psych

MDPI

Article

Accurate Standard Errors in Multilevel Modeling with Heteroscedasticity: A Computationally More Efficient Jackknife Technique

Steffen Zitzmann [1,*,†], Sebastian Weirich [2,†] and Martin Hecht [3]

[1] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, 72072 Tübingen, Germany
[2] Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, 10117 Berlin, Germany
[3] Faculty of Humanities and Social Sciences, Helmut Schmidt University, 22043 Hamburg, Germany
* Correspondence: steffen.zitzmann@uni-tuebingen.de
† These authors contributed equally to this work.

Abstract: In random-effects models, hierarchical linear models, or multilevel models, it is typically assumed that the variances within higher-level units are homoscedastic, meaning that they are equal across these units. However, this assumption is often violated in research. Depending on the degree of violation, this can lead to biased standard errors of higher-level parameters and thus to incorrect inferences. In this article, we describe a resampling technique for obtaining standard errors—Zitzmann's jackknife. We conducted a Monte Carlo simulation study to compare the technique with the commonly used delete-1 jackknife, the robust standard error in M*plus*, and a modified version of the commonly used delete-1 jackknife. Findings revealed that the resampling techniques clearly outperformed the robust standard error in rather small samples with high levels of heteroscedasticity. Moreover, Zitzmann's jackknife tended to perform somewhat better than the two versions of the delete-1 jackknife and was much faster.

Keywords: multilevel modeling; heteroscedasticity; standard error; jackknife; computational efficiency

Citation: Zitzmann, S.; Weirich, S.; Hecht, M. Accurate Standard Errors in Multilevel Modeling with Heteroscedasticity: A Computationally More Efficient Jackknife Technique. *Psych* **2023**, *5*, 757–769. https://doi.org/10.3390/psych5030049

Academic Editor: Okan Bulut

Received: 28 June 2023
Revised: 14 July 2023
Accepted: 19 July 2023
Published: 21 July 2023

1. Introduction

Random-effects models, hierarchical linear models, or multilevel models are powerful tools for analyzing longitudinal, e.g., [1–5], and cross-sectional designs, e.g., [6–11] with lower-level units (e.g., individuals) nested within higher-level units (e.g., groups) in educational psychology and related sciences. In these models, the assumption of homoscedasticity is commonly made by assuming that the variances within groups are homoscedastic [12,13], meaning that they are the same across the groups. Although multilevel models deal with violations of the independence assumption (i.e., the assumption that observations are independent) by taking the nesting of units into account, they are not robust against violations of homoscedasticity, and an effective solution to deal with this issue, such as explicitly modeling heteroscedasticity or using newly developed, more robust methods, e.g., [13], is not always available. As Goldstein [12] noted, heteroscedasticity is often found in naturally occuring groups. For example, males tend to be less similar than females are in, for example, intellectual abilities [14]. Depending on the level of heteroscedasticity, it can lead to biases in standard errors of group-level parameters and thus to incorrect inferences regarding these parameters, e.g., [15]. Although these problems are well-known, and robust methods have long been used in other areas, e.g., [16–18], little progress has been made to adapt these and other methods to multilevel modeling.

A multilevel model is given by the following equations. If we apply the notation of Raudenbush and Bryk [19], the model reads at the individual level:

$$\text{Level 1:} \quad Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \tag{1}$$

for an individual $i = 1, \ldots, n_j$ in a group $j = 1, \ldots, J$. The r_{ij} are normally distributed residuals with variance $\mathrm{Var}(r_{ij})$. This model is homoscedastic, meaning that $\mathrm{Var}\left(r_{ij'}\right) = \mathrm{Var}\left(r_{ij''}\right)$ for different groups j' and j''. At the group level, the intercepts β_{0j} and the slopes β_{1j} are modeled as

$$\begin{aligned} \text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j} \end{aligned} \tag{2}$$

where γ_{00} and γ_{10} are the overall intercept and the overall slope of X_{ij}, γ_{01} and γ_{11} are the main effect of Z_j and the cross-level interaction of X_{ij} and Z_j, and u_{0j} and u_{1j} are normally distributed residuals with variances $\mathrm{Var}(u_{0j})$ and $\mathrm{Var}(u_{1j})$, respectively. The parameters γ_{01}, γ_{11}, $\mathrm{Var}(u_{0j})$, and $\mathrm{Var}(u_{1j})$ are typical group-level parameters of interest. By inserting Equation (2) into Equation (1), the model simplifies to one equation:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{1j} X_{ij} + u_{0j} + r_{ij} \tag{3}$$

In this article, we describe a resampling technique for obtaining standard errors of group-level parameters that has been proposed by Zitzmann [20] as an efficient way to obtain standard errors in multilevel factor score regression; see also [21,22]. However, the technique can also be applied with other types of estimation, such as Maximum Likelihood (ML) estimation, which can be useful particularly when the standard error from ML estimation lacks robustness. Hereafter, we will refer to this specific resampling technique as *Zitzmann's jackknife*. Although the technique has not specifically been developed for the purpose of dealing with heteroscedasticity, it is reasonable to assume that this technique would perform well under heteroscedasticity. In other words, unlike Zitzmann [20], we identify his method as a specific means to deal with heteroscedasticity in multilevel modeling, and thus, we are the first to discuss it more in relation to robust methods (i.e., robust against the violation of homoscedasticity). However, as Zitzmann [20] did not study its performance in the presence of heteroscedasticity, it is yet unknown to what extent this method is really robust. Therefore, we conducted a Monte Carlo simulation study in order to test its performance, in which we also compared this method with the commonly used delete-1 jackknife, the robust standard error in M*plus* [23], and a modified version of the commonly used delete-1 jackknife, as these are the methods that are either readily available in standard software or easily computable. Note that to the best of our knowledge, such a study does not yet exist.

