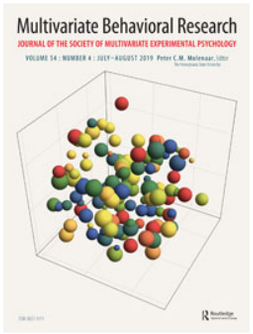

Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Analysis of Variance Models with Stochastic Group Weights

Axel Mayer & Felix Thoemmes

To cite this article: Axel Mayer & Felix Thoemmes (2019) Analysis of Variance Models with Stochastic Group Weights, Multivariate Behavioral Research, 54:4, 542-554, DOI: 10.1080/00273171.2018.1548960

To link to this article: https://doi.org/10.1080/00273171.2018.1548960

Published online: 20 Jan 2019.

Submit your article to this journal

Article views: 51

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hmbr20

MULTIVARIATE BEHAVIORAL RESEARCH
2019, VOL. 54, NO. 4, 542–554
https://doi.org/10.1080/00273171.2018.1548960

Routledge
Taylor & Francis Group

Check for updates

Analysis of Variance Models with Stochastic Group Weights

Axel Mayer[a] and Felix Thoemmes[b]

[a]RWTH Aachen University; [b]Cornell University

ABSTRACT

The analysis of variance (ANOVA) is still one of the most widely used statistical methods in the social sciences. This article is about stochastic group weights in ANOVA models – a neglected aspect in the literature. Stochastic group weights are present whenever the experimenter does not determine the exact group sizes before conducting the experiment. We show that classic ANOVA tests based on estimated marginal means can have an inflated type I error rate when stochastic group weights are not taken into account, even in randomized experiments. We propose two new ways to incorporate stochastic group weights in the tests of average effects – one based on the general linear model and one based on multigroup structural equation models (SEMs). We show in simulation studies that our methods have nominal type I error rates in experiments with stochastic group weights while classic approaches show an inflated type I error rate. The SEM approach can additionally deal with heteroscedastic residual variances and latent variables. An easy-to-use software package with graphical user interface is provided.

KEYWORDS

Adjusted means; analysis of variance; average effects; EffectLiteR; least square means; main effects; marginal means; stochastic group weights

In experimental or quasi-experimental designs with one or multiple categorical predictors the classic analysis of variance (ANOVA) model is still the data-analytic method of choice for many researchers in the social sciences. The method is implemented in all major statistical software packages and its mathematical foundations have remained mostly unchanged since Sir Ronald Aylmer Fisher introduced the method in the beginning of the twentieth century (see e.g., Fisher, 1925). The standard procedure to use ANOVA methods involves specifying the statistical model, the computation of overall hypothesis tests for main effects and interactions using a sums of squares table, and the computation of contrasts based on marginal means (Searle, Speed, & Milliken, 1980) which have also been called least square means in earlier research (Goodnight & Harvey, 1978). In this article, we consider experiments with randomized and nonrandomized treatment assignment, a second categorical predictor and interactions between the two variables. The design is not necessarily balanced which is very common in the social and behavioral sciences (Keselman, Huberty, & Lix, 1998). In such designs, the tests for interactions are unproblematic from both an interpretative and a computational point of view, but methodologists frequently caution against interpreting main effects, because they depend on coding of predictors, type of sum of squares, and in general are considered to be not informative when there are interactions. However, we argue that many researchers are in fact interested in the average effect of a treatment. It is interesting to know if a treatment is effective *on average*, i.e., averaged over the distribution of covariates, even if there is an interaction present. To give a precise example, we may consider a treatment, e.g., to reduce weight. Such a treatment could be *on average* effective, meaning that it reduces weight in participants, at around 10 lbs for the whole course of the treatment. At the same time, it could be true that it reduces weight in males by 12 lbs and for females by 8 lbs. Even though this interaction is interesting, it is still meaningful (and correct) to state that there is weight reduction *on average*. Such statements about average effects are especially important for policy makers who need to know if a treatment they plan to implement on a large scale is effective in addition to learning about interactions, because often they may not even be able to assess these potential moderators.

Average effects are well-defined in the causal inference literature (e.g., in the Neyman-Rubin causal model;

CONTACT Axel Mayer axel.mayer@rwth-aachen.de Department of Psychological Methods, Institute of Psychology, RWTH Aachen University, Jaegerstr. 17/19, D-52066 Aachen, Germany.

Supplemental data for this article can be accessed here.

© 2018 Taylor & Francis Group, LLC

Holland, 1986; Rubin, 1974; Splawa-Neyman, 1990) and are oftentimes the most important quantity in randomized controlled trials. To estimate average effects in an ANOVA framework, researchers need to specify the statistical model, compute the estimated means in each cell based on the model parameters, and then compute the estimated marginal means as a weighted sum of the estimated cell means. If weights are used which are proportional to the frequencies of the factor combinations that are averaged over, the difference between two marginal means corresponds to the average treatment effect.[1] The more frequently used Type III sums of squares (which are the default in some popular statistical software programs) use equal weights across all cells to estimate marginal means. But no matter how we compute average effects, or classic type I, II, III sums of squares main effects, or other weighted combinations of cell means, we always implicitly or explicitly end up using weights in these computations. These weights may sometimes be known, but oftentimes they may also be estimated from observed data. In fact, in standard ANOVA methods, these weights are usually always estimated, but critically, the *uncertainty* in estimating the weights is *not* taken into account.

In this article, we explore consequences of treating weights as known and fixed when they are in fact unknown and stochastic[2] and present two different ways of how to incorporate stochastic group weights into our tests for average effects. We first review the estimation of cell means and marginal means in simple ANOVA designs, then present two statistical models with stochastic group weights which we evaluate and contrast to existing methods using simulation studies. Consequences for different research designs and corresponding analysis methods are discussed.

Fixed versus stochastic group weights

Our focus is to elaborate on the role of stochastic group weights in the context of testing average effects with ANOVA models. To better understand what we mean by stochastic group weights, consider the following hypothetical scenario. Assume we want to investigate the effects of an educational intervention X on the mathematics ability Y of adolescents, while taking into account the school type K. Let $X=1$ denote the group receiving the intervention and let $X=0$ denote the control group. We consider three different school types: public schools $K=0$, private schools $K=1$, and vocational schools $K=2$. In standard ANOVA terminology, this is a 2×3 between subject design. Assume we know the population means and marginal probabilities which are depicted in Table 1.

