Behavior Research Methods (2021) 53:264–277
https://doi.org/10.3758/s13428-020-01407-2

t-Test and ANOVA for data with ceiling and/or floor effects

Qimin Liu[1] · Lijuan Wang[2]

Published online: 15 July 2020

Abstract

Ceiling and floor effects are often observed in social and behavioral science. The current study examines ceiling/floor effects in the context of the *t*-test and ANOVA, two frequently used statistical methods in experimental studies. Our literature review indicated that most researchers treated ceiling or floor data as if these data were true values, and that some researchers used statistical methods such as discarding ceiling or floor data in conducting the *t*-test and ANOVA. The current study evaluates the performance of these conventional methods for *t*-test and ANOVA with ceiling or floor data. Our evaluation also includes censored regression with regard to its capacity for handling ceiling/floor data. Furthermore, we propose an easy-to-use method that handles ceiling or floor data in *t*-tests and ANOVA by using properties of truncated normal distributions. Simulation studies were conducted to compare the performance of the methods in handling ceiling or floor data for *t*-test and ANOVA. Overall, the proposed method showed greater accuracy in effect size estimation and better-controlled Type I error rates over other evaluated methods. We developed an easy-to-use software package and web applications to help researchers implement the proposed method. Recommendations and future directions are discussed.

Keywords Ceiling effect · Floor effect · *t*-Test · ANOVA

Introduction

According to the definitions used in Uttl (2005) and Wang, Zhang, McArdle, and Salthouse (2008), ceiling or floor effects occur when the tests are relatively easy or difficult to the extent that substantial proportions of individuals obtain either the maximum or minimum score. As such, the true extent of their abilities cannot be determined.

Ceiling or floor effects have been observed in various areas of psychology and education research. In developmental psychology, experimental tasks that are too difficult for younger participants can cause a floor effect (Timeo, Farroni, & Maass, 2017). Similarly, performance tasks can be too easy, resulting in a ceiling effect (e.g., Ulber, Hamann, & Tomasello, 2016). This can also occur in educational settings where the performance measure is an educational test (e.g., Dompnier et al., 2015; Fantuzzo, Gadsden, & McDermott, 2011). In clinical research, ceiling and/or floor effects can occur when examining severely symptomatic populations. For example, ceiling effects were observed in symptom measures, and floor effects occurred in resiliency and/or positive affect measures (Muthen, 1990; Priebe et al., 2013). In cognitive psychology, Uttl (2005) provided extensive examples of ceiling effects in widely used memory assessments, with ceiling proportions ranging from at least 25% for 9- to 15-item verbal list learning tasks to more than 50% for the verbal paired-associates learning task.

Ceiling and floor data are censored data: Censoring is a condition in which the values of measurements or observations are only partially known. For example, the only known information about ceiling data is that the true levels are at or above the ceiling threshold. The exact levels are unknown due to ceiling effects. Ceiling or floor effects can be confused with other statistical terms. Two notable examples are the presence of performance asymptotes and of semicontinuous variables. Ceiling effects are different from performance asymptotes (Miller, 1956): The asymptotic

Lijuan Wang is grateful for the support from NIH 1R01HD091235, NIH 1R01HD087319, and NIH 1R01HD088482.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-020-01407-2) contains supplementary material, which is available to authorized users.

✉ Qimin Liu
qimin.liu@vanderbilt.edu

✉ Lijuan Wang
lwang4@nd.edu

[1] Department of Psychology and Human Development, Vanderbilt University, Nashville, TN 37203, USA

[2] Department of Psychology, University of Notre Dame, 390 Debartolo Hall, Notre Dame, IN 46556, USA

values are the largest true values that individuals can demonstrate, whereas ceiling effects imply that the observed scores are lower than the true levels that individuals can demonstrate. A variable with ceiling effects is also different from a semicontinuous variable that combines a continuous distribution with point masses at one or more locations (Olsen & Schafer, 2001): Values in a semicontinuous variable (e.g., alcohol use with many zeros) are all valid values, whereas the ceiling threshold (the maximum observed score) is a proxy for some larger true values (see Wang et al., 2008 for a detailed discussion).

Methodological discussion and development with regard to handling ceiling/floor effects, though sparse, have occurred. Uttl (2005) demonstrated the attenuation in reliability and validity when ceiling effects are present using empirical data. Jennings and Cribbie (2016) also noted that ceiling effects result in weakened reliability and validity. Tobin (1958) proposed the Tobit model to deal with limited-range responses in regression, which uses the likelihood of the censored distribution for parameter estimation and hypothesis testing. Wang et al. (2008) extended the model into the Tobit growth model for longitudinal data analysis with ceiling/floor data using Bayesian estimation, which has been applied in longitudinal studies (e.g., Piccinin et al., 2013). In addition, with regard to confirmatory factor analysis with censored data, Muthen (1990) proposed a method to adjust the correlation matrix using properties of doubly truncated bivariate normal distributions. Schweizer (2016) proposed a method to tackle the variance reduction problem due to ceiling effects in confirmatory factor analysis by multiplying a weight matrix onto the sample covariance matrix. To our knowledge, however, the impact of ceiling/floor effects on the *t*-test and ANOVA and how to statistically deal with ceiling/floor data in this context lack systematic evaluation and discussion. The *t*-test and ANOVA are two of the most commonly used statistical methods in behavioral and social sciences, especially in experimental studies. The high cost of experimental studies thus warrants our current investigation.

To investigate how psychological and educational researchers have statistically handled ceiling/floor data in *t*-tests or ANOVA, a brief literature review was conducted. PsychINFO returned 397 English articles published within a five-year span that mentioned “ceiling effects” or “floor effects,” illustrating the presence of ceiling and floor effects in the literature. Among the articles, we focused on reviewing those that were published in journals with higher impact factors (i.e., five-year impact factor > 2). As examples, we reviewed articles from the *Journal of Experimental Psychology, Psychological Science, American Educational Research Journal, and Child Development.*

After excluding papers on methodology and literature review, 96 substantive articles were reviewed. Thirty-three (34%) of the articles conducted *t*-tests and 50 (53%) conducted ANOVA. Nineteen (57%) of the articles using *t*-tests and 35 (70%) of those using ANOVA treated the ceiling/floor values as if they were true values. That is, researchers completely ignored ceiling/floor effects and simply used the observed scores in the statistical data analysis. Researchers in this case often mentioned ceiling/floor effects only in the discussion section as a plausible explanation for the lack of significant results. Of those who treated the ceiling/floor values as if they were true values, seven articles using the *t*-test and five articles using ANOVA reported the proportions of ceiling/floor data or performed a normality test to evaluate the severity of ceiling/floor effects (e.g., Coman & Berry, 2015; Kim, Peters, & Shams, 2012), whereas the other articles did not report the proportions. Some researchers—nine (27%) in studies using the *t*-test and ten (20%) in studies using ANOVA—attempted to tackle ceiling/floor effects by adjusting the experimental procedures. This was often done by excluding measures that were observed with ceiling/floor effects in their pilot studies (e.g., Chiu & Egner, 2015). Other researchers—four (12%) and five (12%) of the studies using *t*-test and ANOVA, respectively—attempted to statistically handle ceiling/floor effects by simply discarding the ceiling/floor data. One article (3%) where the *t*-test was used employed a modified log-transformation to handle floor and ceiling effects (Sokol-Hessner et al., 2015). Despite the prevalence in the psychological, educational, and behavioral research literature, ceiling and floor effects seem to have rarely been well addressed statistically in the context of the *t*-test and ANOVA.

The current study aims to systematically and quantitatively examine the impact of ceiling/floor effects on *t*-tests and ANOVA and compare different methods for handling these effects. In the remainder of the paper, we first discuss conventional methods and propose an easy-to-use method for handling ceiling/floor effects in *t*-tests and ANOVA. Next, we show the impact of ceiling/floor data on the *t*-test and ANOVA when conventional methods are used and compare the performance of different handling methods with simulated data. Lastly, we provide a real data example to illustrate the application of the proposed method and compare the results from different methods. We conclude the paper with recommendations and future research directions.

Methods for handling ceiling/floor effects in *t*-tests and ANOVA

As discussed earlier in the paper, conventional methods for handling ceiling/floor effects in *t*-tests and ANOVA include treating ceiling/floor data as true values and discarding ceiling/floor data. The former leaves data as they are—censored. The latter results in truncated data. In this section, we first review the *t*-test and ANOVA. We then hypothesize on the impact of two conventional methods, introduce the censored regression model for handling ceiling and floor data, and propose an easy-to-use method that utilizes the properties of truncated normal distributions for the *t*-test and ANOVA with ceiling/floor effects.

