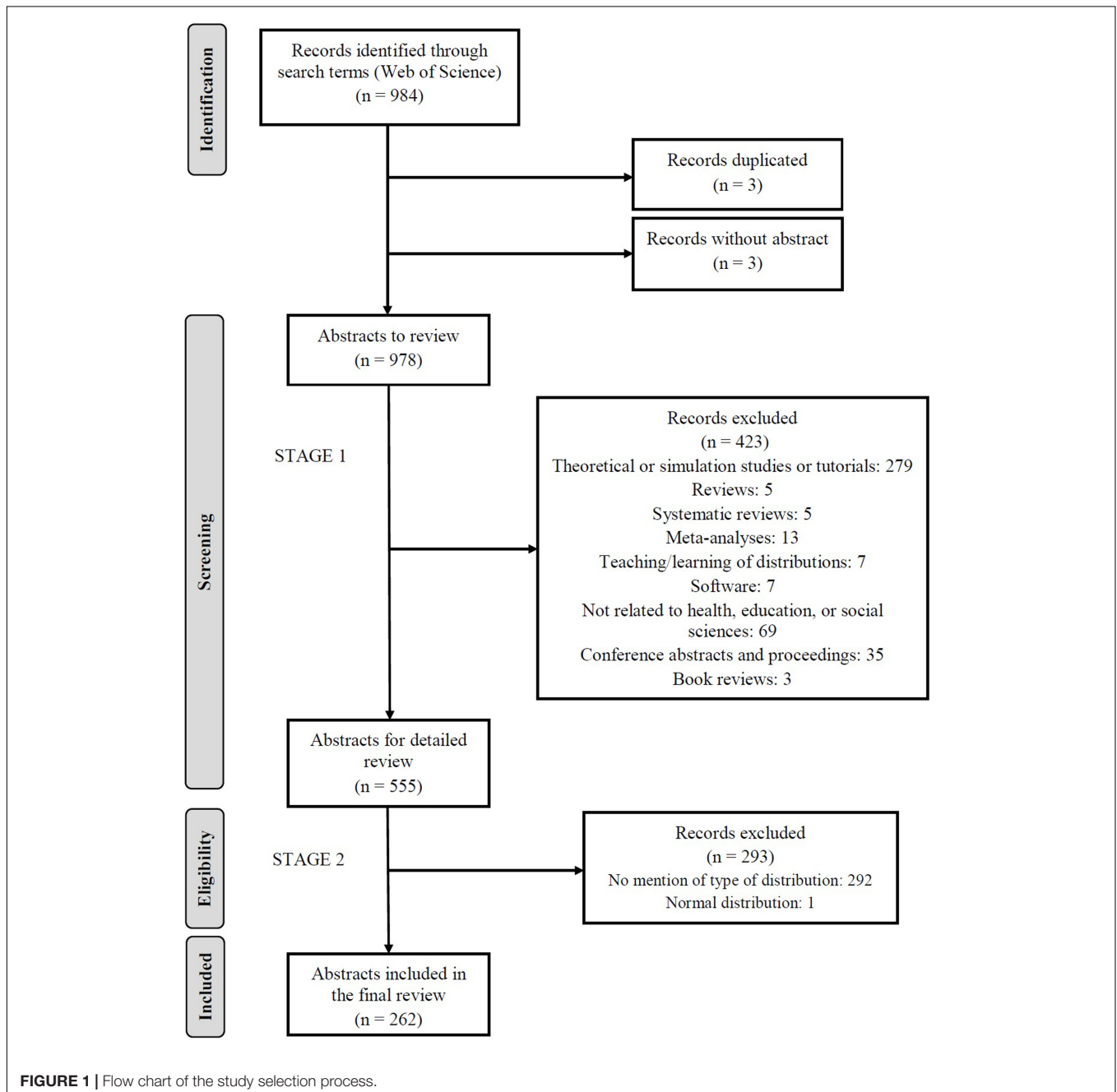
journals without abstracts, and 423 articles according to the abovementioned exclusion criteria (see Selection of Studies for Inclusion in the Review). In stage 2 we eliminated a further 292 abstracts that made no mention of the type of distribution and one which referred to a normal distribution. **Figure 1** summarizes the numbers of records identified and the reasons for exclusion at each stage. It can be seen that of the 984 records retrieved from the WOS, 262 were included in the review (148 from the area of health, 18 from education, and 96 from the social sciences). Seventeen abstracts referred to two distributions, all of which were counted, and therefore a total of 279 distributions were considered.