Comparing the population means between the intervention group and the control group, we see that the mathematics training has a positive effect +10 in public schools and a negative effect −10 in private and vocational schools. The average effect is zero, because 50% of the population attend public schools and 50% attend private or vocational schools, so each of the two effects gets a weight of 0.5. Notice that this is a randomized experiment, and therefore X (intervention status) and K (school type) are independent, i.e., the cell probabilities are obtained as the product of the marginal probabilities.

Now consider two different researchers who both collect data with sample size $N=200$ and conduct an ANOVA to obtain tests for main effects, interactions, and contrasts based on marginal means. The first researcher is using a design with stochastic group weights and the second researcher is using a design with fixed group weights. Researcher 1 recruits a random sample of 200 adolescents, asks them which type of school they attend, and assigns each of them randomly to either the intervention group or the control group with probability 0.5, e.g., by flipping a coin for each participant. Of course, the cell probabilities in the sample are most likely not exactly equal to the population probabilities but they would typically be close. If researcher 1 would repeat the experiment, we would see some variability in cell probabilities in the samples but no systematic deviations from the population values. In the fully stochastic design used by researcher 1, both X and K are stochastic, i.e., the cell frequencies, the marginal frequencies of X, and the

Table 1. Population means and proportions in parentheses for the hypothetical educational intervention example.

	School type K			
	Public $K=0$	Private $K=1$	Vocational $K=2$	
Control group $X=0$	40 (0.25)	30 (0.05)	60 (0.2)	(0.5)
Intervention group $X=1$	50 (0.25)	20 (0.05)	50 (0.2)	(0.5)
	(0.5)	(0.1)	(0.4)	

[1] In some statistical software packages it is not easy to get such average effects (like in SPSS) but in others there exist convenient functions (e.g. R packages lsmeans or emmeans). To the best of our knowledge, EffectLiteR (Mayer & Dietzfelbinger, 2018) is the only software package that now allows for incorporating stochastic group weights.

[2] Notice that our distinction between fixed and stochastic group weights is not related to treating factors as fixed or random, which is sometimes emphasized in the ANOVA literature. In this literature, fixed factors are categorical variables with a fixed number of levels that are all observed in each sample, and random factors are categorical variables, such as the person variable or a school-ID, that have a large number of levels and the realized levels in the sample are a random sample of all possible levels and can therefore be different in each sample. In this sense, we consider fixed factors with stochastic group sizes in this paper.

marginal frequencies of K vary (slightly) from sample to sample. Researcher 2 happens to know the population probabilities and ensures that these are exactly the same in the sample by first recruiting individuals from each school type based on the known proportions and then randomly assigning exactly half of the individuals from each school type to the control and treatment condition, e.g., by drawing assignments out of an urn with the same number of intervention and control group assignments. So, in this sample, there are 100 adolescents from public schools (50 in the control group and 50 in the training group), 20 adolescents from private schools (10 in the control group and 10 in the training group), and 80 adolescents from vocational schools (40 in the control group and 40 in the training group). If researcher 2 would repeat the experiment, we would get identical group sizes and cell probabilities in all samples. In this fully fixed design, the group weights are fixed and therefore also the marginal frequencies of both X and K are fixed as well.

Notice that both researchers conduct a randomized experiment and data come from the same population. Both researchers have the same research question and they use the same statistical method. We argue that the design with stochastic group sizes used by researcher 1 is far more common in social research, because the true population probabilities are rarely known and oftentimes not all predictors are under full control of the experimenter. Nevertheless, the ANOVA method that assumes fixed group weights is frequently used to estimate effects in designs with stochastic group weights.

Adjusted means and marginal means

Before we consider statistical models with stochastic group weights, we briefly introduce the definitions of the main concepts that were described in the previous sections. Considering again the hypothetical scenario with the mathematics training, Table 1 shows the population cell means $E(Y|X=x,\ K=k)$, the cell probabilities $\mathrm{P}(X=x,\ K=k)$, the marginal probabilities of the treatment $\mathrm{P}(X=x)$, and school type $\mathrm{P}(K=k)$. The adjusted means for the control group and the treatment group are defined as:

$$\begin{aligned} AdjM_0 &= E[E(Y|X=0,K)] \\ &= \sum_k E(Y|X=0,K=k)\cdot P(K=k) \\ &= 40\cdot 0.5 + 30\cdot 0.1 + 60\cdot 0.4 \\ &= 47 \end{aligned} \tag{1}$$

$$\begin{aligned} AdjM_1 &= E[E(Y|X=1,K)] \\ &= \sum_k E(Y|X=1,K=k)\cdot P(K=k) \\ &= 50\cdot 0.5 + 20\cdot 0.1 + 50\cdot 0.4 \\ &= 47 \end{aligned} \tag{2}$$

We define the adjusted mean in group $X=x$ as the unconditional expectation $E[E(Y|X=x,K)]$, so we consequently need to use the unconditional probabilities $\mathrm{P}(K=k)$ as weights for the cell means. In comparison, the default setting in many software programs for computing post-hoc contrasts is to use marginal means with equal weights. For our example, this gives

$$\begin{aligned} MM_0^{eq} &= \sum_k E(Y|X=0,K=k)\cdot\frac{1}{3} \\ &= 40\cdot\frac{1}{3} + 30\cdot\frac{1}{3} + 60\cdot\frac{1}{3} \\ &= 43.3\overline{3} \end{aligned} \tag{3}$$

$$\begin{aligned} MM_1^{eq} &= \sum_k E(Y|X=1,K=k)\cdot\frac{1}{3} \\ &= 50\cdot\frac{1}{3} + 20\cdot\frac{1}{3} + 50\cdot\frac{1}{3} \\ &= 40 \end{aligned} \tag{4}$$

To disambiguate what types of weights were used in the computation of the marginal means we use the superscript *eq* to indicate that equal weights are used. This is in line with the definition of (population) marginal means (PMM) suggested by Searle, Speed, and Milliken (1980), who then introduce the widely used term *estimated marginal means* (EMM) for estimates of these marginal means. We avoid the prefix "population" because we think it is misleading: The MM_x^{eq} would refer to a hypothetical population with a uniform distribution of the second categorical variable and not necessarily to the actual population of interest with a potentially different distribution of the second categorical variable. Some software packages allow the user to specify the set of weights used in the computation of marginal means (e.g., the R package emmeans, Lenth, 2018). When proportional weights are specified, the marginal means MM_x^{pro} are identical to the adjusted means as defined above and we can use standard software to obtain estimates for the adjusted means as we show in Appendix A (supplementary materials).