A review of the *t*-test and ANOVA

Given the scope of this paper, we focus on the two-independent-samples *t*-test (referred to simply as "*t*-test" in this paper) and one-way ANOVA. The *t*-test examines the difference between two independent population means. Denote M_1 and M_2 , s_1^2 and s_2^2, and n_1 and n_2 as the sample means, sample variances, and sample sizes of two groups. Welch's *t* statistic can be computed using the following formula:

$$t=\frac{M_2-M_1}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \tag{1}$$

This is a function of the first and second moment estimates of the observed data in each group and the group sample sizes. The critical *t* value can be found from a *t* distribution with a desired alpha level (e.g., 0.05), and the degrees of freedom computed as follows:

$$df=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \tag{2}$$

We use Welch's *t* test statistic instead of the pooled two-sample *t* test statistic because the Welch's *t* test is more robust against violation of the homogeneity of variance (HOV) assumption (Delacre, Lakens, & Leys, 2017; Welch, 1947). Moreover, Cohen's *d*, an effect size measure for the mean difference between two groups, can be computed based on the following formula:

$$\widehat{d}\approx\frac{M_2-M_1}{\sqrt{\frac{n_2s_1^2+n_1s_2^2}{n_1+n_2}}}=t\times\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{3}$$

One-way ANOVA examines differences between means of two or more independent groups (Maxwell, Delaney, & Kelley, 2018). Similar to Welch's *t* test, Brown-Forsythe's *F** statistic (Brown & Forsythe, 1974), which is robust to violation of the HOV assumption, can be computed using the individual group variances as follows:

$$F^*=\frac{\sum_{j=1}^{k}n_j\left(M_j-M\right)^2}{\sum_{j=1}^{k}\left(1-\frac{n_j}{N}\right)s_j^2} \tag{4}$$

Here, N is the total sample size, k refers to the number of groups being compared, M_j and s_j^2 are the sample mean and variance of group j, respectively, and M is the grand sample mean. Like Welch's *t* statistic, *F** is a function of the first and second moment estimates of the observed data in each group and the group sample sizes. The critical *F* value can be approximated from an *F* distribution with a desired significance level, $df_{numerator}=k-1$, and $df_{denominator}=\frac{1}{\sum_{j=1}^{k}\frac{g_j^2}{n_j-1}}$ where $g_j=\frac{\left(1-\frac{n_j}{N}\right)s_j^2}{\sum_{j=1}^{k}\left(1-\frac{n_j}{N}\right)s_j^2}$. An effect size measure for the overall group mean differences is Cohen's f^2 (Maxwell et al., 2018) et al., 2018):

$$\widehat{f}^2=\frac{(k-1)F^*}{N} \tag{5}$$

When group variance and sample size are equal between groups, *F* and *F** are identical, where *F* is the regular *F* test statistic, calculated as the variance of M_j divided by the mean of within-group variances [1]. Subsequently, the effect size estimates from *F* and *F** would be identical given equal variance and equal sample size between groups. When HOV is violated, the effect size estimate from *F* would not be accurate because *F* assumes HOV, and the effect size estimate from *F** may be more suitable.

As shown above, the computation in *t*-test and ANOVA depends upon sample means and sample standard deviations. Therefore, biased mean and standard deviation estimates lead to biased test results and effect size estimates. This is shown below when conventional methods for handling ceiling/floor data are used (i.e., ceiling/floor data are treated as if they were true scores or are removed from a data analysis).

Method 1: Treat ceiling/floor data as if they were true scores

Some researchers ignore ceiling/floor effects and treat ceiling/floor data as if they were true scores in data analyses. The data utilized for analysis from this approach can be seen as censored data (e.g., Wang & Zhang, 2011). More specifically, ceiling effects result in a type of right-censored data, where true scores that are larger than the ceiling threshold *b* (i.e., the maximum score from a test) are not observed but are recorded as *b*. Floor effects lead to a type of left-censored data: True scores that are smaller than floor threshold *a* (i.e., the minimum score from a test) are not observed but are recorded as *a*. When both ceiling

[1] *F** can be systematically smaller than *F* given large samples with small variances where *F* is too liberal. *F** can be systematically larger than *F* given large samples with large variances where *F* is too conservative. Moreover, when cell sample sizes are equal, *F* and *F** are identical, but the denominator degrees of freedom are different. See Maxwell et al., (2018) for more details.

and floor effects occur, the data can be viewed as interval-censored, where true scores that are larger than b (the maximum score) or smaller than a (the minimum score) are not observed and are recorded as b and a, respectively. Let Y be a random variable of the true scores and Y^* be a random variable of the observed scores with floor/ceiling effects. We have

$$y^* = \begin{cases} a, & \text{if } y \leq a \\ y, & \text{if } a < y < b \\ b, & \text{if } y \geq b \end{cases} \tag{6}$$

The impact of censoring on mean and variance estimates has been studied previously (e.g., Cohen, 1959; Greene, 2002). Applying the findings to the context of ceiling/floor effects, for example, with the normality assumption and a finite floor threshold a, for $Y \sim \mathcal{N}(\mu, \sigma^2)$, we can obtain the expected value and variance of the observed data y^* as follows (Greene, 2002):

$$E\left[Y^*\right] = \Psi(a)a + (1 - \Psi(a))(\mu + \sigma\lambda) \tag{7}$$

$$Var\left[Y^*\right] = \sigma^2(1 - \Psi(a))\left[(1 - \delta) + (\alpha - \lambda)^2\Psi(a)\right] \tag{8}$$

$\Psi\left[\frac{(a-\mu)}{\sigma}\right] = \Psi(\alpha) = Prob(y \leq a)$. $\lambda = \frac{\phi(\alpha)}{1-\Psi(\alpha)}$ where ϕ is the standard normal density function. In addition, $\delta = \lambda^2 - \lambda\alpha$. The forms become complex for interval-censored Y^*. Numerically, using Eqs. (7) and (8), when the true mean $\mu = 0$, the true variance $\sigma^2 = 1$ and the proportion of floor data is 20%, the expected mean and the expected variance of observed data y^* are approximately .11 and .69, respectively, which deviate from the true values of $\mu = 0$ and $\sigma^2 = 1$, respectively.

More generally, treating ceiling/floor data as if they were true scores leads to attenuated variance estimates. When ceiling or floor effects exist, the mean of the observed data is expected to be smaller or larger than the true mean, respectively. When both ceiling and floor effects exist, the impact on the observed mean would depend on the proportion of ceiling and floor data. Note that the impact of ceiling/floor effects is "symmetrical" when Y has a symmetrical distribution (e.g., a normal distribution). For example, when $Y \sim \mathcal{N}(\mu, \sigma^2)$, $\mu = 0$ and $\sigma^2 = 1$, and the proportion of ceiling data is 20%, and the expected mean and variance of observed data are approximately −.11 and .69, respectively.

With an attenuated sample variance s^{*2} and a biased sample mean M^* from ceiling/floor data, test statistics based on Eqs. (1) and (4) for *t*-test and ANOVA may also be biased when treating ceiling/floor data as if they were true values. Subsequently, Cohen's d and Cohen's f^2 estimates based on Eqs. (3) and (5) may be biased when ceiling/floor data are treated as true values.

Method 2: Remove ceiling/floor data

Some researchers have handled their ceiling/floor data by removing the ceiling/floor data. The resulting data $y^{'}$ can be viewed as a kind of truncated data:

$$y^{'} = \begin{cases} removed, & \text{if } y \leq a \\ y, & \text{if } a < y < b \\ removed, & \text{if } y \geq b \end{cases} \tag{9}$$

That is, only scores between a and b, not including a and b, are kept for statistical data analyses.