2. Zitzmann's Jackknife

Zitzmann's jackknife is a specific resampling method that may be located within the class of the jackknife techniques. As such, Zitzmann's jackknife shares some features with more common jackknifes, but it also deviates from them. In the following, we describe how this method works, using a minimal example.

Let θ be a group-level parameter such as γ_{01}, γ_{11}, $\mathrm{Var}(u_{0j})$, or $\mathrm{Var}(u_{1j})$ in the above model; Zitzmann's jackknife first computes estimates of this parameter from R subsamples. These subsamples are obtained by omitting d groups. More specifically, they are obtained by dividing the indices of the groups $(1, \ldots, \text{no. of groups})$ into $R = \frac{\text{no. of groups}}{d}$ nonoverlapping subsets, and then using these nonoverlapping subsets of indices to create the subsamples. For example, suppose the total number of groups is 15, and $R = 5$ subsamples should be created (each omitting $d = 3$ groups). The indices of the groups can be arranged as

$$1 \;\; 2 \;\; 3 \;\; 4 \;\; 5 \;\; 6 \;\; 7 \;\; 8 \;\; 9 \;\; 10 \;\; 11 \;\; 12 \;\; 13 \;\; 14 \;\; 15 \tag{4}$$

One natural way to divide these indices into $R = 5$ nonoverlapping subsets is

$$\underbrace{1 \;\; 2 \;\; 3}_{1} \;\; \underbrace{4 \;\; 5 \;\; 6}_{2} \;\; \underbrace{7 \;\; 8 \;\; 9}_{3} \;\; \underbrace{10 \;\; 11 \;\; 12}_{4} \;\; \underbrace{13 \;\; 14 \;\; 15}_{5} \tag{5}$$

For each subset, a subsample of the data is created by omitting the indexed groups. Thus, each resulting subsample includes only 12 groups. The five subsamples are

$$\begin{array}{lll} \text{Subsample} & 1: & 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15 \\ \text{Subsample} & 2: & 1\ 2\ 3\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15 \\ \text{Subsample} & 3: & 1\ 2\ 3\ 4\ 5\ 6\ 10\ 11\ 12\ 13\ 14\ 15 \\ \text{Subsample} & 4: & 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 13\ 14\ 15 \\ \text{Subsample} & 5: & 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12 \end{array} \quad (6)$$

Note that omitting groups implies omitting individuals as well because individuals are nested within groups. However, the procedure operates at the group level in the first place, which differs from single-level resampling techniques.

On each of the five subsamples, the analysis model (i.e., the multilevel model of interest) is run, each yielding an estimate of the group-level parameter:

$$\hat{\theta}_1 \quad \hat{\theta}_2 \quad \hat{\theta}_3 \quad \hat{\theta}_4 \quad \hat{\theta}_5 \quad (7)$$

Based on these estimates, the standard error of the estimate $\hat{\theta}$ from the complete sample is then estimated by the following formula:

$$\text{Zitzmann's jackknifed } \widehat{SE} = \sqrt{(R-1)\sum_{j=1}^{R}\left(\hat{\theta}_j - \bar{\hat{\theta}}\right)^2 / R} \quad (8)$$

where $\bar{\hat{\theta}} = \sum_{j=1}^{R} \hat{\theta}_j / R$ is simply the average across the estimates from the subsamples.

3. Simulation Study

As with all new methods, it is crucial to assess the performance of Zitzmann's jackknife, preferably by means of a simulation—a study in which one or more methods are tested or compared using a large number of artificially generated data. Therefore, we conducted a simulation study to compare the performance of Zitzmann's jackknife with the performances of the commonly used delete-1 jackknife, the robust standard error in M*plus*, and a modification of the delete-1 jackknife.

3.1. Method

3.1.1. Data Generation

The data-generating model was a heteroscedastic version of a simple multilevel model, namely the two-level random intercept-only model. The model reads at the individual level:

$$\text{Level 1:} \quad Y_{ij} = \beta_{0j} + r_{ij} \quad (9)$$

Note that in this model, $\text{Var}\left(r_{ij'}\right) \neq \text{Var}\left(r_{ij''}\right)$ for different groups j' and j'' (i.e., heteroscedasticity). At the group level, the intercepts β_{0j} are modeled as

$$\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + u_{0j} \quad (10)$$

In the context of this model, we call $\text{Var}\left(u_{0j}\right)$ the *variance between groups*. If we insert Equation (10) into Equation (9), the model becomes

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \quad (11)$$

We varied the sample size by varying the number of groups ($J = 20$ vs. 40 vs. 100 vs. 400) and the group size (small vs. medium vs. large). Although there are studies involving fewer than 40 groups of small to large group sizes, around 40 are commonly found in educational psychology [24], and 400 were chosen to study the asymptotic behavior of the methods. The group size was not held perfectly equal across groups in order to generate data that resembled data structures typically found in multilevel research. For each group, the group

size was drawn from a uniform distribution with specified bounds. As a consequence, in the condition with a small group size, the group size varied between $n_j = 7$ and 9, in the condition with a medium-sized group size, it varied between $n_j = 11$ and 17, and in the condition with a large group size, it varied between $n_j = 19$ and 25. More importantly, we varied the degree of violation of the homoscedasticity assumption, which we hereafter also refer to as the level of heteroscedasticity (moderate vs. high vs. very high). To this end, the within-group standard deviation for each group was drawn from a uniform distribution with bounds depending on the level of heteroscedasticity. In the condition with a moderate level of heteroscedasticity, the standard deviation varied between groups between $\sqrt{\text{Var}(r_{ij})} = 70$ and 130. That is, the most homogeneous group had a standard deviation of 70, and the most heterogeneous group had a standard deviation of 130. In the condition with a high level of heteroscedasticity, the standard deviations varied between $\sqrt{\text{Var}(r_{ij})} = 45$ and 170, and in the condition with a very high level of heteroscedasticity, they varied between $\sqrt{\text{Var}(r_{ij})} = 25$ and 200. Thus, overall, the design of the simulation had $4 \times 3 \times 3 = 36$ conditions.

