Quantitative and Computational Methods in Behavioral Sciences

PsychOpen GOLD

Method Dissemination Articles

Informative Hypothesis Testing in the EffectLiteR Framework: A Tutorial

Caroline Keck [1], Axel Mayer [2], Yves Rosseel [1]

[1] *Department of Data Analysis, Ghent University, Ghent, Belgium.* **[2]** *Psychological Methods and Evaluation, Bielefeld University, Bielefeld, Germany.*

Quantitative and Computational Methods in Behavioral Sciences, 2024, Article e13059, https://doi.org/10.5964/qcmb.13059

Received: 2023-10-19 • **Accepted:** 2024-08-26 • **Published (VoR):** 2024-12-23

Handling Editor: Timo von Oertzen, Max Planck Institute for Human Development, Berlin, Germany

Corresponding Author: Yves Rosseel, Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University, Henri Dunantlaan 1, 9000 Ghent, Belgium. E-mail: Yves.Rosseel@UGent.be

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

In this paper, we illustrate how the typical workflow in analyzing psychological data, including analysis of variance and null hypothesis significance testing, may fail to bridge the gap between research questions and statistical procedures. It fails, because it does not provide us with the quantities of interest, which are often average and conditional effects, and it is insufficient, because it does not take the expectations of the researcher about these quantities into account. Using a running example, we demonstrate that the EffectLiteR framework as well as informative hypothesis testing are more suitable to narrow the gap between research questions and statistical procedures. Furthermore, we provide two empirical data examples, one in the context of linear regression and one in the context of the generalized linear model, to further illustrate the use of informative hypothesis testing in the EffectLiteR framework.

Keywords

analysis of variance, ANOVA, null hypothesis significance testing, informative hypothesis testing, constrained statistical inference, average effects, conditional effects

When considering which statistical procedure works best for examining a given hypothesis, researchers are often unaware that there is not always a one-to-one mapping or simple translation between a research question and a statistical procedure. This is because each procedure comes with different advantages and disadvantages and can

only answer questions within a specific, limited context. The unawareness regarding the limits of statistical procedures plays a major role in the replication crisis (see, e.g., Ioannidis, 2005; Yong, 2012) and contributes to what can be called the gap between research questions and statistical procedures (see, e.g., Scheel et al., 2021).

In this paper, we will use a simulated data example to illustrate how the typical workflow in analysing psychological data can fail to bridge the gap between research questions and statistical procedures. We will focus on research questions that involve the evaluation of the effect of a focal categorical predictor X on an outcome variable Y taking into account several covariates. A typical research question in this context may be "Is there a significant effect of X on Y on average?" or "Is the effect of X larger for males than for females?".

The usual approach to analyze this type of data is by using analysis of variance (ANOVA; Rutherford, 2001) in combination with null hypothesis significance testing (NHST). Technically, ANOVA is just a linear regression model with at least one categorical focus predictor. By using the term NHST, we refer to the following typical setting. Under the null hypothesis, we use equality constraints. That is, we test whether one or more quantities (such as model parameters or means) are equal to zero (or another constant) or whether several quantities are equal to each other. Under the alternative hypothesis, we assume that there are no equalities. An example could refer to regression coefficients, where $H_0: \beta_1 = \beta_2$ is tested against $H_1: \beta_1 \neq \beta_2$, or in other words $H_0: \beta_1 - \beta_2 = 0$ is tested against $H_1: \beta_1 - \beta_2 \neq 0$.

ANOVA is one of the most often used statistical procedures in psychology. Unfortunately, using ANOVA in combination with NHST is often insufficient for two main reasons. First, ANOVA does not provide us with the quantities of interest, which are usually average or conditional effects. And second, it does not take into account the expectations of the researcher about these effects of interest, for example their order.

We will introduce two methodological approaches that may be more suitable. These are the EffectLiteR framework (Mayer et al., 2016; Mayer & Dietzfelbinger, 2019) and informative hypothesis testing (IHT; Hoijtink, 2012; Silvapulle & Sen, 2005). We will show how both approaches can be used simultaneously to narrow the gap between research questions and statistical procedures. After these demonstrations, we will present two empirical data examples, one in the context of linear regression and one in the context of the generalized linear model. We provide R (R Core Team, 2020) code as well as further supplemental materials on the OSF project site (see Keck et al., 2024).

By reading this paper, we hope that applied researchers will gain awareness about the pitfalls of using ANOVA together with NHST. Furthermore, we hope that applied researchers will obtain some familiarity with our proposed method, IHT in the EffectLiteR framework, including all its advantages. In order to ease the transition for readers that are currently using traditional ANOVA in combination with NHST, this paper will only

consider the frequentist framework. We acknowledge that ultimately, a fully Bayesian approach may provide even more options and flexibility (in the future).