The impact of truncation on the expected mean and variance of $y^{'}$ has been discussed in the literature (e.g., Aitkin, 1964). Applying the findings to the context of ceiling/floor effects, when ceiling/floor values are removed, the variance of $y^{'}$ is expected to be smaller than the true variance. For data with ceiling or floor effects, the deletion of ceiling or floor values would make the expected mean of $y^{'}$ smaller or larger than the true mean, respectively. Specifically, when $Y \sim \mathcal{N}(\mu, \sigma^2)$ and the ceiling and floor thresholds are b and a, respectively, we derive that Y' has a truncated normal distribution with the following mean and variance based on results from Aitkin (1964).

$$E\left(Y^{'}\right) = E(Y|a < Y < b) = \mu + \sigma\frac{\phi(\alpha) - \phi(\beta)}{\Psi(\beta) - \Psi(\alpha)} \tag{10}$$

$$Var\left(Y^{'}\right) = Var(Y|a < Y < b)$$
$$= \sigma^2\left[1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Psi(\beta) - \Psi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Psi(\beta) - \Psi(\alpha)}\right)^2\right] \tag{11}$$

where $\alpha = \frac{a-\mu}{\sigma}$ and $\beta = \frac{b-\mu}{\sigma}$. Numerically, for example, when the true mean $\mu = 0$, the true variance $\sigma^2 = 1$, and the ceiling proportion is 20%, the mean and variance of the truncated variable $Y^{'}$ are approximately −.35 and .58, respectively.

When ceiling/floor data are removed from data analyses, the sample mean and variance estimates can be biased. Therefore, we expect the test statistics and effect size estimates for *t*-test and ANOVA from Method 2 to be biased.

Method 3: Censored regression for *t*-test and ANOVA with ceiling/floor data

Censored regression has been proposed and demonstrated for regression with censored or limited-range outcomes (Tobin, 1958). In a censored regression model, the outcome variable, Y^*, as described in Eq. (6), is modeled with a censored distribution. The corresponding underlying true model is $y_i = x_i^T B + \epsilon_i$, where y_i is the true score of the dependent variable for person i, x_i^T is the design matrix, and B is a vector of regression coefficients.

Specifically, a censored regression model can be estimated by maximizing the following likelihood function based on a censored distribution, assuming $\epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ (e.g., Henningsen, 2011):

$$\log L = \sum_{i=1}^{N}\left[I_i^a \log\Psi\left(\frac{a - x_i^T B}{\sigma_\epsilon}\right) + I_i^b \log\Psi\left(\frac{x_i^T B - b}{\sigma_\epsilon}\right)\right.$$
$$\left. + \left(1 - I_i^a - I_i^b\right)\left(\log\phi\left(\frac{y_i^* - x_i^T B}{\sigma_\epsilon}\right) - \log\sigma_\epsilon\right] \qquad (12)$$

where $I_i^a = \begin{cases} 1, & \text{if } y_i^* = a \\ 0, & \text{if } y_i^* > a \end{cases}$ and $I_i^b = \begin{cases} 1, & \text{if } y_i^* = b \\ 0, & \text{if } y_i^* < b \end{cases}$. Standard nonlinear optimization algorithms can be used to maximize the log-likelihood function with respect to the parameter vector $(B^T, \sigma_\epsilon)^T$. Likelihood ratio tests can be used for significance testing.

Censored regression has long received attention among methodological researchers. With its capacity to handle censored data in regression, it has the potential to handle ceiling/floor effects in *t*-tests and ANOVA using reparameterization (Tobin, 1958). However, we noted in our literature review that censored regression has rarely been applied to *t*-test or ANOVA with ceiling/floor data in psychological research. To use censored regression for handling ceiling/floor effects, for the *t*-test, the design matrix contains a vector of 1s for the intercept and a vector of 0/1s (dummy coding of the group membership) for the group mean difference coefficient. For *k*-group one-way ANOVA, the formulation of the design matrix is the same except that $k - 1$ vectors of 0/1s are needed for dummy coding the *k* groups. Similarly, maximum likelihood can be used to estimate the regression coefficients. To implement censored regression, R package `censReg` (Henningsen, 2011) can be used to potentially handle ceiling/floor data in *t*-tests and ANOVA.

For a two-independent-samples *t*-test using `censReg`, the regression coefficient of the dummy-coded grouping variable and its inference can be used for comparing two group means. For ANOVA using `censReg`, an omnibus test statistic can be obtained by comparing a full model containing the dummy-coded group variables to the reduced intercept-only model. The difference in the two 2*\times log-likelihood values can be calculated and then compared with the critical χ^2 value with $k - 1$ degrees of freedom. It is worth noting that `censReg` does not provide effect size estimates for the *t*-test or ANOVA. For the *t*-test, we propose using the *t* value of the grouping variable coefficient for Eq. (3) to obtain a Cohen's *d* estimate. For ANOVA, we propose using coefficients in the full and reduced censored regression models to obtain $\widehat{\theta}_j$. Specifically, we have

$$\widehat{\theta}_j = \begin{cases} \widehat{B_0} - \widehat{B_0^+}, & \text{if } j = 1 \\ \widehat{B_0} + \widehat{B}_{j-1} - \widehat{B_0^+}, & \text{if } j > 1 \end{cases} \qquad (13)$$

Here, $\widehat{B_0}$ and $\widehat{B}_{j-1}$ are the intercept and group effect coefficient estimates from the full model. Specifically, $\widehat{B_0}$ and $\widehat{B_0}$ $+ \widehat{B}_{j-1}$ are the group mean estimates of the first group (reference group) and the j_{th} group, respectively. $\widehat{B_0^+}$ is the intercept estimate of the reduced model—the grand mean *M* estimate. Provided $\widehat{\theta}_j's$, the following formula (e.g., Maxwell et al., 2018) can be used to obtain a Cohen's f^2 estimate.

$$\widetilde{\widehat{f}}^2 = \frac{\sum \widehat{\theta}_j^2 / k}{\widehat{\sigma_\epsilon^2}}$$

Method 3 is expected to handle ceiling/floor effects well for the *t*-test and ANOVA when group variances are equal. We also expect that Method 3 can lead to more accurate testing for mean differences and more accurate effect size estimates than Methods 1 and 2. However, it is unclear how sensitive the method is to the HOV violation. This is evaluated in the simulation study.

Method 4: Our proposed approach

Using properties from truncated normal distributions, we propose an easy-to-use method for the *t*-test and ANOVA with ceiling/floor data. Using Eqs. (10) and (11), we derive the mean and variance estimates of true scores for each group with floor and ceiling thresholds *a* and *b*, under the normality assumption. Let $\widetilde{M}\ and\ \widetilde{s}^2$ denote the proposed sample mean and variance estimates of a group, respectively, that adjust for ceiling/floor effects. We have

$$\widetilde{s}^2 = \frac{s'^2}{1 + \dfrac{\widehat{\alpha}\phi\left(\widehat{\alpha}\right) - \widehat{\beta}\phi\left(\widehat{\beta}\right)}{\Psi\left(\widehat{\beta}\right) - \Psi\left(\widehat{\alpha}\right)} - \left(\dfrac{\phi\left(\widehat{\alpha}\right) - \phi\left(\widehat{\beta}\right)}{\Psi\left(\widehat{\beta}\right) - \Psi\left(\widehat{\alpha}\right)}\right)^2} \qquad (14)$$

$$\widetilde{M} = M' + \widetilde{s} \times \frac{\phi\left(\widehat{\beta}\right) - \phi\left(\widehat{\alpha}\right)}{\Psi\left(\widehat{\beta}\right) - \Psi\left(\widehat{\alpha}\right)} \qquad (15)$$

M' and s' are the sample mean and sample standard deviation of the truncated data after removing ceiling and floor data. Recall that $\alpha = \frac{a-\mu}{\sigma}$ and $\beta = \frac{b-\mu}{\sigma}$. Thus, α and β are the standardized floor and ceiling thresholds, respectively. In practice, μ and σ are unknown, and thus α and β need to be estimated. To estimate α and β for each group, we use the proportions of floor and ceiling values of each group. For *l* floor observations out of *n* total observations, $\widehat{\alpha} = \Psi^{-1}(l/n)$. For *r* ceiling observations out of *n* total observations, $\widehat{\beta} = \Psi^{-1}(1 - r/n)$. That is, the standardized floor and ceiling threshold estimates correspond to the *floor proportion* and *1-ceiling proportion* in the standardized normal cumulative distribution function. Thus, to obtain corrected mean and variance estimates using Eqs. (14) and (15), only information about summary statistics including the sample mean of truncated data, sample variance of truncated data, group sample size, and proportions of ceiling/floor data are required from each group. When raw data are available, Eqs. (14) and (15) can also be implemented through the function `*rec.mean.var (y*, floor, ceiling)*` from our *R* package `DACF` on CRAN (Liu & Wang, 2018). The input variable y^* represents a vector of n observations with ceiling/floor effects, and '*floor*' and '*ceiling*' respectively represent the ceiling and floor thresholds. Floor and ceiling percentages are estimated based on the proportions of values at the specified ceiling and floor thresholds, respectively. Then the function gives the following outputs: (1) the calculated ceiling percentage, (2) the calculated floor percentage, (3) the estimated mean after adjusting for ceiling/floor effects, and (4) the estimated variance after adjusting for ceiling/floor effects. Normality is assumed in the estimation.