For each of these conditions, $S = 1000$ datasets were generated in two steps: In the first step, for each group j, the intercept β_{0j} was drawn from a normal distribution with a mean of $\gamma_{00} = 500$ and a variance of $\text{Var}(u_{0j}) = 100$. For example, in the condition with 20 groups, 20 intercepts were generated. Moreover, for each group, a value of the standard deviation within the group was drawn. For example, in the condition with a moderate level of heteroscedasticity, this standard deviations had a value between $\sqrt{\text{Var}(r_{ij})} = 70$ and 130. Finally, a value of the group size n_j was drawn for each group.

In the second step, the individual observations were drawn from a normal distribution with mean β_{0j} and variance $\text{Var}(r_{ij})$. Each resulting dataset was analyzed with a two-level random intercept model in which the variances within groups were assumed to be equal across groups (i.e., homoscedasticity assumption), which is the default and often also the only model in many software programs.

3.1.2. Model Estimation

Model estimation was carried out by ML estimation in which the likelihood function or the log-likelihood is maximized [25]. To briefly illustrate this procedure, suppose for a moment that Y is grand mean-centered (i.e., dropping the intercept α), group sizes are equal (i.e., dropping the index j in n_j), and homoscedasticity holds. Then, the log-likelihood of the two-level random intercept-only model can be expressed as

$$\begin{aligned} \log L = c - J(n-1)\log\left(\text{Var}(r_{ij})\right)/2 - J\log(g)/2 - \sum_{j=1}^{J}\sum_{i=1}^{n} Y_{ij}^2/\left(2\text{Var}(r_{ij})\right) \\ + n\left(g - \text{Var}(r_{ij})\right)\sum_{j=1}^{J} \bar{Y}_{\bullet j}^2/\left(2\text{Var}(r_{ij})g\right) \end{aligned} \tag{12}$$

where c is a constant term, and g is used as an abbreviation for $n\text{Var}(u_{0j}) + \text{Var}(r_{ij})$. Analytically maximizing this log-likelihood yields the following ML estimates:

$$\widehat{\text{Var}}(r_{ij}) = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n} Y_{ij}^2 - n\sum_{j=1}^{J}\bar{Y}_{\bullet j}^2}{J(n-1)} \tag{13}$$

$$\widehat{\text{Var}}(u_{0j}) = \sum_{j=1}^{J}\bar{Y}_{\bullet j}^2/J - \widehat{\text{Var}}(r_{ij})/n \tag{14}$$

While assuming homoscedasticity in the actual analysis model (i.e., the homoscedastic two-level random intercept-only model), Y was noncentered, and the group sizes were

allowed to differ between groups. Therefore, instead of the analytical solution from above, we used the more flexible M*plus* software in order to solve the optimization problem

$$\left(\hat{\gamma}_{00}, \widehat{\mathrm{Var}}(r_{ij}), \widehat{\mathrm{Var}}(u_{0j})\right) = \underset{\left(\hat{\gamma}_{00}, \widehat{\mathrm{Var}}(r_{ij}), \widehat{\mathrm{Var}}(u_{0j})\right)}{\arg\max} \log L \tag{15}$$

by maximizing the log-likelihood through numerical optimization; see [26].

Standard errors were obtained by three different methods. Zitzmann's jackknife was implemented as illustrated above, however, with a fixed number of $R = 20$ subsamples. This number was taken from previous work, e.g., [20], which showed that it was sufficient for the jackknife to perform well.

In addition to Zitzmann's jackknife, we implemented the commonly used delete-1 jackknife, which is also used by the R-package BIFIEsurvey [27]. Using the symbol J for the total number of groups, the jackknifed standard error is obtained by aggregating the estimates from J subsamples created by omitting only one group:

$$\text{jackknifed } \widehat{SE} = \sqrt{(J-1)\sum_{j=1}^{J}\left(\hat{\theta}_j - \hat{\theta}\right)^2 / J} \tag{16}$$

where $\hat{\theta}$ is the estimate from the complete dataset (A. Robitzsch, personal communication, 6 December 2022), which is a major deviation from Zitzmann's jackknife. See also the second formula in Kolenikov [28], which is equivalent to the presented one in the case of nonstratified data.

Moreover, we retained the robust standard error from M*plus*, which is essentially a sandwich standard error that is rather robust against violations of the normality assumption and other assumptions but not necessarily also against violations of homoscedasticity (for a formal expression of this standard error, see [29]; see also [13,30]). Indeed, it has been argued that "unless the number of groups is very large, the robust standard errors are not up to that task" (p. 439, [15]).

Finally, as another benchmark, we implemented a modified version of the delete-1 jackknife, which differs from the commonly used delete-1 jackknife in that $\bar{\hat{\theta}}$ (i.e., the average across the estimates from the subsamples) is used in place of the estimate from the complete dataset, which makes the modified version more similar to Zitzmann's jackknife and identical to Zitzmann's jackknife in situations where only one group is omitted.