Simulated Data Example

The running example focuses on a clinical researcher who is interested in the effect of a new drug in combination with cognitive behavioral therapy (CBT) to reduce depression. The main hypothesis is that CBT in combination with either the old or the new drug is more effective than CBT only, and that the new drug with CBT is more effective than the old drug with CBT. The simulated data set can be found on the OSF project site under the name "runningExampleData.csv" (see Keck et al., 2024).

The researcher sets up a non-randomized experiment with three treatment groups; one group receiving CBT only ($X = 0$), one group receiving CBT together with the old drug ($X = 1$) and one group receiving CBT together with the new drug ($X = 2$). The total sample size is $n = 1000$. As covariates, the continuous variable depression pre-test (Z) and the dichotomous variable "treatment experience" (K) are considered. The latter indicates whether any treatment has been received before ($K = 1$) or not ($K = 0$). The outcome variable Y is the depression post-test. Note that both Z and Y will be treated as manifest variables and higher scores denote better mental health. Furthermore, note that our running example is a simplification of a non-randomized experiment. In a real world setting, more covariates would have to be taken into account.

The expectation of the researcher is that $\mu_2^{adj} > \mu_1^{adj} > \mu_0^{adj}$, where μ_2^{adj} and μ_1^{adj} correspond to the treatment experience and pre-test adjusted means of Y for the combination of CBT with the new and the old drug, respectively, and μ_0^{adj} corresponds to the treatment experience and pre-test adjusted mean of Y for CBT with no drugs. However, as will be explained in more detail later on, there are different ways to adjust means that researchers may not always be aware of.

After the data collection, the researcher will typically first have a look at some descriptive statistics. Table 1 shows the estimates of various (conditional) expectations and adjusted means in the simulated example. In line with our data generation, we find that there are no (significant) baseline differences in depression pre-test Z between the levels of X and K. This is also reflected by an F-test: $F(5, 994) = 0.2724$, $p = .928$, where we compared a model with main effects for X and K and their interaction effect (X:K) to the intercept-only model. Often, the researcher will also visualize the data, for example as can be seen in Figures 1 and 2. Figure 1 shows a boxplot of the post-test Y for the different treatment groups X. Figure 2 depicts the linear regression of post-test Y on pre-test Z in the different combinations of treatment group X and treatment experience K. We see that the slopes do not differ much between the grid elements. In other words, there seems to be no three-way interaction between either X or K and the two continuous variables Y and Z. Again, this can be confirmed using an F-test; $F(2, 988) = 1.637$, $p = .199$.

Table 1

Estimates of Various (Conditional) Expectations and Adjusted Means in the Simulated Example

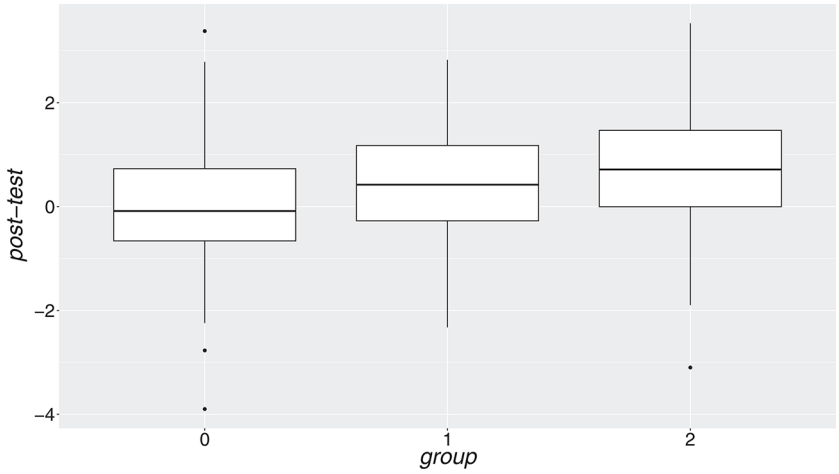
Group	$X = 0$	$X = 1$	$X = 2$	$X = .$
$K = 0$	$\hat{E}(Y \mid X = 0, K = 0) = 0.008$	$\hat{E}(Y \mid X = 1, K = 0) = 0.386$	$\hat{E}(Y \mid X = 2, K = 0) = 0.410$	$\hat{E}(Y \mid K = 0) = 0.257$
	$\hat{E}(Z \mid X = 0, K = 0) = 0.006$	$\hat{E}(Z \mid X = 1, K = 0) = -0.017$	$\hat{E}(Z \mid X = 2, K = 0) = -0.036$	$\hat{E}(Z \mid K = 0) = -0.016$
	$P(X = 0, K = 0) = 0.200$	$P(X = 1, K = 0) = 0.150$	$P(X = 2, K = 0) = 0.200$	$P(K = 0) = 0.550$
$K = 1$	$\hat{E}(Y \mid X = 0, K = 1) = 0.091$	$\hat{E}(Y \mid X = 1, K = 1) = 0.596$	$\hat{E}(Y \mid X = 2, K = 1) = 0.913$	$\hat{E}(Y \mid K = 1) = 0.695$
	$\hat{E}(Z \mid X = 0, K = 1) = -0.005$	$\hat{E}(Z \mid X = 1, K = 1) = 0.123$	$\hat{E}(Z \mid X = 2, K = 1) = -0.039$	$\hat{E}(Z \mid K = 1) = -0.013$
	$P(X = 0, K = 1) = 0.100$	$P(X = 1, K = 1) = 0.050$	$P(X = 2, K = 1) = 0.300$	$P(K = 1) = 0.450$
$K = .$	$\hat{E}(Y \mid X = 0) = 0.036$	$\hat{E}(Y \mid X = 1) = 0.438$	$\hat{E}(Y \mid X = 2) = 0.712$	
	$\widehat{AdjM}_0 = 0.044$	$\widehat{AdjM}_1 = 0.475$	$\widehat{AdjM}_2 = 0.643$	
	$\hat{E}(Z \mid X = 0) = 0.002$	$\hat{E}(Z \mid X = 1) = 0.018$	$\hat{E}(Z \mid X = 2) = -0.038$	
	$P(X = 0) = 0.300$	$P(X = 1) = 0.200$	$P(X = 2) = 0.500$	