Our proposed mean and variance estimates can be used in computing the t statistics, i.e., Eq. (1), and F^* statistics, i.e., Eq. (4), for the t-test and one-way ANOVA, respectively. Under the normality assumption, asymptotically, our method should produce accurate mean and variance estimates, because mean and variance estimates form sufficient statistics to describe normally distributed random variables. Asymptotically, we expect our method with corrected mean and variance estimates to yield accurate estimates for the t statistic and the F^* statistic. With improved estimates for the t statistic and F^* statistic, our method is expected to yield less biased results than Methods 1 and 2 for the effect size estimates (Cohen's d and f^2). As our method uses Welch's t test and Brown-Forsythe's F^* test for ANOVA, our method is expected to perform well when the homogeneity of variance assumption is violated for the t test or ANOVA.

The proposed method calculates the degrees of freedom based on the after-truncation sample sizes. The rationale was that the proposed method utilizes full information only from data points of $n - r - l$ participants and partial information from data points of $r + l$ participants of a group for the mean and variance estimation. Specifically, the corrected mean and variance estimates (Eqs. 14 and 15) are functions of mean and variance estimates using after-truncation data ($n - r - l$ participants) and the standardized floor and ceiling threshold estimates. The thresholds are estimated using the ceiling and floor percentage estimates based on data points of $n - r$ and $n - l$ participants, respectively. This is a relatively conservative approach for calculating the degrees of freedom, which can help control the type I error rate. This feature can be beneficial, especially given the "replication crisis" in psychological and behavioral research.

Simulation Study 1: *t*-Test with ceiling data

Our first simulation study was designed to evaluate the performance of the aforementioned methods for the two-independent-groups mean t-test with ceiling data. As discussed in Method 1 above, the results with ceiling data are generalizable to floor data when the true population distribution is symmetrical.

In this simulation study, four factors were manipulated: the population effect size ($d = 0$, .2, .5, .8, corresponding to the null, small, medium, and large effects), population standard deviation ratio between two groups ($SDR = 1$ and 1.5, corresponding to the scenarios where HOV is met and violated, respectively), sample size per group ($n = n_1 = n_2 = 25$, 50, 100, 200, 500), and ceiling proportion of the reference group (Group 1 $CP = 10\%$, 20%, 30%). In total, we have $4 \times 2 \times 4 \times 3 = 96$ conditions included in Simulation Study 1. Additional conditions that examine the impact of greater heterogeneity of variance ($SDR = 2$), unbalanced design ($n_2/n_1 = ½$ or 2 with $n_1 = 50$ or 200), simultaneous ceiling and floor effects (10% ceiling and 20% floor or 15% ceiling and 15% floor with $n = 50$ or 200), and non-normal distribution (lognormal outcomes) are included. The simulation study design, data generation methods, and simulation results of those additional conditions are included in the online supplemental materials, as the patterns are consistent with the results we present here. The number of replications for each condition was 1000.

We used the following evaluation criteria for evaluating the performance of the methods: (1) Accuracy of effect size estimation measured by bias (when the true effect size is null), $\overline{\widehat{d}}-d$, or relative bias (when the true effect size is non-null), $\frac{\overline{\widehat{d}}-d}{d}$. An estimator with its relative bias larger than 10% (non-ignorable bias) is considered less than desirable (Muthén & Muthén, 2002). (2) Type I error rate with a satisfactory range from 2.5% to 7.5% (Bradley, 1978). (3) Coverage probability of 95% confidence intervals containing the true population mean difference with a satisfactory range from 92.5% to 97.5%. The simulation study was conducted in *R*.

Data generation

Group 1 (reference group) true data (free of ceiling effects) were generated from the standard normal distribution, $Y_1 \sim \mathcal{N}(0, 1)$. With Eqs. (1) and (3), for a given Cohen's d and SDR, Group 2 (treatment group) true data were generated from $Y_2 \sim \mathcal{N}\left(d \times \sqrt{\frac{1+SDR^2}{2}}, SDR^2\right)$. Reference t test statistics and reference Cohen's d estimates were recorded using the generated true data.

We then introduced ceiling effects, with the ceiling threshold determined by the standardized inverse cumulative normal density function with 1-*CP*. For example, when Group 1 *CP* = 20% and 30%, the ceiling thresholds *b* are .842 and .524, respectively. The same ceiling threshold is used across the two groups. This aims to simulate a more realistic scenario: the same measure with the same limited range of scores is used in both the control and treatment groups. Accordingly, Group 2 may have a higher ceiling proportion than Group 1 (see Table 1 for the ceiling proportions of Group 2, ranging from 10% to 61%). For example, when *d* is positive, Group 2 has a larger population mean and thus Group 2 has a larger ceiling proportion than Group 1 in a given simulation condition. For another instance, when SDR is greater than 1, the ceiling proportion of Group 2 in a condition is larger than that of Group 1 because Group 2 scores are distributed more widely than Group 1 scores. The ceiling proportions are faithful to those observed in our empirical literature review, as mentioned in the introduction. Methods 1–4 (i.e., 1: treating ceiling data as if they were true values, 2: removing ceiling data, 3: using censored regression for handling ceiling effects, and 4: our proposed approach) were applied to analyze the data with ceiling effects.

Results

Results with *n*=50 or *n*=200 are summarized in Tables 2 (type I error rates and coverage rates) and 3 (average bias in Cohen's *d* estimates). For the conditions with the other sample sizes, the results shared a similar pattern and thus are included in the online supplemental document.

When HOV was met, the Type I error rates from Methods 1–3 were satisfactory under the studied conditions (see Table 2 when *d* = 0 and *SDR* =1). When HOV was violated, the type I error rates from Methods 1–3 were inflated (see Table 2 when *d* = 0 and *SDR* =1.5). For example, these can be as high as 33.3%, 93.3%, and 11.4% when the ceiling proportion of the reference group (*CP)* is 20% for Methods 1–3, respectively. The inflation was more severe with increased ceiling proportions or increased group sample size. As ceiling proportions increase, the biases in both mean and variance estimates increase, resulting in more severe inflation of type I error rates. As sample size increases, the biases in the estimates become more visible as the confidence interval widths become narrower. Among Methods 1–3, Method 2 (removing ceiling data) yielded the most inflated type I error rates, followed by Method 1 (treating ceiling data as if they were true values) and then Method 3 (censored regression). Our proposed method (Method 4) became slightly conservative when ceiling proportions increased. However, Method 4 was the only studied method that had a type I error rate ranging between .025 and .075 across most of the studied conditions (the only exception was the *t*-test with 30% ceiling data in the reference group when sample size per group was 25).

When HOV was met, the coverage rates from Methods 1 (treating ceiling data as if they were true values) and 2 (removing ceiling data) were not satisfactory (under-coverage) under most of the studied conditions (see Table 2 when *d* ≠0 and *SDR* =1). For example, these can be as low as 57.2% and 18.6% when the *CP* of Group 1 = 20% and *d* = .5 for Methods 1 and 2, respectively. The coverage rates deviated more from the nominal value 95% as the ceiling proportion increased. The deviations also increased when the HOV assumption was not met (see Table 2 when *d* ≠0 and *SDR* = 1.5). Between Methods 1 and 2, Method 2 performed worse in coverage rates across the studied conditions. Censored regression (Method 3) yielded satisfactory coverage rates when HOV was met. However, when HOV was violated, Method 3 had less than ideal coverage probabilities (see Table 2 when *d* ≠0 and *SDR* = 1.5). For example, these can be as low as