All four methods were implemented in R [31].

3.1.3. Evaluation Criteria

For each type of standard error, two indicators of accuracy were computed from the simulated datasets per simulation condition. The percentage of bias of the standard error was estimated as

$$\%\text{ bias} = 100\% \times \left(\sum_{s=1}^{S}\widehat{SE}_s / S - SE^*\right) \Big/ SE^* \tag{17}$$

where $\widehat{SE}_s$ is the standard error obtained from the sth data set, and SE^* is a proxy for the true standard error and computed as the empirical standard deviation across the parameter estimates from the datasets. The standard error ratio was computed as

$$SE\text{ ratio} = \sum_{s=1}^{S}\widehat{SE}_r / SE^* \tag{18}$$

Note that whereas the exact number of datasets used for computing the percentage of bias and the standard error ratio per condition was less than 1000 because datasets for which ML estimation yielded outliers (i.e., values beyond Q3 plus 1.5 times the interquartile range, in analogy to boxplot conventions) were discarded, all 1000 datasets were used to compute the proxy for the true standard error.

We expected that the two resampling techniques would outperform the robust standard error in rather small samples with high levels of heteroscedasticity because the resampling techniques inherently account for heteroscedasticity, whereas the robust standard error does not.

In our evaluations, we focused on the group-level parameter in the model, which is the variance between groups ($\mathrm{Var}(u_{0j})$). The findings from this study are the subject of the next section.

3.2. Results

Table 1 gives an overview of the sources of variation in the bias of the standard error. Main effects and interactions that explained a substantial amount of variability (i.e., an eta-squared η^2 greater than 0.01) are printed in bold type. Only these substantial eta-squared values will be interpreted next. Table 2 shows the detailed results for the percentage of bias for all three methods as a function of the simulation condition.

As can be seen from Table 1, all factors except the number of groups showed substantial eta-squared values. Among the main effects of these factors, the main effect of the method factor showed the largest eta-squared, meaning that this factor influenced the percentage of bias most. The differences between methods were most pronounced between the resampling techniques and the robust standard error. A look into Table 2 reveals that the differences between Zitzmann's jackknife and the two delete-1 jackknifes were generally small, with Zitzmann's jackknife tending to provide somewhat less biased standard errors. Overall, Zitzmanns' jackknife provided on average a bias of 4.8% (averaged over simulation conditions), which can be considered acceptable [32]. The average bias of the commonly used delete-1 jackknife was 15.1% and thus somewhat larger. The average bias of the robust standard error was 94.1% and thus clearly unacceptably high, and the average bias of the modified delete-1 jackknife was 13.2% and thus very similar to the bias of the commonly used delete-1 jackknife. The factor with the second largest eta-squared value was the group size (see Table 1). With an increasing group size, the percentage of bias reduced (see Table 2). What is more important for the present investigation, the level of heteroscedasticity also explained a substantial amount of variance, although its eta-squared was only small. The more the within-group standard standard deviation varied between groups, the more biased was the standard error.

Four of the two-way interaction effects showed substantial eta-squared values. These were the three interactions involving the method factor, with the two sample size $\times$ method interactions being the ones with the largest eta-squared values. When the sample size increased, the differences between the methods decreased. Specifically, whereas the sample size did not affect the performance of the resampling techniques much, it reduced the bias for the robust standard error considerably, and thus, the differences between the resampling techniques and the robust standard error decreased as well. More important, the level of heteroscedasticity $\times$ method interaction showed a substantial but small eta-squared. A larger variance in the within-group standard deviation between groups tended to lead to larger differences between methods, particularly between the resampling techniques and the robust standard error. There was also a number of groups $\times$ level of heteroscedasticity interaction, indicating that the impact of heteroscedasticity on the bias depended on the sample size. However, the impact was not consistent across methods, as indicated by the number of groups $\times$ level of heteroscedasticity $\times$ method interaction.

An identical picture emerged with regard to the standard error ratio. Zitzmanns' jackknife exhibited an average standard error ratio of 1.05 across all conditions, whereas the average standard error ratio of the commonly used delete-1 jackknife was 1.15, the average standard error ratio of the robust standard error was 1.94, and the average standard error ratio of the modified delete-1 jackknife was 1.13. For the detailed results, see Tables 1 and 2.

Table 1. Simulation study results: sources of variation.

	η^2
Percentage of Bias	
Main effects	
No. of Groups	0.00
Group Size	**0.04**
Level of Heteroscedasticity	**0.01**
Method	**0.54**
Two-Way Interaction Effects	
No. of Groups × Group Size	0.00
No. of Groups × Level of Heteroscedasticity	**0.01**
Group Size × Level of Heteroscedasticity	0.00
No. of Groups × Method	**0.25**
Group Size × Method	**0.07**
Level of Heteroscedasticity × Method	**0.01**
Three-Way Interaction Effects	
No. of Groups × Group Size × Level of Heteroscedasticity	**0.01**
No. of Groups × Group Size × Method	0.00
No. of Groups × Level of Heteroscedasticity × Method	**0.01**
Group Size × Level of Heteroscedasticity × Method	0.00
Four-Way Interaction Effects	
No. of Groups × Group Size × Level of Heteroscedasticity × Method	0.00
Standard Error Ratio	
Main effects	
No. of Groups	0.00
Group Size	**0.04**
Level of Heteroscedasticity	**0.01**
Method	**0.54**
Two-Way Interaction Effects	
No. of Groups × Group Size	0.00
No. of Groups × Level of Heteroscedasticity	**0.01**
Group Size × Level of Heteroscedasticity	0.00
No. of Groups × Method	**0.25**
Group Size × Method	**0.07**
Level of Heteroscedasticity × Method	**0.01**
Three-Way Interaction Effects	
No. of Groups × Group Size × Level of Heteroscedasticity	**0.01**
No. of Groups × Group Size × Method	0.00
No. of Groups × Level of Heteroscedasticity × Method	**0.01**
Group Size × Level of Heteroscedasticity × Method	0.00
Four-Way Interaction Effects	
No. of Groups × Group Size × Level of Heteroscedasticity × Method	0.00

Note. Substantial eta-squared values are indicated in bold.