Average effects and marginal effects

Usually we are not only interested in the (conditional) expectations in treatment groups but also in effects, i.e., differences between such conditional expectations.

The average effect of treatment $X=1$ compared to $X=0$ is defined as the difference between the two adjusted means:

$$\begin{aligned} AVE_{10} &= AdjM_1 - AdjM_0 \\ &= E[E(Y|X=1,K)] - E[E(Y|X=0,K)] \\ &= 47 - 47 \\ &= 0. \end{aligned} \tag{5}$$

Equivalently, we can write the average effects as the unconditional expectation of the conditional effects:

$$\begin{aligned} AVE_{10} &= E[E(Y|X=1,K) - E(Y|X=0,K)] \\ &= \sum_k CE_{10}(K=k) \cdot P(K=k) \\ &= 10 \cdot 0.5 - 10 \cdot 0.1 - 10 \cdot 0.4 \\ &= 0 \end{aligned} \tag{6}$$

where $CE_{10}(K=k)$ stands for the $(K=k)$-conditional treatment effect defined as

$CE_{10}(K=k) = E(Y \mid X=1, K=k) - E(Y \mid X=0, K=k)$.

In a similar vein, we introduce terms for the differences between marginal means, also keeping the superscript that refers to the weighting scheme used in the computation of marginal means. For equal weights, we get marginal effects defined as:

$$\begin{aligned} ME_{10}^{eq} &= MM_1^{eq} - MM_0^{eq} \\ &= -3.33. \end{aligned}$$

Similarly, for proportional weights we get:

$$\begin{aligned} ME_{10}^{pro} &= MM_1^{pro} - MM_0^{pro} \\ &= 0. \end{aligned}$$

The latter one is of course again equal to the average effect AVE_{10}, while the former is equal to the ANOVA main effect using Type III sums of squares. Notice that we avoid the less clear defined term *main effect* in the remainder of this article, because the definition of main effects differs between the different so-called types of sums of squares. Maxwell and Delaney (2004) nicely illustrate that main effects can be written as weighted sums of cell means and the different types of sums of squares use different weights. For an overview of the on-going controversy about which type of sums of squares should be used in which situation, see Hector, von Felten, and Schmid (2010).

Models with stochastic group weights

Having introduced the most important concept at the population level, we now turn to presenting two different statistical models with stochastic group weights that both can be used to estimate adjusted means as well as average effects based on a sample. They can also be used to obtain overall hypothesis tests as is common for ANOVA-like methods. The first model is based on the general linear model (GLM) where the group weights are added to the model. The second model is a special case of the general EffectLiteR model (Mayer, Dietzfelbinger, Rosseel, & Steyer, 2016) and uses maximum likelihood estimation for a multigroup structural equation model with stochastic group weights. Both models along with the computation of all parameters and hypothesis tests of interest are implemented in the R software package EffectLiteR (Mayer & Dietzfelbinger, 2018). The R code for analyzing the teaching intervention example is provided in Appendix A.

GLM with stochastic group weights

The GLM is one of the most often used statistical models and its parameters are typically estimated via ordinary least squares. Many statistical software packages use the GLM in the background for their ANOVA routines but oftentimes do not show the parameters by default. The basic equation for the GLM is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is a vector of N observations of the dependent variable, $\mathbf{X}$ is a design matrix with fixed constants, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is a vector of residuals. ANOVA, analysis of covariance (ANCOVA), and multiple regression are special cases of the GLM. For our analysis of the randomized mathematics training, we first need to choose a parameterization for the categorical predictors. In the ANOVA framework, it is suggested that researchers use a coding with orthogonal contrasts because the usual interpretation of type III sums of squares requires orthogonality. For the analysis of average effects, the coding scheme is not important as long as we choose a saturated parameterization for the GLM. Then all possible coding schemes yield identical results. A saturated parameterization means that we are estimating six regression coefficients (either using treatment contrasts, or effect contrasts or whatever contrasts we like) for six conditional expectations in our 2×3 between subject design. Then we can obtain estimates for the conditional expectations

$\hat{E}(Y|X=x,K=k)=\hat{\boldsymbol{\mu}}_{xk}$ as a function $f(\hat{\boldsymbol{\beta}})$ of the estimated $\hat{\boldsymbol{\beta}}$ coefficients. The estimates for the conditional expectations $\hat{\boldsymbol{\mu}}$ are the predicted values from the GLM. The variance–covariance matrix of $\hat{\boldsymbol{\mu}}$ denoted by $Var(\hat{\boldsymbol{\mu}})$ can be obtained by using the multivariate delta method based on the variance–covariance matrix of regression coefficients $Var(\hat{\boldsymbol{\beta}})=\sigma^2_{\epsilon,}(\mathbf{X}'\mathbf{X})^{-1}$. But in order to compute the adjusted means and the average effects, we also need the group weights. Point estimates of the group weights can easily be obtained by using the vector of sample proportions $\hat{\boldsymbol{p}}$ resulting in the vector of all parameters $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{p}}')'$. The variance–covariance matrix of the sample proportions is given by:

$$\mathrm{Var}(\hat{\boldsymbol{p}})=\left[\mathrm{diag}(\boldsymbol{p})-\boldsymbol{p}\cdot\boldsymbol{p}'\right]/N$$

where the diag() operator turns a vector into a diagonal matrix. Kiefer & Mayer (in press) show that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{p}}$ are independent and therefore the variance–covariance matrix of all parameters $\hat{\boldsymbol{\theta}}$ is block diagonal:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}})=\begin{bmatrix} Var(\hat{\boldsymbol{\beta}}) & \mathbf{0} \\ \mathbf{0} & Var(\hat{\boldsymbol{p}}) \end{bmatrix}$$

Based on the full parameter vector $\boldsymbol{\theta}$ and its variance–covariance matrix, we can estimate adjusted means, marginal means, average effects, and marginal effects while taking into account uncertainty in the estimation of sample proportions. For example, once we have estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{p}}$ the average effect is computed as shown in Equation 6 as

$$\hat{AVE}_{10}=\sum_k\left[\hat{\boldsymbol{\mu}_{1k}}-\hat{\boldsymbol{\mu}_{0k}}\right]\cdot\hat{\boldsymbol{p}}_k.$$

We can also use this information to compute overall hypothesis tests using either a large sample Wald χ^2 statistic (with asymptotic χ^2 distribution) or a finite sample F statistic (with approximate F distribution).