Across the reviewed studies the most common distributions were gamma ($n = 57$), negative binomial ($n = 51$), multinomial ($n = 36$), binomial ($n = 33$), lognormal ($n = 29$), and exponential ($n = 20$). The beta distribution fitted to very few data sets ($n = 5$). Other distributions identified but which had not been considered as search terms were the Poisson ($n = 12$), Weibull ($n = 2$), Pareto ($n = 1$), Lomax ($n = 1$), and exGaussian ($n = 1$). In addition to these distributions, 31 abstracts only indicated that the distribution was non-normal. **Figure 2** shows the percentage of the different types of distribution across the articles included in the review.

DISCUSSION

The aim of this systematic review of papers published between 2010 and 2015 was to determine the frequency of appearance of the most common non-normal distributions used in the health, educational, and social sciences. The results show that the most frequent distributions are the gamma and the negative binomial, followed by the multinomial, the binomial, the lognormal, and the exponential. The multinomial and binomial distributions show a good fit to data derived from discrete measurement scales, whereas the gamma and negative binomial distributions fit to variables related to health costs or income in social research. These findings extend those obtained by Micceri (1989) and Blanca et al. (2013), who analyzed the

FIGURE 2 | Percentage of the different distributions across the articles retrieved from the Web of Science (WOS) database for the period 2010–2015.

distributional characteristics of real data and noted that non-normal distributions are commonly found when working with psychological variables and psychometric measures. Knowing which distributions are most common is important because the type of distribution is a key aspect to consider when choosing an analytical technique.

When researchers know that the distribution which fits their data is non-normal, they should consider using alternatives to classical procedures. One way of modeling the response variable in order to find the type of distribution that best represents the data is to apply what are known as generalized additive models for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005). This method provides the foundations for further analyses (Campitelli et al., 2016). Other data analysis procedures include robust statistical methods (Wilcox, 2012), generalized linear models (McCullagh and Nelder, 1989) and their extension to mixed models (Stroup, 2013), and linear quantile mixed models (Geraci and Bottai, 2014).

As regards the limitations of this study the search was limited to a specific set of distributions, those considered to be the most common, and it is possible that the type of distributions identified by the review was biased somewhat by the search terms used. However, with the descriptors used we located the most well-known distributions from the exponential family. In order to access the full range of distributions, including the less common ones, we would have had to have applied the search term ‘*distribution,*’ which would have yielded many more types of distribution with a low or very low percentage across studies. One distribution of the exponential family that is of interest but which is not analyzed in our study is the Poisson distribution. This distribution was not included in the systematic review because it is more directly associated with count data and, in such cases, the negative binomial distribution can be used as an alternative (Faddy and Smith, 2011; Smith and Faddy, 2016).

Another limitation is that it is difficult to know whether the data are actually distributed as identified in the title and/or abstract. That is, researchers may have simply assumed particular non-normal distributions based on histograms or frequency distributions, or on a prior decision to apply a particular statistical technique or software. Empirical studies do not always indicate the distribution shape, or the procedure used to identify which distribution fits the data, and neither is a rationale usually given for why a particular non-normal distribution was used. Ideally, studies would report this kind of information so that other researchers from the same applied field have clear knowledge about the distributional properties of the variables under study.