Table 2. Simulation study results: percentage of bias and standard error ratio.

			Percentage of Bias				Standard Error Ratio			
No. of Groups	**Group Size**	**Level of Heteroscedasticity**	**Zitzmann's Jackknife**	**Commonly Used Delete-1 Jackknife**	**Robust Standard Error**	**Modified Delete-1 Jackknife**	**Zitzmann's Jackknife**	**Commonly Used Delete-1 Jackknife**	**Robust Standard Error**	**Modified Delete-1 Jackknife**
20	Small	Moderate	−5.5	−2.6	174.3	−5.5	0.95	0.97	2.74	0.95
		High	−10.8	−7.6	205.9	−10.8	0.89	0.92	3.06	0.89
		Very High	−13.4	−10.8	174.0	−13.4	0.87	0.89	2.74	0.87
	Medium	Moderate	−4.5	−0.5	148.5	−4.5	0.96	1.00	2.49	0.96
		High	−9.0	−5.9	164.7	−9.0	0.91	0.94	2.65	0.91
		Very High	−14.3	−11.4	143.0	−14.3	0.86	0.89	2.43	0.86
	Large	Moderate	−6.3	−3.6	90.7	−6.3	0.94	0.96	1.91	0.94
		High	−13.0	−10.7	95.6	−13.0	0.87	0.89	1.96	0.87
		Very High	−8.9	−6.7	147.9	−8.9	0.91	0.93	2.48	0.91
40	Small	Moderate	3.8	16.4	158.2	11.3	1.04	1.16	2.58	1.11
		High	−4.8	5.9	167.7	2.0	0.95	1.06	2.68	1.02
		Very High	−4.2	7.5	167.8	2.7	0.96	1.08	2.68	1.03
	Medium	Moderate	−4.0	3.5	98.7	0.0	0.96	1.03	1.99	1.00
		High	2.6	14.3	136.8	10.7	1.03	1.14	2.37	1.11
		Very High	−8.9	1.5	155.1	−2.8	0.91	1.02	2.55	0.97
	Large	Moderate	−2.0	6.3	57.1	2.7	0.98	1.06	1.57	1.03
		High	−3.1	4.6	80.1	1.3	0.97	1.05	1.80	1.01
		Very High	−1.8	6.8	98.9	3.4	0.98	1.07	1.99	1.03
100	Small	Moderate	11.0	31.0	101.2	28.2	1.11	1.31	2.01	1.28
		High	12.9	32.1	122.3	28.7	1.13	1.32	2.22	1.29
		Very High	11.5	31.6	144.9	28.4	1.12	1.32	2.45	1.28
	Medium	Moderate	9.2	20.8	62.5	20.0	1.09	1.21	1.62	1.20
		High	13.5	34.2	75.2	29.9	1.14	1.34	1.75	1.30
		Very High	14.0	40.2	100.7	35.0	1.14	1.40	2.01	1.35
	Large	Moderate	3.1	11.9	19.3	11.2	1.03	1.12	1.19	1.11
		High	8.1	26.1	38.2	23.7	1.08	1.26	1.38	1.24
		Very High	3.7	17.8	50.5	15.3	1.04	1.18	1.50	1.15
400	Small	Moderate	28.3	27.8	49.5	28.4	1.28	1.28	1.49	1.28
		High	34.7	40.7	60.0	42.6	1.35	1.41	1.60	1.43
		Very High	40.3	49.2	75.2	52.5	1.40	1.49	1.75	1.53
	Medium	Moderate	7.7	7.7	5.8	7.9	1.08	1.08	1.06	1.08
		High	23.6	34.5	17.6	36.0	1.24	1.35	1.18	1.36
		Very High	29.5	44.9	29.1	49.4	1.30	1.45	1.29	1.49
	Large	Moderate	4.6	2.2	−9.3	1.9	1.05	1.02	0.91	1.02
		High	13.4	61.9	−10.7	63.8	1.13	1.62	0.89	1.64
		Very High	11.4	22.8	−8.3	24.8	1.11	1.23	0.92	1.25

4. Real Data Example

To further illustrate the different methods for obtaining standard errors, we applied them to real data. The data belong to the Dutch PISA 2006 study and were already used by Fox [33], who provided a link to download them (see also the BIFIEsurvey package, where they are called "data.pisaNLD"). They contain achievement data of $J = 154$ classes with the class size ranging from $n_j = 13$ to 30. Importantly, there is some variation in the within-class variance, as indicated by a scatter of the within-class standard deviations of 9.35, which means that homoscedasticity does not hold.

Like in the simulation, we fit a homoscedastic two-level random intercept model using M*plus* and focused on the variance between groups. The maximum likelihood estimate of this parameter was 5421.94. Zitzmann's jackknife yielded a standard error of 613.26, the commonly used delete-1 jackknife a standard error of 506.26, the robust standard error was 520.58, and the modified delete-1 jackknife resulted in a standard error of 506.29. As is evident from these values, the different methods differed, with Zitzmann's jackknife clearly standing out, which could have been anticipated based on the findings from the simulation. Given that in the simulation, Zitzmann's jackknife performed best, we tend to prefer the standard error from this method in order to quantify the uncertainty in the estimate of the variance between groups.