Table 1 Treatment group ceiling proportions per population effect size and *SDR* (*CP* is the ceiling proportion of the reference group [Group 1])

t-Test: Group 2 ceiling proportion														
	SDR = 1				*SDR* = 1.5									
	D = 0	*d* = 0.2	*d* = 0.5	*d* = 0.8	*d* = 0	*d* = 0.2	*d* = 0.5	*d* = 0.8						
CP = 0.1	10%	14%	22%	32%	20%	24%	30%	37%						
CP = 0.2	20%	26%	37%	48%	29%	33%	41%	49%						
CP = 0.3	30%	37%	49%	61%	36%	41%	49%	57%						
ANOVA: Ceiling proportions of Group 2 (G2) and Group 3 (G3)														
	SDR = 1								*SDR* = 1.5					
	$f^2 = 0$	$f^2 = 0.01$		$f^2 = 0.0625$		$f^2 = 0.16$			$f^2 = 0$	$f^2 = 0.01$		$f^2 = 0.0625$		$f^2 = 0.16$
	θ = 0	θ = .12	θ = −.12	θ = .31	θ = −.31	θ = .49	θ = −.49		θ = 0	θ = .13	θ = −.13	θ = .33	θ = −.33	θ = .53	θ = −.53
	G2,G3	G2	G3	G2	G3	G2	G3	G2,G3	G2	G3	G2	G3	G2	G3
CP = 0.1	10%	12%	8%	16%	6%	21%	4%	20%	22%	17%	26%	14%	31%	11%
CP = 0.2	20%	24%	17%	30%	13%	36%	9%	29%	32%	26%	37%	22%	42%	18%
CP = 0.3	30%	34%	26%	41%	20%	49%	16%	36%	40%	33%	45%	28%	50%	24%

Table 2 Type I error rates and coverage probabilities in *t*-test with ceiling data

CP (Group 1) =		0%	10%				20%				30%			
		Reference	1	2	3	4	1	2	3	4	1	2	3	4
SDR = 1, *n* = 50	d = 0	.053	.057	.050	.060	.035	.050	.049	.051	.032	.049	.049	.056	.031
	d = .2	.934	.937	.935	.927	.952	.939	**.924**	.931	.962	**.916**	**.909**	.936	.975
	d = .5	.944	**.912**	**.836**	.937	.964	**.848**	**.709**	.942	.967	**.695**	**.633**	.944	.968
	d = .8	.938	**.835**	**.554**	.940	.964	**.559**	**.320**	.941	.969	**.215**	**.222**	.949	.969
SDR = 1, *n* = 200	d = 0	.046	.043	.038	.042	.036	.044	.040	.043	.032	.038	.055	.040	.030
	d = .2	.947	.939	**.899**	.944	.961	**.916**	**.808**	.948	.963	**.850**	**.728**	.951	.974
	d = .5	.946	**.852**	**.467**	.951	.959	**.572**	**.186**	.954	.962	**.187**	**.077**	.950	.959
	d = .8	.940	**.531**	**.035**	.938	.957	**.043**	**.004**	.934	.952	**.001**	**.003**	.939	.941
SDR = 1.5, *n* = 50	d = 0	.046	.071	**.313**	.065	.028	**.120**	**.391**	**.064**	.025	**.170**	**.438**	**.092**	.025
	d = .2	.943	**.880**	**.503**	.934	.954	**.787**	**.380**	**.917**	.959	**.655**	**.314**	**.898**	.962
	d = .5	.952	**.727**	**.191**	.934	.969	**.418**	**.080**	**.916**	.975	**.140**	**.048**	**.897**	.975
	d = .8	.954	**.379**	**.023**	.931	.970	**.050**	**.006**	**.892**	.969	**.003**	**.005**	**.836**	.957
SDR = 1.5, *n* = 200	d = 0	.050	**.167**	**.845**	.072	.033	**.333**	**.933**	**.114**	.026	**.495**	**.959**	**.175**	.025
	d = .2	.954	**.639**	**.021**	.927	.969	**.316**	**.002**	**.857**	.973	**.111**	**.001**	**.773**	**.976**
	d = .5	.946	**.196**	**.000**	**.880**	.962	**.009**	**.000**	**.781**	.963	**.000**	**.000**	**.654**	.961
	d = .8	.954	**.002**	**.000**	**.817**	.968	**.000**	**.000**	**.680**	.962	**.000**	**.000**	**.512**	.940

Note 1: 1 = Method 1 (treating ceiling data as if they were true values); 2 = Method 2 (removing ceiling data); 3 = Method 3 (censored regression); 4 = Method 4 (our proposed method)

Note 2: When $d = 0$, the statistic is the empirical Type I error rate. Otherwise, it is the coverage rate

Note 3: Type I error rates that are outside the 2.5–7.5% range and coverage rates that are outside the 92.5–97.5% range are bolded

78.1% when the *CP* of the reference group is 20%, $d = .5$, and $SDR = 1.5$. Our proposed method (Method 4) yielded good coverage rates across almost all the studied conditions (see Table 2).

In terms of the average bias or average relative bias in Cohen's d estimates (Table 3), overall, Method 2 had the most biased estimates, followed by Methods 1 and 3. The relative biases from Method 2 were above 10% in most of the studied conditions. For example, for the conditions with $d = .2$, $n = 200$, and *CP* of the reference group = 10%, the relative biases in Cohen's d estimates from Method 2 were −18% and −185.0% when HOV was met ($SDR = 1$) and violated ($SDR = 1.5$), respectively. When HOV was met and the *CP* of the reference group ≤ 20%, the effect size estimates from Methods 1 and 3 were acceptable (e.g., the highest relative bias was −7.5% and −6.3% from Methods 1 and 3, respectively, when $d = .8$). However, the violation of HOV can lead to biased effect size estimate from Methods 1 and 3. For example, the relative biases were as high as −60.0% and −25.0% from Methods 1 and 3, respectively, when $d = .2$, $n = 200$, $SDR = 1.5$, and the *CP* of the reference group was as low as 10%. As the ceiling proportion increased, the effect size estimates from Methods 1–3 became more biased. Our proposed method (Method 4) yielded the most accurate effect size estimates across all the studied methods under all the studied conditions. Furthermore, the relative bias from Method 4 was all under 10%.

Simulation Study 2: ANOVA with ceiling data

Our second simulation study evaluated the performance of the methods for three-group ANOVA with ceiling data. In this study, four factors were manipulated: population effect size ($f^2 = 0, .01, .0625, .16$, corresponding to null, small, medium, large effect sizes), population standard deviation ratio between the group with a positive treatment effect and the other groups ($SDR = 1$ and 1.5, representing the scenarios where HOV is met and violated, respectively), sample size per group ($n = n_1 = n_2 = n_3 = 25, 50, 100, 200, 500$), and ceiling proportion of the reference group (*CP* = 10%, 20%, 30%). Table 1 shows the ceiling proportions for the treatment groups at different population effect sizes. The ceiling proportions ranged from 4% to 50%. In total, we had 96 conditions in Simulation Study 2, and the number of replications for each condition was 1000. Similar to Simulation Study 1, we include additional conditions in the supplemental materials to investigate the impact of greater heterogeneity of variance (SDR = 2), unbalanced design ($n_1 = n_2 \times 2$ or $n_1 = n_2/2$ with $n_1 = 50$ or 200), simultaneous ceiling and floor effects (10% ceiling and 20% floor

Table 3 Average bias in Cohen's *d* estimates in *t*-test with ceiling data

CP (Group 1) =		0%	10%				20%				30%			
		Reference	1	2	3	4	1	2	3	4	1	2	3	4
SDR = 1, *n* = 50	d = 0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	d = .2	5.0	0.0	**−15.0**	5.0	5.0	0.0	**−25.0**	0.0	5.0	−5.0	**−30.0**	0.0	5.0
	d = .5	0.0	−2.0	**−18.0**	0.0	2.0	−4.0	**−24.0**	−2.0	0.0	−8.0	**−30.0**	−6.0	2.0
	d = .8	0.0	−2.5	**−20.0**	−2.5	1.3	−7.5	**−28.8**	−6.3	1.3	**−12.5**	**−32.5**	**−10.0**	1.3
SDR = 1, *n* = 200	d = 0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	d = .2	0.0	0.0	**−15.0**	0.0	0.0	−5.0	**−25.0**	−5.0	0.0	−10.0	**−35.0**	−5.0	0.0
	d = .5	2.0	−2.0	**−18.0**	0.0	2.0	−6.0	**−28.0**	−2.0	2.0	−10.0	**−34.0**	−6.0	2.0
	d = .8	0.0	−2.5	**−21.3**	−2.5	0.0	−7.5	**−28.8**	−6.3	0.0	**−13.8**	**−35.0**	**−11.3**	0.0
SDR = 1.5, *n* = 50	d = 0	0.0	−0.1	−0.3	−0.1	0.0	−0.2	−0.4	−0.1	0.0	−0.2	−0.5	−0.1	0.0
	d = .2	5.0	**−55.0**	**−190.0**	**−20.0**	5.0	**−85.0**	**−235.0**	**−45.0**	5.0	**−110.0**	**−265.0**	**−60.0**	5.0
	d = .5	2.0	**−28.0**	**−96.0**	**−12.0**	4.0	**−42.0**	**−118.0**	**−22.0**	4.0	**−56.0**	**−132.0**	**−32.0**	4.0
	d = .8	1.3	**−20.0**	**−73.8**	−10.0	6.2	**−31.3**	**−88.8**	**−18.8**	7.5	**−41.3**	**−98.8**	**−27.5**	6.2
SDR = 1.5, *n* = 200	d = 0	0.0	−0.1	−0.3	0.0	0.0	−0.2	−0.4	−0.1	0.0	−0.2	−0.5	−0.1	0.0
	d = .2	5.0	**−60.0**	**−185.0**	**−25.0**	5.0	**−90.0**	**−230.0**	**−45.0**	5.0	**−115.0**	**−260.0**	**−60.0**	5.0
	d = .5	0.0	**−28.0**	**−96.0**	**−14.0**	4.0	**−44.0**	**−118.0**	**−24.0**	4.0	**−56.0**	**−134.0**	**−34.0**	4.0
	d = .8	1.3	**−20.0**	**−73.8**	**−10.0**	6.2	**−31.3**	**−90.0**	**−20.0**	6.2	**−41.3**	**−100.0**	**−27.5**	7.5