Note. The last column of the table shows the expectations and probabilities where we only condition on K. The last row shows the expectations and probabilities where we only condition on X. The cell probabilities are fixed by design. The adjusted means ($AdjM$) are computed according to EffectLiteR, which will be explained in more detail later on.

Quantitative and Computational Methods in Behavioral Sciences
2024, Article e13059. https://doi.org/10.5964/qcmb.13059

Figure 1

Boxplot of Depression Post-Test Y Grouped by the Three Levels of X



ANOVA Versus EffectLiteR

For the data generation as well as the data analysis, the following model is used[1] :

$$\begin{aligned}\text{post-test}_i = {} & \beta_0 + \beta_1 \cdot \text{group1}_i + \beta_2 \cdot \text{group2}_i + \beta_3 \cdot \text{treatexp1}_i + \beta_4 \cdot \text{pre-test}_i + \\ & \beta_5 \cdot \text{group1}_i \cdot \text{treatexp1}_i + \beta_6 \cdot \text{group2}_i \cdot \text{treatexp1}_i + \\ & \beta_7 \cdot \text{group1}_i \cdot \text{pre-test}_i + \beta_8 \cdot \text{group2}_i \cdot \text{pre-test}_i + \\ & \beta_9 \cdot \text{treatexp1}_i \cdot \text{pre-test}_i + \varepsilon_i \,. \end{aligned} \tag{1}$$

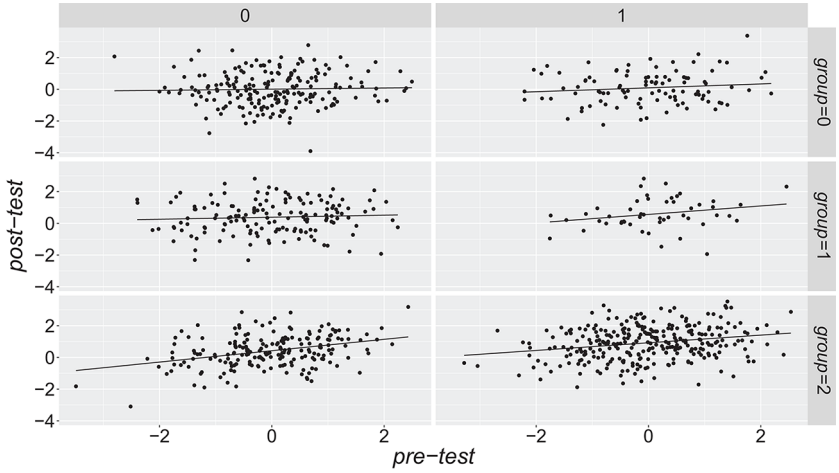
Note that this model does not include a three-way interaction term. The control group is $X = 0$. The variables group1_i and group2_i are dummy variables which indicate by a value of 1, if a subject belongs to group $X = 1$ or $X = 2$, respectively, and are 0 otherwise. Similarly, the variable treatexp1_i is a dummy variable, which indicates whether a subject has treatment experience ($K = 1$) or not ($K = 0$).