5. Summary and Discussion

Standard multilevel modeling requires that the variances within groups be homoscedastic. However, in research, this assumption may be violated, leading to biased standard errors of group-level parameters and incorrect inferences. We described Zitzmann's jackknife, a resampling technique for obtaining standard errors, and conducted a simulation study to compare the technique with the commonly used delete-1 jackknife, the robust standard error in M*plus*, and a modified delete-1 jackknife. In line with our expectation, all three resampling techniques outperformed the robust standard error regarding percentage of bias and standard error ratio in rather small samples with high levels of heteroscedasticity. Somewhat surprisingly, Zitzmann's jackknife tended to exhibit a better performance than the different versions of the delete-1 jackknife, although the differences between these methods were small. Note that we implemented Zitzmann's jackknife with 20 subsamples, whereas the delete-1 jackknifes used as many subsamples as there were groups. Therefore, similar standard errors and thus similar performances could have been expected in conditions with 20 groups but not necessarily in conditions with more groups.

It is interesting to note that Zitzmann's jackknife was much faster than the delete-1 jackknifes because, except in conditions with 20 groups, the delete-1 jackknifes required that many more subsamples be generated and analyzed. For example, in conditions with 100 groups, 100 subsamples had to be analyzed. In this case, Zitzmann's jackknife reduced the computing time by a factor of about five! That is, Zitzmann's jackknife was about five times faster than the delete-1 jackknifes, which shows that Zitzmann's jackknife is computationally more efficient (while performing similarly). Such an efficient resampling technique is particularly beneficial in situations in which a complicated model is fit to a large dataset, as is the case in international large-scale assessment studies ([34], but see [35] for one promising approach for reducing the run-time of multilevel models).

In the simulation study and the real data example, Zitzmann's jackknife was used with only 20 subsamples. Therefore, it might be speculated that the scatter of the standard error estimate would generally be larger for Zitzmann's jackknife compared to the delete-1 jackknifes, as the number of subsamples involved in this jackknife is smaller. However, this did not show up in the simulation, where Zitzmann's jackknife tended to scatter even less than the delete-1 jackknifes (see Table A1 in Appendix A).

Although it has repeatedly been found in relatively simple multilevel models and consequently argued that Zitzmann's jackknife can yield correct standard errors with 20 subsamples [20,22], a larger number might nevertheless be required in more advanced multilevel models. It would be interesting to investigate the number of subsamples that is

required in models involving multiple indicators, e.g., [36–38], or (latent) moderation and nonlinear effects, e.g., [39].

Often, samples are not proportional to the population (i.e., sampling weights are available). Future research could extend Zitzmann's jackknife in such a way that the technique can also be applied to this type of sample.

To sum up, we showed that Zitzmann's jackknife can be an attractive choice when homoscedasticity is violated, especially when other methods are computationally not very efficient or not readily available in standard software.

Author Contributions: S.Z.: writing. S.W.: conducting the simulation. M.H.: writing and lead. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available on request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Further Results

The following table shows an overview of the mean and the range of the values obtained by each standard error method.

Table A1. Additional simulation study results: mean and range (in brackets) of standard errors.