Note 1: 1 = Method 1 (treating ceiling data as if they were true values); 2 = Method 2 (removing ceiling data); 3 = Method 3 (censored regression); 4 = Method 4 (our proposed method)

Note 2: Relative biases larger than 10% are bolded

Note 3: Absolute bias is reported for the null effect condition; percentage relative bias is reported otherwise

or 15% ceiling and 15% floor with *n* =50 or 200), and non-normal distribution (lognormal outcomes).

We used the following evaluation criteria for evaluating the performance of the methods: (1) accuracy of effect size estimation measured by bias (when the true effect size is null) or relative bias (when the true effect size is non-null), and (2) type I error rates. The simulation study was conducted in *R*.

Data generation

We first generated the true data that were free of ceiling effects. We generated Group 1 (the reference group, i.e., $\theta_1 = 0$) true data from the standard normal distribution, $\mathcal{N}(0, 1)$, θ_j represents the deviation of group *j* mean from the grand mean. For convenience, we set Groups 2 and 3 to have a positive and negative treatment effect of equal magnitude, i.e., $\theta_2 = -\theta_3$. In addition, we set Group 2 standard deviation based on the *SDR* value and fixed Group 3 standard deviation to 1. Given a Cohen's f^2 and *SDR*, Group 2 and 3 true data were generated accordingly.

Reference *F** test statistics and reference Cohen's f^2 estimates were recorded using the generated true data. We then introduced ceiling effects to the data using a similar procedure as that described in Simulation Study 1. Methods 1–4 were applied to analyze the data with ceiling effects.

Results

Tables 4 (for type I error rates) and 5 (for average bias or average relative bias in effect size f^2 estimates) summarize the simulation results with *n* = 50 and *n* = 200. As the results showed a similar pattern for the conditions with the other sample sizes, those results are included in the online supplemental document. For the additional conditions in the supplemental materials, the patterns that emerged are consistent with the conditions presented here.

Type I error rates from Methods 1–3 were satisfactory under the studied conditions when HOV was met (see Table 4 when *SDR* = 1). When HOV was violated, inflated type I error rates from Methods 1–3 were observed (see Table 4 when *SDR* = 1.5). For example, when the *CP* of the reference group was 20%, the type I error rates were as high as 36%, 96%, and 13% for Methods 1–3, respectively. The inflation was more severe at higher ceiling proportions. Similar to the *t*-test results, among Methods 1–3, Method 2 (removing ceiling data) yielded the most inflated type I error rates, followed by Method 1 (treating ceiling data as if they were true values) and then Method 3 (censored regression). Method 4 was the only studied method with a type I error rate ranging between .025 and .075 across most of the studied conditions (the only exception was that a type I error rate of 18% was observed when *n* = 25, *SDR* = 1, and *CP* = 30%)

Table 4 Type I error rates in ANOVA with ceiling data

CP (Group 1) =	0%	10%				20%				30%			
	Reference	1	2	3	4	1	2	3	4	1	2	3	4
$SDR = 1$, $n = 50$	.05	.05	.05	.05	.04	.05	.05	.05	.03	.05	.05	.05	.03
$SDR = 1$, $n = 200$	.04	.05	.04	.05	.03	.05	.05	.04	.03	.05	.05	.05	.03
$SDR = 1.5$, $n = 50$	.05	.07	.32	.05	.04	**.11**	**.40**	.06	.03	**.16**	**.46**	**.09**	.03
$SDR = 1.5$, $n = 200$	.06	**.18**	**.88**	.07	.03	**.36**	**.96**	**.13**	.03	**.56**	**.97**	**.21**	.03

Note 1: 1 = Method 1 (treating ceiling data as if they were true values); 2 = Method 2 (removing ceiling data); 3 = Method 3 (censored regression); 4 = Method 4 (our proposed method)

Note 2: Type I error rates that are outside the 2.5–7.5% range are bolded

Similar to the findings from Simulation Study 1, f^2 estimates were least accurate with Method 2 (removing ceiling data; see Table 5). For example, for the conditions with $f^2 = .0625$, $n = 200$, and *CP* of the reference group = 10%, the relative biases in f^2 estimates from Method 2 were −27.3% and −70.8% when HOV was met ($SDR = 1$) and violated ($SDR = 1.5$), respectively. When the CP of the reference group < 30% and HOV was met ($SDR = 1$), f^2 estimates from Methods 1 and 3 were acceptable. However, when the CP of the reference group = 30%, biases in f^2 estimates from Method 1 became non-ignorable (e.g., the relative bias was −12.1% when $f^2 = 0.0625$ and $n = 200$). When HOV was violated, Methods 2 and 3 produced biased effect size estimates. For example, when $n = 200$, *CP* of the reference group = 20%, and $f^2 = 0.0625$, the relative biases from Methods 2 and 3 were −44.6% and −12.3%, respectively. Our proposed method (Method 4) yielded the most accurate effect size estimates across all the studied methods under all the studied ANOVA conditions. Moreover, the relative bias from Method 4 was under 10% for most of the studied conditions except when $CP = 30\%$, $f^2 = .01$, $n = 50$, and $SDR = 1$, where the relative bias from Method 4 was 16.7%.

Table 5 Average bias in Cohen's f^2 estimates in ANOVA with ceiling data

CP (Group 1) =		0%	10%				20%				30%			
		Reference	1	2	3	4	1	2	3	4	1	2	3	4
SDR = 1, $n = 50$	$f^2 = 0$	0.014	0.014	0.016	0.014	0.014	0.014	0.017	0.015	0.015	0.014	0.02	0.015	0.017
	$f^2 = .01$	0.024	0.0	−4.2	4.2	4.2	0.0	0.0	4.2	8.3	−4.2	4.2	8.3	**16.7**
	$f^2 = .0625$	0.079	−2.5	**−22.8**	0.0	0.0	−6.3	**−30.4**	1.3	0.0	**−11.4**	**−34.2**	1.3	1.3
	$f^2 = .16$	0.173	−1.7	**−23.1**	0.6	0.6	−6.4	**−33.5**	1.2	0.0	**−11.6**	**−39.9**	1.7	0.6
SDR = 1, $n = 200$	$f^2 = 0$	0.003	0.003	0.004	0.003	0.003	0.003	0.004	0.003	0.004	0.003	0.005	0.004	0.004
	$f^2 = .01$	0.013	0.0	**−15.4**	0.0	0.0	−7.7	**−23.1**	0.0	0.0	−7.7	**−23.1**	0.0	0.0
	$f^2 = .0625$	0.066	−3.0	**−27.3**	0.0	0.0	−7.6	**−37.9**	0.0	0.0	**−12.1**	**−45.5**	0.0	0.0
	$f^2 = .16$	0.165	−2.4	**−27.3**	0.0	−1.2	−7.3	**−38.8**	0.0	−1.2	**−13.3**	**−46.7**	0.0	−1.2
SDR = 1.5, $n = 50$	$f^2 = 0$	0.013	0.016	0.04	0.016	0.014	0.019	0.054	0.018	0.015	0.022	0.067	0.021	0.017
	$f^2 = .01$	0.024	**−25.0**	4.2	0.0	0.0	**−29.2**	**50.0**	−4.2	4.2	**−29.2**	**95.8**	−8.3	8.3
	$f^2 = .0625$	0.079	**−26.6**	**−60.8**	1.3	−5.1	**−38.0**	**−59.5**	−7.6	−5.1	**−46.8**	**−55.7**	**−16.5**	−3.8
	$f^2 = .16$	0.174	**−17.2**	**−58.0**	6.3	−4.6	**−27.0**	**−65.5**	−1.7	−5.2	**−35.6**	**−68.4**	−8.6	−4.6
SDR = 1.5, $n = 200$	$f^2 = 0$	0.003	0.006	0.028	0.005	0.003	0.009	0.042	0.006	0.004	0.013	0.054	0.008	0.004
	$f^2 = .01$	0.014	**−42.9**	0.0	**−14.3**	0.0	**−50.0**	**71.4**	**−28.6**	0.0	**−50.0**	**128.6**	**−35.7**	0.0
	$f^2 = .0625$	0.065	**−30.8**	**−70.8**	0.0	−4.6	**−44.6**	**−70.8**	**−12.3**	−4.6	**−53.8**	**−67.7**	**−21.5**	−3.1
	$f^2 = .16$	0.164	**−19.5**	**−63.4**	4.9	−5.5	**−29.9**	**−72.0**	−3.7	−6.1	**−39.0**	**−76.8**	**−11.0**	−5.5