Multigroup SEM approach with stochastic group weights (EffectLiteR)

Another model with stochastic group weights that can be used for ANOVA-like methods is the EffectLiteR model as presented in Mayer et al. (2016). The EffectLiteR model uses a multigroup structural equation model for computation instead of the GLM. Using a SEM approach comes with several advantages: It allows for including latent variables, it can naturally handle heteroscedastic variances, and modern (robust) estimators and model fit measures are readily available. The complete model consists of a group-specific measurement model relating manifest variables $\boldsymbol{y}$ to latent variables $\boldsymbol{\eta}$, a group-specific structural model specifying structural relations among (latent) variables, and a part for the group weights (cf., Equations 1–3 in Mayer et al., 2016). The general multigroup structural equation model with stochastic group sizes is given by:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\nu}_g+\boldsymbol{\Lambda}_g\boldsymbol{\eta}+\boldsymbol{\epsilon} \\ \boldsymbol{\eta} &= \boldsymbol{\alpha}_g+\boldsymbol{B}_g\boldsymbol{\eta}+\boldsymbol{\zeta} \\ \log(n_g) &= \boldsymbol{\kappa}_g \end{aligned}$$

where $\boldsymbol{\nu}_g$ is a vector of measurement intercepts; $\boldsymbol{\Lambda}_g$ is a matrix of loadings; $\boldsymbol{\alpha}_g$ is a vector of structural intercepts; $\boldsymbol{B}_g$ is a matrix of structural coefficients; ϵ, is a vector of measurement error variables with zero mean vector; $\boldsymbol{\zeta}$ is a vector of structural residuals with zero mean vector; and $\boldsymbol{\kappa}_g$ is a parameter for the log-transformed expected group frequency n_g of group g.

For the purpose of this article, we do not need the measurement model because there are no latent variables, the group-specific structural model can be substantially simplified because there are no continuous covariates in the analysis and therefore, we just estimate the means and variances of Y in the six groups formed by combinations of values of X and K:

$$Y=\eta=\mu_{xk}+\zeta \text{ structural model for group } (X=x,K=k)$$

The model for the group weights in EffectLiteR is a saturated Poisson model as implemented in lavaan (Rosseel, 2012), a R (R Core Team, 2018) package for structural equation modeling:

$$\log(n_{xk})=\kappa_{xk} \quad \text{group sizes for group } (X=x,K=k)$$

We chose a Poisson model with the canonical log link because it leads to a relatively easy-to compute likelihood and does not require constrained optimization. Similar models using a multinomial model for group sizes is available in Mplus (Muthén & Muthén, 1998–2012) via the KNOWNCLASS option or can be implemented in a Bayesian framework (Mayer, Umbach, Flunger, & Kelava, 2017). The parameters of the structural model and the model for group weights are estimated simultaneously using maximum likelihood, where the full likelihood consists of a part for the structural equation model and a part for the group weights (see Mayer et al., 2016). The variance–covariance matrix of all parameter estimates can be obtained following standard maximum-likelihood theory. Using the EffectLiteR approach also provides us with estimates for the cell means and proportions which then can be used to compute adjusted means, marginal means, average effects, marginal effects, and overall hypothesis tests just like with the GLM with

stochastic group weights but using a more general approach with a different estimator. Using the parameters of the model, the average effect is again estimated like in Equation 6:

$$\hat{AVE}_{10} = \sum_k [\hat{\mu_{1k}} - \hat{\mu_{0k}}] \cdot \hat{p_k}$$

where $\hat{p_k}$ is computed based on the $\hat{\boldsymbol{\kappa}}$ parameters applying standard computational rules for probabilities.

Simulation study

We conducted various simulation studies in order to evaluate the statistical properties and especially the type I error rates of the classic and the newly proposed estimators. Our main interest is to investigate the consequences of using a method that assumes fixed group weights when they are in fact stochastic and vice versa. To illustrate our main point, we start with a small-scale simulation that can easily be reproduced by the interested reader within a few minutes. The full R code for the simulation is given in Appendix B (supplementary materials). Later we further investigate the role of sample size, unequal group weights, size of the interaction, number of cells, heteroscedastic residual variances, nonrandomized designs, and different estimators in five more extensive additional simulations studies.

Design

In our small-scale simulation, we use a balanced 2×2 design with randomized treatment assignment. The marginal probabilities and the conditional expectations are given by:

	$K=0$	$K=1$	
$X=0$	24	26	(0.5)
$X=1$	36	14	(0.5)
	(0.5)	(0.5)	

The standard deviation of Y in each cell is 10 and the sample size is $N=200$. Notice that the average effect of X in this example is equal to zero and there is an interaction between X and K. The effect size measures eta-square and partial eta-square for the interaction term are $\eta^2_{X:K} = 0.22; \eta^2_{\text{par};X:K} = 0.27$. For the partial eta-square, the main effects of X and K are partialled out. Data were generated in two different ways: For the data generating mechanism with stochastic group sizes, there is some variability in the cell proportions over repeated sampling. This mimics a randomized experiment as described for Researcher 1 in the introduction, in which the marginal frequencies are stochastic. In contrast, for the data generating mechanism with fixed group sizes it is ensured that the cell proportions are always exactly equal to the population proportions. This follows the reasoning of an experiment as conducted by Researcher 2 in the introduction, in which the marginal frequencies are fixed. Other than that the two data generating mechanisms are identical. The data, generated either as fixed or as stochastic, are analyzed using both a linear model with fixed group weights and a linear model with stochastic group weights. We use 2000 replications for our simulation and focus on the inference for the average effect of X.

Results

Not surprisingly, the two linear models give identical point estimates and show negligible absolute bias around 0.06 for all conditions. However, the two models differ with regard to statistical inference: When the data are generated as stochastic, but the analysis model uses fixed group weights, we find that the standard errors are deflated (indicated by a relative standard error bias of −15%), the type I error rate is too high (around 9% instead of the nominal 5%), and the coverage is too low (only 90.6% of the confidence intervals include the true average effect of zero instead of the nominal 95%). The opposite pattern evolves, when the data are generated as fixed and the analysis model uses stochastic group weights. In this case, the standard errors are inflated (indicated by a relative standard error bias of +18%), the type I error rate is too low (around 2% instead of the nominal 5%), and the coverage is too high (98.0% of the confidence intervals include the true average effect of zero instead of the nominal 95%). When the data generating mechanism matches with the analysis model, i.e., fixed data are analyzed with a model that uses fixed group weights and stochastic data are analyzed with a model that uses stochastic group weights, the inference is as expected: The standard errors show negligible relative bias under 2%, type I error rate is at the nominal level of 5%, and the coverage is also at the nominal level of 95%. In sum, this simulation illustrates that erroneously using a model with fixed group weights when the data is stochastic is too liberal, while erroneously using a model with stochastic group weights when the data is fixed is too conservative.