Finally, and as noted in the introduction, the known distributions most widely used in simulation studies are the lognormal and the exponential, although discrete distributions such as the binomial and the multinomial have also been analyzed. In light of the results of this systematic review, future simulation studies examining the robustness and power of different statistical tests should also use the gamma and negative binomial distributions, the two most common forms according to our review. This is important because simulation studies need to include the distributions used in real-world data. Thus, we
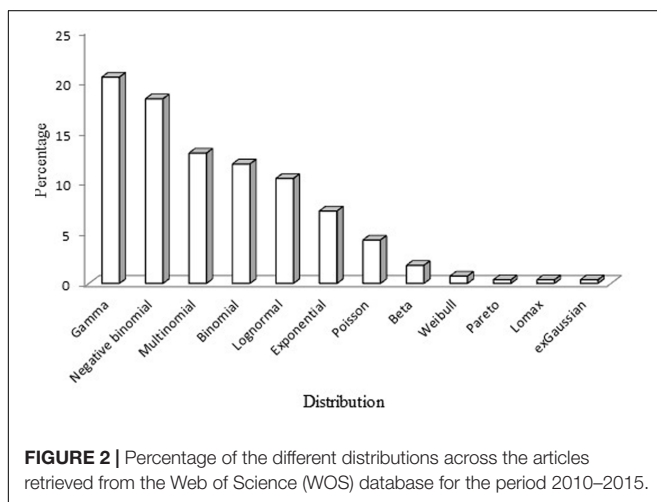
suggest that researchers who conduct Monte Carlo studies should generate data according to the distributions that are most relevant to the empirical reality of different disciplines.

AUTHOR CONTRIBUTIONS

RB was responsible for planning and executing the research activity and for drafting the manuscript, was involved in selecting the search terms to be used in the systematic review, acted as the second reviewer of the systematic review, and wrote the final version of the manuscript. MB was the first reviewer of the systematic review and offered a review of the manuscript's content. JA was involved in selecting the search terms to be used in the systematic review and supervised the drafting of the

manuscript. JG-B supervised the methods of systematic review and the final version of the manuscript, and acted as the third reviewer in the event that the first two reviewers could not reach an agreement regarding the type of distribution. All authors agree to be accountable for the content of the work, and have approved the final version to be published.

FUNDING

REFERENCES

Algina, J., and Keselman, H. J. (1998). A power comparison of the Welch-James and improved general approximation test in the split-plot design. *J. Educ. Behav. Stat.* 23, 152–159. doi: 10.3102/10769986023002152

Arnau, J., Bendayan, R., Blanca, M. J., and Bono, R. (2014). Should we rely on the Kenward–Roger approximation when using linear mixed models if the groups have different distributions? *Br. J. Math. Stat. Psychol.* 67, 408–429. doi: 10.1111/bmsp.12026

Arnau, J., Bono, R., Blanca, M. J., and Bendayan, R. (2012). Using the linear mixed model to analyze non-normal data distributions in longitudinal designs. *Behav. Res. Methods* 44, 1224–1238. doi: 10.3758/s13428-012-0196-y

Bauer, D. J., and Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: performance of alternatives specifications and methods of estimation. *Psychol. Methods* 16, 373–390. doi: 10.1037/a0025813

Bebbington, P. E., McBride, O., Steel, C., Kuipers, E., Radovanovic, M., Brugha, T., et al. (2013). The structure of paranoia in the general population. *Br. J. Psychiatry* 202, 419–427. doi: 10.1192/bjp.bp.112.119032

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., and Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology* 9, 78–84. doi: 10.1027/1614-2241/a000057

Bono, R., Arnau, J., Blanca, M. J., and Alarcón, R. (2016). Sphericity estimation bias for repeated measures designs in simulation studies. *Behav. Res. Methods* 48, 1621–1630. doi: 10.3758/s13428-015-0673-1

Campitelli, G., Macbeth, G., Ospina, R., and Marmolejo-Ramos, F. (2016). Three strategies for the critical use of statistical methods in psychological research. *Educ. Psychol. Meas.* 1–15. doi: 10.1177/0013164416668234