No. of Groups	Group Size	Level of Heteroscedasticity	Zitzmann's Jackknife	Commonly Used Delete-1 Jackknife	Robust Standard Error	Modified Delete-1 Jackknife
20	Small	Moderate	7.94 [0.24, 25.04]	8.20 [0.24, 27.00]	43.60 [0.11, 5483.15]	7.94 [0.24, 25.04]
		High	8.95 [0.38, 34.49]	9.29 [0.39, 38.24]	80.27 [0.54, 13,619.31]	8.95 [0.38, 34.49]
		Very High	10.64 [0.44, 43.39]	10.94 [0.45, 43.49]	97.68 [0.02, 11,262.94]	10.64 [0.44, 43.39]
	Medium	Moderate	6.26 [0.32, 23.06]	6.52 [0.36, 23.21]	31.14 [0.39, 2914.86]	6.26 [0.32, 23.06]
		High	6.99 [0.26, 26.09]	7.25 [0.33, 26.22]	80.67 [0.99, 32,518.31]	6.99 [0.26, 26.09]
		Very High	7.97 [0.30, 27.06]	8.24 [0.35, 27.35]	55.10 [1.05, 9190.37]	7.97 [0.30, 27.06]
	Large	Moderate	5.39 [0.15, 16.42]	5.54 [0.19, 16.47]	20.29 [0.28, 1989.66]	5.39 [0.15, 16.42]
		High	6.07 [0.21, 21.12]	6.22 [0.26, 21.13]	30.06 [0.13, 3480.56]	6.07 [0.21, 21.12]
		Very High	6.78 [0.26, 24.38]	6.95 [0.32, 26.08]	41.92 [2.67, 5184.94]	6.78 [0.26, 24.38]
40	Small	Moderate	7.09 [0.44, 19.82]	8.17 [0.42, 31.19]	25.39 [3.22, 856.23]	7.61 [0.39, 24.63]
		High	8.21 [0.55, 28.54]	9.38 [0.49, 52.29]	31.41 [1.05, 1220.91]	8.77 [0.46, 29.82]
		Very High	9.73 [0.36, 33.66]	10.99 [0.42, 43.88]	39.13 [4.00, 1769.04]	10.39 [0.37, 32.50]
	Medium	Moderate	5.55 [0.22, 16.65]	6.01 [0.25, 26.70]	15.13 [3.17, 291.09]	5.77 [0.22, 16.14]
		High	6.70 [0.36, 23.57]	7.66 [0.26, 31.28]	20.36 [4.95, 456.23]	7.20 [0.23, 20.99]
		Very High	7.20 [0.25, 22.30]	8.25 [0.25, 36.89]	31.87 [2.41, 1724.39]	7.67 [0.24, 23.67]
	Large	Moderate	4.82 [0.25, 12.11]	5.28 [0.24, 17.44]	10.33 [3.50, 307.89]	5.05 [0.23, 12.53]
		High	5.53 [0.24, 18.03]	6.03 [0.27, 21.95]	13.05 [3.88, 125.25]	5.79 [0.24, 18.38]
		Very High	6.24 [0.21, 18.01]	6.89 [0.21, 30.22]	20.27 [4.37, 3016.62]	6.58 [0.19, 18.45]
100	Small	Moderate	6.11 [0.58, 18.35]	7.69 [0.52, 60.41]	12.02 [4.14, 98.06]	7.25 [0.52, 36.06]
		High	7.50 [0.54, 21.06]	9.53 [0.68, 49.02]	16.50 [4.87, 201.95]	8.97 [0.67, 38.04]
		Very High	8.52 [0.48, 22.50]	10.69 [0.34, 75.00]	22.12 [5.82, 667.85]	10.17 [0.33, 46.24]
	Medium	Moderate	4.84 [0.46, 12.21]	5.81 [0.48, 36.44]	8.12 [3.14, 101.48]	5.52 [0.48, 24.94]
		High	6.12 [0.23, 14.36]	7.78 [0.33, 41.88]	10.96 [3.87, 153.51]	7.30 [0.31, 31.02]
		Very High	6.88 [0.31, 20.27]	9.05 [0.35, 45.07]	13.70 [4.52, 175.57]	8.43 [0.34, 32.05]
	Large	Moderate	4.02 [0.29, 9.69]	4.70 [0.28, 25.80]	5.28 [2.49, 53.66]	4.51 [0.28, 20.42]
		High	4.93 [0.38, 11.91]	6.33 [0.40, 38.68]	7.43 [3.01, 306.17]	5.90 [0.40, 24.25]
		Very High	5.54 [0.30, 13.83]	6.76 [0.26, 34.78]	9.58 [3.50, 78.43]	6.43 [0.26, 25.71]
400	Small	Moderate	5.18 [0.43, 12.16]	7.31 [0.54, 110.38]	6.65 [2.82, 106.79]	6.90 [0.53, 67.71]
		High	6.71 [0.40, 13.97]	9.70 [0.49, 118.64]	10.03 [3.52, 186.19]	9.27 [0.49, 77.28]
		Very High	7.75 [0.70, 17.71]	11.70 [0.74, 105.57]	14.22 [4.13, 1380.48]	11.04 [0.74, 79.20]
	Medium	Moderate	3.53 [0.84, 9.12]	4.27 [0.92, 59.38]	3.88 [2.06, 94.53]	4.12 [0.92, 42.63]
		High	4.92 [0.52, 9.88]	6.81 [0.57, 77.77]	5.59 [1.67, 181.11]	6.48 [0.56, 53.01]
		Very High	5.88 [0.48, 12.41]	8.57 [0.59, 73.78]	7.47 [2.81, 162.03]	8.10 [0.58, 59.99]
	Large	Moderate	2.52 [1.02, 6.16]	3.09 [1.05, 19.42]	2.28 [1.54, 11.35]	2.93 [1.04, 18.23]
		High	3.65 [0.59, 7.87]	6.07 [0.88, 40.46]	3.25 [1.80, 36.95]	5.43 [0.88, 33.39]
		Very High	4.32 [0.64, 8.77]	6.46 [0.68, 57.02]	4.44 [2.22, 62.80]	5.98 [0.68, 44.91]