Additional simulations

In five extensive additional simulations studies, we further investigate under which conditions the mismatch

Table 2. Overview of the five extensive simulation studies.

	Values				
Parameter	Simulation 1	Simulation 2	Simulation 3	Simulation 4	Simulation 5
Design	2×3 Design	2×2 Design	$3 \times 2 \times 2$ Design (same as 3×4 Design)	2×2 Design	2×2 Design
Group weights	Fixed, stochastic	Fixed, stochastic	Fixed, stochastic	Fixed, stochastic	Fixed, stochastic
Interaction	No, small, medium, large, very large	No, small, large, very large	(Varied implicitly as a function of the standard deviation of ε)	No, small, large, very large	No, small, large, very large
Sample size N	100, 200, 500	200	200	200	200
Marginal frequencies K	50/10/40	50/50, 60/40, 80/20, 90/10	30/20/20/30	50/50	50/50
Marginal frequencies X	50/50	50/50	30/40/30	50/50, 80/20	50/50
Cor(X, K)	0	0	0	0	0, 0.2, 0.6
Standard deviation ε	10	10	5, 10, 20, 40	10 (homog.), 8 ($X = 0$) and 12 ($X = 1$) (heterog.)	10
True average effect	0 ($X = 1$ vs. $X = 0$)	0 ($X = 1$ vs. $X = 0$)	0 ($X = 1$ vs. $X = 0$) 0 ($X = 2$ vs. $X = 0$)	0 ($X = 1$ vs. $X = 0$)	0 ($X = 1$ vs. $X = 0$)
Evaluation criterion	Type I error rate	Type I error rate	Type I error rate	Type I error rate	Type I error rate
Nrep per cell	1000	1000	1000	1000	1000
Conditional expectations	See Table C1	See Table D1	See Table E1	See Table F1	See Table G1
Models	elr_sem_sto, elr_sem_fix, elr_lm_sto, elr_lm_fix, emmeans	elr_sem_sto, elr_sem_fix, elr_lm_sto, elr_lm_fix, emmeans	elr_sem_sto, elr_sem_fix, elr_lm_sto, elr_lm_fix, emmeans	elr_sem_sto, elr_sem_fix, elr_lm_sto, elr_lm_fix, emmeans	elr_sem_sto, elr_sem_fix, elr_lm_sto, elr_lm_fix, emmeans

Note. Five models were analyzed in all five simulation studies: A multigroup structural equation model with stochastic group weights (elr_sem_sto), a multigroup structural equation model with fixed group weights (elr_sem_fix), a general linear model with stochastic group weights (elr_lm_sto), a general linear model with fixed group weights (elr_lm_fix), and estimated marginal means which are also based on the general linear model with fixed group weights (emmeans). The first four models were estimated with the R package EffectLiteR (Mayer & Dietzfelbinger, 2018) and the estimated marginal means were tested with the R package emmeans (Lenth, 2018).

between data generating mechanism and statistical analysis model yields a substantial bias. Therefore, we additionally manipulate the size of the interaction, sample size, unbalancedness of group weights, the number of cells, the structure of residual variances, and the correlation between the factors. Table 2 gives an overview of the parameters used in the five simulation studies.

We analyze the data with five different models: Estimated marginal means with proportional weights (an analysis that always assumes fixed margins), GLM with and without stochastic group weights, and multigroup structural equation models with and without stochastic group weights. For the structural equation modeling approaches, we use Wald χ^2 tests and for the other approaches we use approximate F tests to test the null hypothesis of no average treatment effects. All models can be estimated in the R programming language (R Core Team, 2018) using different packages: For estimated marginal means, we use the emmeans package (Lenth, 2018) which is the successor of the lsmeans package (Lenth, 2016), and for the newly proposed methods with and without stochastic group weights we use the EffectLiteR package (Mayer & Dietzfelbinger, 2018). In all conditions, the average effect of the treatment X is equal to zero and since all the evaluation criteria in the simulations lead to similar conclusions, we only report type I error rates. We briefly report the most important findings for all five simulation studies and provide the details and the results in Appendices C–G (supplementary materials).

Simulation 1

Simulation 1 is based on a randomized, unbalanced 2×3 design and focuses on the effects of interaction size and sample size on type I error rates (see Appendix C for details). The cell means are constructed in such a way that the adjusted means remain equal across conditions and just the size of the interaction is varied. As in our small-scale simulation, we find that the models that erroneously assume fixed group weights have an inflated type I error rate. This finding crucially depends on the size of the interaction: For medium interactions ($\eta^2_{X:K} = 0.14; \eta^2_{\text{par};X:K} = 0.39$), the type I error rate is around 10%, for large interactions ($\eta^2_{X:K} = 0.27; \eta^2_{\text{par};X:K} = 0.59$), the type I error rate is around 20%, and for very large interactions ($\eta^2_{X:K} = 0.39; \eta^2_{\text{par};X:K} = 0.72$), the type I error rate is around 30%. The inflation of type I error rates is negligible in conditions with no or small interactions. Correctly specified models have nominal type I error rates in all conditions. Interestingly, these findings are totally independent of sample size. We also do not find differences between using Wald χ^2 tests as for the SEM based models or F tests as for the GLM based models.

Simulation 2

Simulation 2 is based on a randomized 2×2 design and focuses on the effects of unbalancedness on type I error rates (see Appendix D for details). Unbalancedness is manipulated by using different marginal group proportions for K ranging from balanced (50/50) to highly unbalanced (10/90). Since the interaction size turned out to be a crucial parameter, we also manipulated the size of the interaction in simulation 2. The cell means are again constructed in such a way that the adjusted means remain equal across conditions and that the average effect of X is zero in all conditions. The key findings are the same as in our previous simulations and are not repeated here. Interestingly, the amount of unbalancedness is irrelevant for the inflation of type I error rates. The only significant factor is again the size of the interaction.