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* 70, 213–220. doi: 10.1037/h0026256

Deluchi, K. L., and Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: working with many zero values. *Am. J. Psychiatry* 161, 1159–1168. doi: 10.1176/appi.ajp.161.7.1159

Diaz-Serrano, L. (2005). Labor income uncertainty, skewness and homeownership: a panel data study for Germany and Spain. *J. Urban Econ.* 58, 156–176. doi: 10.1016/j.jue.2005.03.003

Donnell, A. J., Belanger, H. G., and Vanderploeg, R. D. (2011). Implications of psychometric measurement for neuropsychological interpretation. *Clin. Neuropsychol.* 25, 1097–1118. doi: 10.1080/13854046.2011.599819

Faddy, M. J., and Smith, D. M. (2011). Analysis of count data with covariate dependence in both mean and variance. *J. Appl. Stat.* 38, 2683–2694. doi: 10.1080/02664763.2011.567250

Fang, L., and Louchin, T. M. (2013). Analyzing binomial data in a split-plot design: classical approach or modern techniques? *Commun. Stat. Simul. Comput.* 42, 727–740. doi: 10.1080/03610918.2011.650264

Geraci, M., and Bottai, M. (2014). Linear quantile mixed models. *Stat. Comput.* 24, 461–479. doi: 10.1007/s11222-013-9381-9

Halpern, R., Nadkarni, A., Kalsekar, I., Nguyen, H., Song, R., Baker, R. A., et al. (2013). Medical costs and hospitalizations among patients with depression treated with adjunctive atypical antipsychotic therapy: an analysis of health insurance claims data. *Ann. Pharmacother.* 47, 933–945. doi: 10.1345/aph.1R622

Ho, R. T. H., Fong, T. C. T., and Cheung, I. K. M. (2014). Cancer-related fatigue in breast cancer patients: factor mixture models with continuous non-normal distributions. *Qual. Life Res.* 23, 2909–2916. doi: 10.1007/s11136-014-0731-7

Jiang, D., and Oleson, J. J. (2011). Simulation study of power and sample size for repeated measures with multinomial outcomes: an application to sound direction identification experiments (SIDIE). *Stat. Med.* 30, 2451–2466. doi: 10.1002/sim.4302

Keselman, H. J., Kowalchuk, R. K., and Boik, R. J. (2000). An examination of the robustness of the empirical Bayes and other approaches for testing main and interaction effects in repeated measures designs. *Br. J. Math. Stat. Psychol.* 53, 51–67. doi: 10.1348/000711000159178

Kowalchuk, R. K., Keselman, H. J., Algina, J., and Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educ. Psychol. Meas.* 64, 224–242. doi: 10.1177/0013164403260196

Kuo-Chin, L. (2010). Goodness-of-fit tests for modeling longitudinal ordinal data. *Comput. Stat. Data Anal.* 54, 1872–1880. doi: 10.1016/j.csda.2010.02.013

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lei, M., and Lomax, R. G. (2005). The effect of varying degrees on nonnormality in structural equation modeling. *Struct. Equ. Model.* 12, 1–27. doi: 10.1207/s15328007sem1201_1

Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal distributions across the sciences: keys and clues. *BioScience* 51, 341–352. doi: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Lix, L. M., Algina, J., and Keselman, H. J. (2003). Analyzing multivariate repeated measures designs: a comparison of two approximate degrees of freedom procedures. *Multivar. Behav. Res.* 38, 403–431. doi: 10.1207/s15327906mbr3804_1

McCrae, C. S., Bramoweth, A. D., Williams, J., Roth, A., and Mosti, C. (2014). Impact of brief cognitive behavioral treatment for insomnia on health care utilization and costs. *J. Clin. Sleep Med.* 10, 127–135. doi: 10.5664/jcsm.3436