Note 1: 1 = Method 1 (treating ceiling data as if they were true values); 2 = Method 2 (removing ceiling data); 3 = Method 3 (censored regression); 4 = Method 4 (our proposed method)

Note 2: Effect size estimates with their relative biases larger than 10% are bolded. Relative biases were computed using the reference values

Note 3: Absolute bias is reported for the null effect condition; percentage relative bias is reported otherwise

Illustration with an empirical data analysis

To illustrate the methods, a subset of the real data from Salthouse (2004) and Wang et al. (2008) were used. Wechsler Memory Scale III Word List subsets were administered to participants (*N* = 608) aged 19 to 97 in three sessions. In each session, the task was for participants to recall 12 unrelated words that were presented immediately before the task. The procedure was performed four times using the same words in the same order. For our purposes, we only used the trial 4 data from the first session. Additionally, the sample was divided into three age groups to examine cross-sectional age differences in memory: younger adult group aged 18–39 (*n* = 135); middle-aged adult group aged 40–59 (*n* = 236); and older adult group aged 60–97 (*n* = 237).

Table 6 displays the ceiling proportions and group mean and standard deviation estimates for the three age groups using Methods 1, 2, and 4. The younger adult group had a higher ceiling proportion, followed by the middle-aged adult and older adult groups. The younger adult group had higher mean estimates than the other two groups. The older adult group had greater variance estimates than the other two groups. An *F* test was conducted to compare the sample variance of the middle age group to that of the older adult group. The results revealed that the variance of the middle age group was significantly smaller than that of the older adult group (95% confidence interval estimate of the variance ratio was [.48, .88]). Thus, the HOV assumption was violated in the current example. All four methods were applied to compare the means of the middle-aged adult group and the older adult group with a *t* test, and to compare the three group means with ANOVA. The main results of the analysis are shown in Table 7.

In *t*-tests, all methods except for Method 2 (removing ceiling data) yielded statistically significant results. Middle-aged adults had significantly different average scores from older adults. Method 2 resulted in a considerably smaller effect size estimate. Although results from Method 1 (treating data as if they were true values), Method 3 (censored regression), and our proposed method (Method 4) agreed in statistical significance, the widths of the confidence interval estimates differed. Our proposed method is implemented in our *R* package \`DACF\`. For the *t* test, \`lw.t.test(x1, x2, floor, ceiling)\` takes in \`x1\` and \`x2\`, vectors of group 1 and group 2 data, respectively. In addition, 'floor' and 'ceiling' represent the floor and ceiling thresholds, respectively, such as the minimum and maximum scores of the measurement scale. For example, here we used \`lw.t.test(mid, old, 0, 12)\`, where 'mid' and 'old' contain the data vectors of scores observed for middle-aged and older adults, respectively.

For ANOVA, Brown-Forsythe *F** tests were conducted. All methods reported a statistically significant mean difference among the three groups. Treating ceiling data as if they were true values produced the largest *F** value, whereas removing ceiling data resulted in the smallest *F** value. For censored regression, because a likelihood ratio test (chi-square test) was conducted to compare the group means, deviance is shown in place of the *F** statistic for censored regression in Table 7. Removing ceiling data produced the smallest effect size estimates, whereas censored regression and treating data as if they were true values resulted in effect size estimates that were close to those from our proposed method. Based on the simulation results under the HOV violation scenarios, our proposed method (Method 4) is recommended for both the *t*-test and ANOVA. For ANOVA, \`lw.f.star (data, formula, floor, ceiling)\` takes in a data frame of a column for the observed dependent variable scores and a column for the levels of the grouping factor. Here, 'formula' represents the modeling relationship, e.g., scores ~ age. Again, a user needs to specify the ceiling and floor thresholds. Here we used \`lw.f.star(dat, scores~age, 0, 12)\`, where 'dat' is a dataframe with one variable named 'scores' containing the observed scores from all groups, and another variable named 'age' containing the categorized age information (i.e., 1, 2, 3, representing younger-, middle-, and older-aged adult groups, respectively) for the respective participant.

In the implementation of our proposed methods, both functions output test statistics (*t* value and *F** value, respectively), *p* values, and effect size estimates (Cohen's *d* and f^2 estimates, respectively). In addition, our *t* test function outputs 95% confidence interval estimates for the group mean differences. To help researchers more easily use the proposed approach, we

Table 6 Descriptive statistics and group mean and standard deviation estimates of the empirical example

Age Group	*n*	Ceiling proportion	Mean			Standard deviation		
			1	2	4	1	2	4
18–39	135	43.7%	10.89	10.03	11.43	1.36	1.26	2.00
40–69	236	31.4%	10.33	9.73	10.66	1.47	1.25	1.79
70–97	237	15.6%	9.46	8.99	9.62	1.96	1.77	2.23

Note 1: 1 = Method 1 (ceiling data treated as if they were true values); 2 = Method 2 (ceiling data were removed); 4 = Method 4 (our proposed estimation method)

Table 7 *t*-Test and ANOVA results of the empirical example

		1	2	3	4
T-Test	t	−8.27	.48	−7.71	−6.49
	CI	(−1.77, −1.09)	(−0.29, .48)	(−2.48, −1.47)	(−2.36, −1.25)
	p	.00	.63	.00	.00
	$\widehat{d}$	−.89	.05	−.83	−.83
ANOVA	F^*	41.94	28.34	*70.33*	38.86
	p	.00	.00	.00	.00
	$\widehat{f}^2$	.14	.09	.14	.13

Note 1: 1 = Method 1 (treating ceiling data as if they were true values); 2 = Method 2 (removing ceiling data); 3 = Method 3 (censored regression); 4 = Method 4 (our proposed method)

Note 2: The *italicized* statistic is the deviance computed from 'censReg' outputs

developed an R Shiny application, which can be accessed at https://qmliu.shinyapps.io/DACFE/.

Discussion

Ceiling/floor effects can have negative impact on *t*-test and ANOVA when inappropriate statistical methods are used. As demonstrated in our simulation studies, the test results and effect size estimates of the *t*-test and ANOVA are often misleading when ceiling/floor data are treated as if they were true values or when they are removed from statistical analyses. Thus, it is important for researchers to attend to ceiling/floor effects in their statistical data analyses. The *t*-test and ANOVA, the two most widely used statistical techniques, are not among the exceptions.

To handle ceiling/floor effects in *t*-test and ANOVA, we introduced more appropriate methods including censored regression and the proposed method for normally distributed continuous outcomes. Our simulation results showed that censored regression provided less misleading test results and more accurate effect size estimates than the conventional methods. However, under HOV violation, the performance of censored regression for handling ceiling/floor effects was less than satisfactory. With greater HOV violation, censored regression yielded worse results. This is because standard censored regression was designed with the HOV assumption. Having an unbalanced design can exacerbate the impact of HOV violation. Overall, we found that Methods 1–3 (treated as if they were true values; removed from data analyses; censored regression) yielded worse performance under unbalanced designs than balanced design, and/or under greater HOV violation than less HOV violation. Future research can investigate approaches for modifying the regular censored regression model to relax the HOV assumption (e.g., allowing heterogeneous residual variances across groups).