When comparing the results from Simulations 1 and 2, we find that the inflation of type I error rates are considerably higher in Simulation 1. For example, in the very large interaction condition, the type I error rate in Simulation 1 is around 30%, whereas it is around 13% in Simulation 2. There are two key differences between the simulation studies that could cause this discrepancy: First, it could be due to the number of values of K. Since we need to average over K, more group weights enter the computation of the average effect in Simulation 1. Second, the average effect of K is higher in Simulation 1, leading to a higher amount of overall variance explained (R^2) and therefore to much greater effect sizes of the interaction in terms of partial η^2. To figure out which causes the discrepancy in type I error rates between Simulation 1 and 2, we conducted a third simulation study with more levels of K, in which we also manipulated the average effect of K.

Simulation 3

Simulation 3 is based on a randomized $3 \times 2 \times 2$ design and focuses on the consequences of the average effect of K on type I error rates for the average effect of X (see Appendix E for details). Notice that the overall amount of variance explained in the outcome (R^2) changes as well by manipulating the average effect of K. In contrast to previous simulations, X can take on three values and the reported tests for average effects are joined tests that there is no average effect of $X = 1$ vs. $X = 0$ and there is no average effect of $X = 2$ vs. $X = 0$. Since we focus on the effects of X, we need to average over the 2×2 combinations of values of the other two categorical variables K_1 and K_2. This is computationally equivalent to a 3×4 design with a K variable that can take on four values. The data are generated in such a way that the absolute size of the interaction and the conditional effects of X remain the same across a low R^2 and high R^2 condition. In the low R^2 condition, there is no average effect of K while there is a strong average effect of K in the high R^2 condition. Because the absolute size of the interaction and the residual error variance remain the same, the partial η^2 for the interaction term is the same in both low and high R^2 conditions, but the η^2 for the interaction term can be quite different between low and high R^2 conditions. The size of the interaction is manipulated indirectly by using different values for the residual standard deviation SD(ε).

The results clearly illustrate that the key factor for the inflation of type I error rates is the partial η^2. Within the large interaction condition, the type I error rate is almost identical for the low and high R^2 conditions, even though the η^2 is much higher for the high R^2 condition. However, the partial η^2 is identical for both conditions which explains that they have almost identical type I error rates. The number of levels of K is not important as can be seen by comparing results across the simulations: In conditions with similar partial η^2 values we always find comparable type I error rates, no matter if we average over two (as in Simulation 2), three (as in Simulation 1) or four (as in Simulation 3) values of K.

Simulation 4

Simulation 4 uses a 2×2 design like Simulation 2. Its main focus is on comparing our first proposed approach based on the general linear model to our second approach based on the multigroup SEM (see Appendix F for details). The SEM-based approach uses a multigroup model and can easily incorporate unequal residual variances across groups. Therefore, we expect to see an advantage of the SEM approach in conditions with heteroscedastic residual variances. The considered scenarios are similar to those in Simulation 2, but instead of manipulating the marginal group proportions of K, which did not make a difference, we manipulated the marginal group proportions of X (with two conditions, 50/50 and 20/80), because it is well-known in the literature that violation of homoscedasticity is critical with respect to the standard error of estimated regression coefficients when group sizes are unequal (e.g., Berry, 1993; Kroehne, 2009). In the heteroscedastic variances condition, the standard deviation of the residuals in group $X = 0$ was 8 and in the group $X = 1$

it was 12, while the standard deviation of the residuals was 10 in all groups in the homoscedastic condition.

The results clearly show that only the SEM-based approaches yield nominal type I error rates in the conditions with unequal group sizes and heteroscedastic residual variances. The GLM-based approaches have way too low type I error rates in these conditions. Of course, the SEM approach with stochastic group sizes is best in the conditions with stochastic data generation and the SEM approach with fixed group sizes is best in the conditions with fixed data generation, especially in the large interaction conditions. Interestingly, the standard GLM based approach with fixed group sizes also has nominal type I error rate in the condition in the bottom right corner in Figure F1 (very large interaction, unequal group sizes, stochastic group sizes, heteroscedastic residual variances). This is a coincidence and only shows up because the two misspecifications in the GLM cancel each other out exactly in this condition: While the erroneous assumption of fixed group sizes leads to an inflated type I error rate, the erroneous assumption of homogeneous residual variances leads to a deflated type I error rate.

Simulation 5

Simulation 5 also uses a 2×2 design like Simulations 2 and 4. Its main focus is on generalizing our previous results to nonrandomized research designs. We keep the marginal group proportions of both X and K equal but modified the correlation between the two factors (no, small, large): In the randomized condition there is no correlation between X and K. The cell probabilities are the product of the marginal probabilities and are therefore 0.25 each. In the conditions with small and large correlations between X and K, the cell probabilities are no longer obtained as the product of the marginal probabilities. We chose cell probabilities in such a way that they result in a correlation of 0.2 (small correlation) and a correlation of 0.6 (large correlation). The details are shown in Table G1. In addition, we modified the data generation (fixed and stochastic) and the size of the interaction.

The results show a similar pattern as in the previous simulations (see Figure G1): The methods using fixed group weights have nominal type I error rates in the conditions with fixed data generation and the methods using stochastic group weights have nominal type I error rates in the conditions with stochastic data generation. And this pattern does not change in nonrandomized conditions, i.e., is independent of the correlation between X and K. Consequently, the issue of stochastic group weights does not only show up in randomized experiments but also in nonrandomized experiments in the same manner.

Discussion

In this article, we discussed the issue of stochastic group sizes in factorial ANOVA designs. Even though stochastic group sizes are probably the norm, rather than the exception, in social science research, they are typically ignored in the data analysis. In fact, Keselman et al. (1998) show in their review that 72% of the between-subject factorial designs in educational research are unbalanced. Yet, the standard ANOVA approach is used almost exclusively, and we argue that there is essentially no awareness of this issue in the psychological community. This approach, as we have explained, does not by default integrate the uncertainty that arises from stochastic cell sizes.

Using a simple simulation study, we have shown that ignoring the uncertainty in the sampling process, leads to a negative bias in standard errors, and resulting from this to type I error inflation, and coverage intervals that are lower than their nominal level. In further simulation studies, we were able to demonstrate that the magnitude of the interaction is a critical component as to how much bias is introduced. Generally speaking, the larger the interaction in the factorial design, the more bias will be introduced when the stochastic nature of the sampling procedure is ignored. An interesting aspect of these simulations studies is that this type of bias also happens in randomized studies, and is not much influenced by sample size, or the particular design (e.g., number of groups, or imbalance of the design).