References

1. Skrondal, A.; Rabe-Hesketh, S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*; CRC Press: Boca Raton, FL, USA, 2004.
2. Hamaker, E.L.; Kuiper, R.M.; Grasman, R.P.P.P. A critique of the crosslagged panel model. *Psychol. Methods* **2015**, *20*, 102–116. [CrossRef] [PubMed]
3. Jongerling, J.; Laurenceau, J.P.; Hamaker, E.L. A multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivar. Behav. Res.* **2015**, *50*, 334–349. [CrossRef] [PubMed]
4. Voelkle, M.C.; Oud, J.H.L.; Davidov, E.; Schmidt, P. An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychol. Methods* **2012**, *17*, 176–192. [CrossRef]
5. Hecht, M.; Zitzmann, S. A computationally more efficient Bayesian approach for estimating continuous-time models. *Struct. Equ. Model.* **2020**, *27*, 829–840. [CrossRef]
6. Rabe-Hesketh, S.; Skrondal, A.; Pickles, A. Generalized multilevel structural equation modeling. *Psychometrika* **2004**, *69*, 167–190. [CrossRef]
7. Lüdtke, O.; Marsh, H.W.; Robitzsch, A.; Trautwein, U.; Asparouhov, T.; Muthén, B.O. The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychol. Methods* **2008**, *13*, 203–229. [CrossRef] [PubMed]
8. Preacher, K.J.; Zyphur, M.J.; Zhang, Z. A general multilevel SEM framework for assessing multilevel mediation. *Psychol. Methods* **2010**, *15*, 209–233. [CrossRef]
9. Zitzmann, S.; Lüdtke, O.; Robitzsch, A. A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivar. Behav. Res.* **2015**, *50*, 688–705. [CrossRef]
10. Stapleton, L.M.; Yang, J.S.; Hancock, G.R. Construct meaning in multilevel settings. *J. Educ. Behav. Stat.* **2016**, *41*, 481–520. [CrossRef]
11. Zitzmann, S.; Wagner, W.; Hecht, M.; Helm, C.; Fischer, C.; Bardach, L.; Göllner, R. How many classes and students should ideally be sampled when assessing the role of classroom climate via student ratings on a limited budget? An optimal design perspective. *Educ. Psychol. Rev.* **2022**, *34*, 511–536. [CrossRef]
12. Goldstein, H. Heteroscedasticity and Complex Variation. In *Wiley StatsRef: Statistics Reference Online*; Wiley: Hoboken, NJ, USA, 2014.
13. Huang, F.L.; Wiedermann, W.; Zhang, B. Accounting for heteroskedasticity resulting from between-group differences in multilevel models. *Multivar. Behav. Res.* **2022**, *58*, 637–657. [CrossRef]
14. Feingold, A. Sex differences in variability in intellectual abilities: A new look at an old controversy. *Rev. Educ. Res.* **1992**, *62*, 61–84. [CrossRef]
15. Maas, C.J.M.; Hox, J.J. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Comput. Stat. Data Anal.* **2004**, *46*, 427–440. [CrossRef]
16. Hayes, A.F.; Cai, L. Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behav. Res. Methods* **2007**, *39*, 709–722. [CrossRef]
17. Long, J.S.; Ervin, L.H. Using heteroscedasticity consistent standard errors in the linear regression model. *Am. Stat.* **2000**, *54*, 217–224. [CrossRef]
18. White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **1980**, *48*, 817–838. [CrossRef]
19. Raudenbush, S.W.; Bryk, A.S. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed.; Advanced quantitative techniques in the social sciences; Sage: Thousand Oaks, CA, USA, 2002.
20. Zitzmann, S. A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivar. Behav. Res.* **2018**, *53*, 612–632. [CrossRef]
21. Zitzmann, S.; Helm, C. Multilevel analysis of mediation, moderation, and nonlinear effects in small samples, using expected a posteriori estimates of factor scores. *Struct. Equ. Model.* **2021**, *28*, 529–546. [CrossRef]
22. Zitzmann, S.; Lohmann, J.F.; Krammer, G.; Helm, C.; Aydin, B.; Hecht, M. A Bayesian EAP-based nonlinear extension of Croon and van Veldhoven's model for analyzing data from micro-macro multilevel designs. *Mathematics* **2022**, *10*, 842. [CrossRef]
23. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*, 7th ed.; Muthén & Muthén: Los Angeles, CA, USA, 2012.
24. Maas, C.J.M.; Hox, J.J. Sufficient sample sizes for multilevel modeling. *Methodology* **2005**, *1*, 85–91. [CrossRef]
25. Hox, J.J. *Multilevel Analysis: Techniques and Applications*; Routledge: New York, NY, USA, 2010.
26. Bollen, K.A. *Structural Equations with Latent Variables*; Wiley: New York, NY, USA, 1989.
27. Robitzsch, A.; Oberwimmer, K. Package 'BIFIEsurvey'. 2022.
28. Kolenikov, S. Resampling variance estimation for complex survey data. *Stata J.* **2010**, *10*, 165–199. [CrossRef]
29. Liang, K.Y.; Zeger, S.L. data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22. [CrossRef]
30. Cheong, Y.F.; Fotiu, R.P.; Raudenbush, S.W. Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. *J. Educ. Behav. Stat.* **2001**, *26*, 411–429. [CrossRef]
31. R Development Core Team. *R: A Language and Environment for Statistical Computing*; The R Foundation: Ames, IA, USA, 2016.
32. Muthén, L.K.; Muthén, B.O. How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model.* **2002**, *9*, 599–620. [CrossRef]
33. Fox, J.P. Multilevel IRT Modeling in practice with the package mlirt. *J. Stat. Softw.* **2007**, *20*, 1–16. [CrossRef]

34. Lohmann, J.; Zitzmann, S.; Voelkle, M.; Hecht, M. A primer on continuous-time modeling in educational research: An exemplary application of a continuous-time latent curve model with structured residuals (CT-LCM-SR) to PISA data. *Large-Scale Assess. Educ.* **2022**, *10*, 5. [CrossRef]
35. Hecht, M.; Gische, C.; Vogel, D.; Zitzmann, S. Integrating out nuisance parameters for computationally more efficient Bayesian estimation—An illustration and tutorial. *Struct. Equ. Model.* **2020**, *27*, 483–493. [CrossRef]
36. Lüdtke, O.; Marsh, H.W.; Robitzsch, A.; Trautwein, U. A 2×2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychol. Methods* **2011**, *16*, 444–467. [CrossRef]
37. Zitzmann, S.; Lüdtke, O.; Robitzsch, A.; Marsh, H.W. A Bayesian approach for estimating multilevel latent contextual models. *Struct. Equ. Model.* **2016**, *23*, 661–679. [CrossRef]
38. Zitzmann, S.; Weirich, S.; Hecht, M. Using the effective sample size as the stopping criterion in Markov chain Monte Carlo with the Bayes Module in M*plus*. *Psych* **2021**, *3*, 336–347. [CrossRef]
39. Preacher, K.J.; Zhang, Z.; Zyphur, M.J. Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychol. Methods* **2016**, *21*, 189–205. [CrossRef] [PubMed]