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Edn. London: Chapman & Hall. doi: 10.1007/978-1-4899-3242-6

McKenry, P. C., Clark, K. A., and Stone, G. (1999). Evaluation of a parent education program for divorcing parents. *Fam. Relat.* 48, 129–137. doi: 10.2307/585076

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* 105, 156–166. doi: 10.1037/0033-2909.105.1.156

Oberfeld, D., and Franque, T. (2013). Evaluating the robustness of repeated measures analyses: the case of small sample sizes and nonnormal data. *Behav. Res. Methods* 45, 792–812. doi: 10.3758/s13428-012-0281-2

Oosthuizen, M. D., and Phipps, W. D. (2012). A preliminary standardisation of the Bohnen et al. version of the Stroop Color-Word Test for Setswana-speaking university students. *South Afr. J. Psychol.* 42, 411–422. doi: 10.1177/008124631204200313

Qazi, S., DuMez, D., and Uckun, F. M. (2007). Meta analysis of advanced cancer survival data using lognormal parametric fitting: a statistical method to identify effective treatment protocols. *Curr. Pharm. Des.* 13, 1533–1544. doi: 10.2174/138161207780765882

Rigby, R. A., and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Appl. Stat.* 54, 507–554. doi: 10.1111/j.1467-9876.2005.00510.x

Shang-Wen, Y., and Ming-Hua, H. (2010). Estimation of air traffic longitudinal conflict probability based on the reaction time of controllers. *Saf. Sci.* 48, 926–930. doi: 10.1016/j.ssci.2010.03.016

Smith, D. M., and Faddy, M. J. (2016). Mean and variance modeling of under- and overdispersed count data. *J. Stat. Softw.* 69, 1–23. doi: 10.18637/jss.v069.i06

Soler, H., Vinayak, P., and Quadagno, D. (2000). Biosocial aspects of domestic violence. *Psychoneuroendocrinology* 25, 721–739. doi: 10.1016/S0306-4530(00)00022-6

Stroup, W. W. (2013). *Generalized Linear Mixed Models. Modern Concepts, Methods and Applications.* Boca Raton, FL: Taylor and Francis.

Stuhldreher, N., Wild, B., Koenig, H. H., Konnopka, A., Zipfel, S., and Herzog, W. (2015). Determinants of direct and indirect costs in anorexia nervosa. *Int. J. Eat. Disord.* 48, 139–146. doi: 10.1002/eat.22274

Swartout, K. M., Thompson, M. P., Koss, M. P., and Su, N. (2015). What is the best way to analyze less frequent forms of violence? The case of sexual aggression. *Psychol. Violence* 5, 305–313. doi: 10.1037/a0038316

Szinovacz, M. E., and Egley, L. C. (1995). Comparing one-partner and couple data on sensitivity marital behaviors: the case of marital violence. *J. Marriage Fam.* 57, 995–1010. doi: 10.2307/353418

Ulrich, R., and Miller, J. (1993). Information processing models generating lognormally distributed reaction times. *J. Math. Psychol.* 37, 513–525. doi: 10.1006/jmps.1993.1032

Van der Linden, W. J. (2006). A log-normal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

Vasiliadis, H. M., Dionne, P. A., Preville, M., Gentil, L., Berbiche, D., and Latimer, E. (2013). The excess healthcare costs associated with depression and anxiety in elderly living in the community. *Am. J. Geriatr. Psychiatry* 21, 536–548. doi: 10.1016/j.jagp.2012.12.016

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd Edn. San Diego, CA: Academic Press.

Wu, K., and Wu, L. (2007). Generalized linear mixed models with informative dropouts and missing covariates. *Metrika* 66, 1–18. doi: 10.1007/s00184-006-0083-6

Zhou, X., Lin, H., and Johnson, E. (2009). Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *J. R. Stat. Soc. B Stat. Methodol.* 70, 1029–1047. doi: 10.1111/j.1467-9868.2008.00669.x