We have proposed two new ways that allow to incorporate stochastic sampling plans, either using GLM methods that include the information on stochastic group sizes, or using a maximum likelihood-based multigroup SEM approach. Both of these approaches account for the uncertainty in the sampling process, and yield unbiased estimates of the standard errors, and therefore nominal type I error and coverage rates. The SEM-based approach can additionally deal with heteroscedastic residual variances and latent variables. Our recommendation therefore follows naturally, that we urge researchers who use stochastic group sizes, to rely on methods that estimate the uncertainty that comes with these methods, and provide trustworthy inferential statistics.

Generalizations

Nonrandomized experiments: In this article, we focused on randomized experiments and briefly showed in Simulation 5 that the issue of stochastic group weights generalizes to nonrandomized experiments. Stochastic group weights can be present in all types of studies in which researchers want to estimate average or conditional effects, while taking into account interactions between a cause and categorical covariates. The proposed procedure of estimating the stochastic group weights and incorporating them in the computation of average effects is directly applicable in nonrandomized settings as well. However, in nonrandomized experiments there can be a stochastic dependency between the treatment variable and the other categorical (and continuous) variables. Therefore, considering additional categorical and continuous variables in the analysis is essential not only for gaining efficiency but also to obtain unbiased estimates for the average effects in such studies. In order for the average effects to have a causal interpretation in nonrandomized experiments, additional causality assumptions such as strong ignorability (Rosenbaum & Rubin, 1983) or another causality condition needs to be fulfilled (see also Imbens, 2004; Steyer, Gabler, von Davier, & Nachtigall, 2000).

Continuous covariates: We looked at experimental designs with categorical covariates only and did not consider examples with continuous covariates. However, the proposed approach can be generalized to such models including continuous covariates. A simple way to incorporate continuous covariates in the analysis is to add grand-mean centered continuous covariates (Aiken & West, 1996) as regressors either to the GLM-based approach or to the SEM-based approach and then proceed as suggested in this article. However, mean centering the continuous covariates again treats the mean of the covariates as fixed and known and ignores uncertainty in the estimation of the population mean. Just like ignoring uncertainty in the estimation of group weights, this may result in an inflated type I error rate for the average effects (Kroehne, 2009; Liu, West, Levy, & Aiken, 2017; Sampson, 1974). Instead, estimates for the mean of the continuous variable can be added along with a standard error and can then take into account for computing average effects. Mayer et al. (2016) fully describe an example with a nonrandomized 3×2 design and a continuous (latent) covariate and also provide general formulas for computing average effects in such designs.

Comparisons to other methods

Post-stratification: A statistical technique that is closely related to the present article is post-stratification (Cochran, 1977; McHugh & Matts, 1983). In post-stratification, the researcher stratifies persons based on a pretreatment variable, estimates treatment effects within the strata and then uses a weighted average of these strata estimates for the overall average treatment effect estimate (Gelman & Little, 1997; Miratrix, Sekhon, & Yu, 2013). The definition of the average treatment effect also originates from the causal inference literature and is identical with the definition we used in the present paper. However, the rationale for deriving standard errors and confidence intervals for the average effect is different: The post-stratification literature uses the Neyman-Rubin causal model (Holland, 1986) to derive expressions for the standard errors of the sample average treatment effect (SATE) and the population average treatment effect (PATE) based on potential outcomes (Imbens & Rubin, 2015; Miratrix, Sekhon, & Yu, 2013; Splawa-Neyman, (1990)). In contrast, we build on the ANOVA and regression literature to derive standard errors for the average effect based on the estimated regression coefficients and group sizes. Miratrix and colleagues claim that their "[…] post-stratified estimator is identical to a fully saturated ordinary linear regression with the strata as dummy variables and all strata-by-treatment interactions—i.e. a two-way analysis-of-variance analysis with interactions." (p. 382). So, their post-stratified estimator for the SATE is identical to our GLM-based estimator with fixed group sizes and yields valid inferences when the group sizes are fixed – or, equivalently, when only the sample average effect is of interest. In the settings, we considered the main interest is not in the SATE but in generalizing to a population, i.e., in estimating the PATE. In this case, we presume that our approaches with stochastic group sizes are very similar (if not identical) to the post-stratified estimator for the PATE (Imbens, 2011; Miratrix et al., 2013), even though the derivations and formulas are very different.

Weighted effects coding: In situations where the proportions of cases in each group in the sample can be considered to represent the corresponding proportion of cases in the population, Cohen, Cohen, West, and Aiken (2003) recommend using weighted effects coding (see Chapter 8.4 on page 328ff). Weighted effects coding is straightforward to apply when there is only a single factor. In this case, the regression coefficients represent the difference between the group means and the weighted (overall) mean. However, weighted effects coding becomes much more complicated when there are continuous or further categorical variables in the model, because interaction terms cannot simply be created by multiplying the values of the two variables that make up the interaction (Nieuwenhuis, te

Grotenhuis, & Pelzer, 2017). Our approach differs from weighted effects coding in multiple ways: First, we define our effects by comparing the groups with a reference group and not with the weighted overall mean. Second, we provide clear definitions of average effects in the presence of interactions based on conditional expectations and show how they can be computed. Third, we take uncertainty in the estimation of group weights into account which is not done in weighted effects coding.

Stochastic regression: In stochastic regression, the regressors are modeled as stochastic and the joint distribution of the outcome and the regressors is considered (as opposed to the conditional distribution of the outcome in ordinary least-squares regression). For regressions with identity link function, the estimation and inference for the regression coefficients are identical for both approaches (see, e.g., Casella, G., & Berger, R. L. (2002). Statistical inference. 2nd Edition. Pacific Grove, CA: Duxbury.). While there is no difference for the regression coefficients, the key aspect of the present manuscript is the (subsequent) analysis of average or main effects. And for this subsequent analysis it can make a difference whether we treat the regressors (group weights) as fixed or as stochastic as we show in our simulations. So, we need to distinguish between treating the regressors as fixed/stochastic in the first step (i.e., when estimating the regression coefficients; this is what stochastic regression is concerned about) and in the second step (i.e., when estimating the average effects; this is the focus of the present article).