Our proposed method, in comparison, is robust to the HOV violation regardless of design balance, owing to the use of the unpooled sample variances. In addition, mean and variance estimates form sufficient statistics to describe a normally distributed random variable. Under the normality assumption, asymptotically, our method with the corrected mean and variance estimates yields accurate estimates for the *t* statistic and the F^* statistic and effect size estimates (Cohen's d and f^2). One potential concern with our method is that the corrected test statistics comprise the corrected sample moments, but there is uncertainty in the moment estimates. However, as evidenced by the satisfactory coverage rates and the well-controlled type I error rates, the standard error estimates from the proposed statistics did not find this to be an issue. Thus, uncertainty in the moment estimates was appropriately quantified by our method. Based on the simulation results with finite samples, our proposed method generally handled ceiling/floor effects better than the conventional methods (treating ceiling/floor data as if they were true values or removing ceiling/floor data) for the balanced and unbalanced designs. Furthermore, our proposed method performed better than or as well as censored regression. In particular, overall, our method (Method 4) had better-controlled type I error rates than all the other studied methods across different conditions. While our proposed methods demonstrated satisfactory type I error rates and coverage rates across a wide range of simulated conditions, future studies should develop further mathematical proofs regarding the null distributions and the test statistics given our proposed corrections in the sample moments to enhance the generalizability.

Both censored regression and our proposed method are not without assumptions. A common major assumption is that true scores are assumed to be normally distributed. The likelihood function of censored regression is based upon the normal distribution density function. In our proposed methods, group means and variances are estimated using the properties of truncated normal distributions. Thus, violation of normality in true scores may lead to misleading results from censored regression and from our proposed method. This assumption is vital: in our simulation with lognormal data, both censored regression and our proposed approach yielded suboptimal performance. Future research can extend our proposed method to handle ceiling and floor effects while relaxing the normality assumption.

Ceiling/floor effects can be prevented or dealt with in earlier stages of research prior to statistical data analyses. In the experimental design phase, when a researcher selects an ability/attitude instrument, the researcher should consult existing literature to investigate whether ceiling/floor effects could occur. It may be beneficial for the researcher to avoid using or revising an instrument that is likely to produce ceiling/floor effects. This is because ceiling/floor effects by their nature lead to loss of information in the observed data: true scores that are above the maximum or minimum thresholds are observed at the thresholds. Thus, when a researcher has to use an instrument that is subject to ceiling/floor effects, the researcher should plan for a larger sample size. It is worth noting that in some cases, ceiling/floor observations may be informative to the researcher. For example, when the researcher wishes to evaluate a new invention for improving math ability, the changes in the ceiling/floor proportions before and after the invention may be informative in some context. During the data analysis phase, we strongly recommend that researchers report the proportions of ceiling/floor data whenever relevant. For *t*-tests and ANOVA, we also recommend our proposed method (Method 4) for analyzing data with ceiling/floor effects.

In summary, ceiling/floor effects can lead to biased results in tests of mean differences when improper statistical methods, such as treating ceiling/floor values as true values or removing ceiling/floor values, are used. Thus, we introduced and proposed more appropriate methods: censored regression and our proposed method. Via simulation studies, we found that our proposed method was robust against the HOV violation and often yielded more accurate and valid *t*-test and ANOVA results for data with ceiling/floor effects. We hope that our R Shiny app will make it easy for researchers to apply the proposed method for handling ceiling/floor effects in *t*--tests and ANOVA.

References

Aitkin, M. A. (1964). Correlation in a singly truncated bivariate normal distribution. *Psychometrika*, *29*(3), 263–270. https://doi.org/10.1007/BF02289723

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, *69*(346), 364. https://doi.org/10.2307/2285659

Chiu, Y.-C., & Egner, T. (2015). Inhibition-Induced Forgetting: When More Control Leads to Less Memory . *Psychological Science* , *26*(1), 27–38. https://doi.org/10.1177/0956797614553945

Cohen, A. C. J. (1959). Simplified estimators for the normal distribution when samples are single censored or truncated. *Technometrics*, *1*(3), 217–237. https://doi.org/10.2307/1266442

Coman, A., & Berry, J. N. (2015). Infectious Cognition: Risk Perception Affects Socially Shared Retrieval-Induced Forgetting of Medical Information . *Psychological Science* , *26*(12), 1965–1971. https://doi.org/10.1177/0956797615609438

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Dompnier, B., Darnon, C., Meier, E., Brandner, C., Smeding, A., & Butera, F. (2015). Improving Low Achievers' Academic Performance at University by Changing the Social Value of Mastery Goals. *American Educational Research Journal*, *52*(4), 720–749. https://doi.org/10.3102/0002831215585137

Fantuzzo, J. W., Gadsden, V. L., & McDermott, P. A. (2011). An Integrated Curriculum to Improve Mathematics, Language, and Literacy for Head Start Children. *American Educational Research Journal*, *48*(3), 763–793. https://doi.org/10.3102/0002831210385446

Greene, W. H. (2002). Econometric Analysis. In *Econometric Analysis*.

Henningsen A. (2011). Censreg: Censored Regression (Tobit) Models. R package version 0.5, http://CRAN.R-project.org/package=censReg

Jennings, M. A., & Cribbie, R. A. (2016). Comparing Pre-Post Change Across Groups: Guidelines for Choosing between Difference Scores, ANCOVA, and Residual Change Scores. *Journal of Data Science*, *14*, 205–230.

Kim, R., Peters, M. A. K., & Shams, L. (2012). 0 + 1 > 1: How Adding Noninformative Sound Improves Performance on a Visual Task . *Psychological Science* , *23*(1), 6–12. https://doi.org/10.1177/0956797611420662

Liu, Q., & Wang, L. (2018). DACF: Data Analysis with Ceiling and/or Floor Data. CRAN

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (3rd ed.). New York: Routledge.

Miller GA. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2):81–97. https://doi.org/10.1037/h0043158

Muthen, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 131–143.

Muthén, L. K., & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Olsen, M. K., & Schafer, J. L. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association*, *96*(454), 730–745. https://doi.org/10.1198/016214501753168389

Piccinin, A. M., Muniz-Terrera, G., Clouston, S., Reynolds, C. A., Thorvaldsson, V., Deary, I. J., … Spiro, A. (2013). Coordinated analysis of age, sex, and education effects on change in MMSE scores. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *68*(3), 374–390.

Priebe, K., Kleindienst, N., Zimmer, J., Koudela, S., Ebner-Priemer, U., & Bohus, M. (2013). Frequency of intrusions and flashbacks in patients with posttraumatic stress disorder related to childhood sexual abuse: An electronic diary study. *Psychological Assessment*, *25*(4), 1370–1376. https://doi.org/10.1037/a0033816

Salthouse, T. A. (2004). Localizing age-related individual differences in a hierarchical structure. *Intelligence*, *32*(6), 541–561. https://doi.org/10.1016/j.intell.2004.07.003

Schweizer, K. (2016). A confirmatory factor model for the investigation of cognitive data showing a ceiling effect: an example. In *Quantitative Psychology Research* (pp. 187–197). Springer International Publishing.

Sokol-Hessner, P., Lackovic, S. F., Tobe, R. H., Camerer, C. F., Leventhal, B. L., & Phelps, E. A. (2015). Determinants of Propranolol's Selective Effect on Loss Aversion. *Psychological Science*, *26*(7), 1123–1130. https://doi.org/10.1177/0956797615582026

Timeo, S., Farroni, T., & Maass, A. (2017). Race and Color: Two Sides of One Story? Development of Biases in Categorical Perception. *Child Development*, *88*(1), 83–102. https://doi.org/10.1111/cdev.12564

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, *26*(1), 24–36. https://doi.org/10.2307/1907382

Ulber, J., Hamann, K., & Tomasello, M. (2016). Extrinsic Rewards Diminish Costly Sharing in 3-Year-Olds. *Child Development*, *87*(4), 1192–1203. https://doi.org/10.1111/cdev.12534

Uttl, B. (2005). Measurement of Individual Differences. *Psychological Science*, *16*(6), 460–467. https://doi.org/10.1111/j.0956-7976.2005.01557.x

Wang, L., & Zhang, Z. (2011). Estimating and Testing Mediation Effects with Censored Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(1), 18–34. https://doi.org/10.1080/10705511.2011.534324

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating Ceiling Effects in Longitudinal Data Analysis. *Multivariate Behav Res*, *43*(3), 476–496. https://doi.org/10.1080/00273170802285941

Welch, B. L. (1947). The generalisation of student's problems when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35. https://doi.org/10.1093/BIOMET/34.1-2.28

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.