Limitations

Our study tried to argue for the general point of being mindful of stochastic group sizes, but there are some nuances that we did not explore in this article. We only considered designs in which all factors had fixed samples sizes, or in which all factors had stochastic sample sizes. Clearly, there can be situations in which one factor has fixed sample sizes, and the other factor has stochastic sample sizes, which we could refer to as mixed sample size design. For example, if researcher 2 in our example in the article would recruit individuals from each school type based on the exact known proportions and would then randomly assign individuals using a coin toss to the control and treatment condition, then X would be stochastic and K would be fixed. In such a mixed design, we need to think about the nature of the variable that we average over for computing the average effect. So, when we compute the average effect of X, we average over the distribution of fixed school type K and therefore can use fixed group weights. When we compute the average effects of K (e.g., for $K=1$ vs $K=0$ and $K=2$ vs. $K=0$), we average over the stochastic X and consequently need stochastic group weights. We did not consider these designs directly, but based on our results, we strongly conjecture that ignoring the stochastic nature of one factor, would yield similar issues to the one that we described in the article. For more complex experimental designs that eventually have structurally empty cells or other distinct features, we also need to carefully think about the weighting scheme we would like to use and whether the weights are stochastic or fixed.

In this article, we neither fully discussed the causal interpretation and the causal definitions for the effects, nor the causality conditions that are required for obtaining causal effects. However, the definitions that we used for average effects and adjusted means originate from the causal inference literature (e.g., Imbens & Rubin, 2015; Pearl, 2009; Rubin, 1974; Steyer, Mayer, & Fiege, 2014). Given that there are no unobserved confounders, the average effect is the unconditional expectation of the individual causal effects and the adjusted means are the expectations of the individual potential or true outcomes in each treatment condition.

Our focus in the presentation of our approaches was on studies in which the sample is a random sample from the population. In survey research where more complicated sampling procedures are used, the sampling design needs to be taken into account not only for the estimation of regression coefficients but also for the estimation of the group weights. Otherwise we get biased estimates for the group weights. Oftentimes the population proportions (from the Census or alike) are used to determine the sampling scheme (e.g., oversampling of minorities) in a large-scale survey. In this case, we recommend directly using the population proportions and treating them as fixed, instead of using the biased estimates from the survey. An alternative is to correct the biased estimates and recompute the population proportions taking into account the sampling scheme.

As we have illustrated throughout the article, there are many situations where stochastic group weights are more meaningful and we believe that they are an important addition to the statistical toolbox of researchers. However, there are also situations where fixed group weights are more appropriate. In balanced designs, the group sizes have usually been fixed by the experimenter and can therefore be considered fixed.

Also, in cases where the unbalancedness is not per se meaningful but results from dropout or other difficulties in the data selection process, fixed group sizes can be reasonable. In such cases, the researcher can also consider to use (fixed) equal group weights instead of the observed unequal group weights. Another situation where fixed group weights are adequate is when there is information on the true marginal probabilities of the categorical variables that we average over. For instance, when we know, e.g., from Census data, the true proportions of male and females in a population of interest, we can choose to use these true proportions in our computations of adjusted means and average effects and consequently treat them as fixed. To conclude, there are designs where fixed group weights are more appropriate and there are designs where stochastic group weights are more appropriate and with this article we give researchers the opportunity to adequately address both types of designs in their analysis.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant MA 7702/1-1 from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

Aiken, L. S., & West, S. G. (1996). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage Publications

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed). New York: Wiley.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ, USA: Laurence Erlbaum Associates, Publishers.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, London: Oliver & Boyd.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, *23*, 127–135.

Goodnight, J. H., & Harvey, W. R. (1978). *Least squares means in the fixed effects general linear model*. SAS Institute, Incorporated.

Hector, A., von Felten, S., & Schmid, B. (2010). Analysis of variance with unbalanced data: An update for ecology & evolution. *Journal of Animal Ecology*, *79*, 308–316. doi: 10.1111/j.1365-2656.2009.01634.x

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29.

Imbens, G. W. (2011). Experimental design for unit and cluster randomized trials. In *International initiative for impact evaluation, Cuernavaca*

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. doi:10.1017/CBO9781139025751

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., …, Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386. doi: 10.3102/00346543068003350

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika*, *63*(2), 145–163. doi:10.1007/BF02294772

Kiefer, C., & Mayer, A. (in press). Average effects based on regressions with logarithmic link function: A new approach with stochastic covariates. *Psychometrika*.

Kroehne, U. (2009). *Estimation of average causal effects in quasi-experimental designs: Non-linear constraints in structural equation models* (Unpublished doctoral dissertation). Friedrich-Schiller-University, Jena, Germany.

Lenth, R. V. (2018). *emmeans: Estimated marginal means, aka least-squares means*, R package version 1.1. https://CRAN.R-project.org/package=emmeans

Lenth, R. V., (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 6. 1–33. doi: 10.18637/jss.v069.i01

Liu, Y., West, S. G., Levy, R., & Aiken, L. S. (2017). Tests of simple slopes in multiple regression models with an interaction: Comparison of four approaches. *Multivariate Behavioral Research*, 1–20.

Maxwell, S., & Delaney, H. (2004). *Designing experiments and analyzing data: A model comparison perspective* Mahwah, NJ: Lawrence Erlbaum.

McHugh, R., & Matts, J. (1983). Post-stratification in the randomized clinical trial. *Biometrics*, *39*(1), 217–225.

Mayer, A., & Dietzfelbinger, L. (2018). *EffectLiteR: An R package for estimating average and conditional effects*. Retrieved from https://github.com/amayer2010/EffectLiteR.

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behavioral Research*, *51*(2–3), 374–391. doi:10.1080/00273171.2016.1151334

Mayer, A., Umbach, N., Flunger, B., & Kelava, A. (2017). Effect analysis using nonlinear structural equation mixture modeling. *Structural Equation Modeling*, *24*(4), 556–570. doi:10.1080/10705511.2016.1273780

Miratrix, L. W., Sekhon, J. S., & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B*, *75*(2), 369–396.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.) [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Nieuwenhuis, R., te Grotenhuis, M., & Pelzer, B. J. (2017). Weighted effect coding for observational data with wec. *The R Journal*, *9*, 477–485.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511803161

Core Team, R. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, *69*(347), 682–689.

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, *34*, 216–221.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments: Essays on principles. Section 9. *Statistical Science*, *5*(4), 465–480.

Steyer, R., Gabler, S., von Davier, A., & Nachtigall, C. (2000). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online*, *5*, 55–86.

Steyer, R., Mayer, A., & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 606–631). Dordrecht, Netherlands: Springer. doi:10.1007/978-94-007-0753-5_295