Then, the researcher will formulate the hypotheses of interest. As in our example, the focus is usually on the "main effect" of the treatment and following the classical NHST approach, $H_0{:}\mu_2^{adj} = \mu_1^{adj} = \mu_0^{adj}$ will be tested against $H_1{:}\mu_2^{adj} \neq \mu_1^{adj} \neq \mu_0^{adj}$. However, as mentioned before, the μ's may correspond to different types of adjusted means, which

1) Note that EffectLiteR uses a different notation with so-called intercept and effect functions based on gammas instead of betas. The interested reader is referred to Keck et al. (2021).

Figure 2

Slopes of the Linear Regression of the Post-Test Y on the Pre-Test Z for the Different Combinations of Group X and Treatment Experience K



will be explained in the next section. Furthermore, the alternative hypothesis does not correspond to the initial expectation of the researcher, where the adjusted means are ordered: $\mu_2^{adj} > \mu_1^{adj} > \mu_0^{adj}$. We will also argue that the researcher is actually interested in the so-called average effect of the treatment, as will be explained in more detail later on. After specifying the hypotheses, the researcher will fit the model. In the sequel, we will first give a theoretical overview of ANOVA, before coming back to our running example and the results of the fitted model.

ANOVA

ANOVA is one of the most popular statistical techniques in the social and behavioral sciences. It is a framework or collection of methods based on a linear regression model, where at least one predictor is categorical in nature. Usually, this categorical predictor describes the different conditions of an experiment, or the different (treatment) groups in an intervention study. The ANOVA framework (Edwards, 1993) includes one-way and multi-way ANOVA, univariate and multivariate (M)ANOVA, ANOVA using within-subjects and/or between-subjects factors, and AN(C)OVA where covariates are included in the model.

The problem with ANOVA is that we only obtain regression coefficients as well as main and interaction effects, which are often difficult to interpret. Especially the interpretation of a main effect in the presence of an interaction effect is far from trivial.

This is because different sum of squares (SS) can be used in ANOVA for hypothesis testing. Depending on the SS, the main effect is defined in a different way and thus, a different null hypothesis is tested (for an overview, see, e.g., Fox, 2016; Graefe et al., 2022; Maxwell et al., 2018). There are several types of SS, namely Type I, Type II and Type III.[2] To understand the main differences between these different types, consider a model of the form $Y \sim A + B + A{:}B$. An ANOVA table based on Type I SS will contain the following three model comparisons: $Y \sim A$ versus $Y \sim 1$,[3] $Y \sim A + B$ versus $Y \sim A$, and $Y \sim A + B + A{:}B$ versus $Y \sim A + B$. In other words, Type I SS corresponds to an incremental procedure, where single terms are added to the model one by one and the model is then compared to the previous model without the new term. For Type II, the ANOVA table will contain the following model comparisons: $Y \sim A + B$ versus $Y \sim B$, $Y \sim A + B$ versus $Y \sim A$, and $Y \sim A + B + A{:}B$ versus $Y \sim A + B$. The main characteristic of the Type II procedure is the principle of marginality: If a term is removed from the model, then all higher-order terms involving this term will be removed too. Finally, the Type III procedure leads to the following set of model comparisons: $Y \sim A + B + A{:}B$ versus $Y \sim B + A{:}B$, $Y \sim A + B + A{:}B$ versus $Y \sim A + A{:}B$, and $Y \sim A + B + A{:}B$ versus $Y \sim A + B$. It is always the full model versus a model where a single term is deleted.

Type III SS are used per default in many popular software programs like SPSS (IBM Corp, 2020). Thus, researchers will typically use Type III SS without further deliberation, even though there is a great controversy in the literature about when to use which SS (e.g., Hector et al., 2010; Herr & Gaebelein, 1978; Macnaughton, 1998). Graefe et al. (2022) conducted simulation studies considering the different types of SS in balanced, proportional and non-orthogonal designs. They found that in balanced designs, using either one of the three SS yields main effects that can be interpreted unambiguously. However, in proportional designs, this is only true when using Type I and II SS. In case of Type III SS, the main effect is biased if there are interactions. Finally, in non-orthogonal designs, the main effect is always biased when using Type I SS. And when there are interactions, Types II and III also yield biased main effects. Nevertheless, for the sake of illustration, we will use ANOVA to analyse our simulated dataset.

Using the centered version of the pre-test variable Z, the model for our running example is fitted as follows in R:

2) There are also Type IV SS, but they are rarely used.

3) The '1' at the right-hand side represents the intercept.

Listing 1

Specification of ANOVA Models Including All Two-Way Interactions

```
# this R code can be found under 01Anova.R on the OSF project site

# read in data
Data <- read.csv("runningExampleData.csv")

# center pre-test for better interpretation of regression coefficients
Data$pretest.cent <- Data$pretest - mean(Data$pretest)
# treat group and treatexp as categorical variables
Data$group <- factor(Data$group)
Data$treatexp <- factor(Data$treatexp)

# load packages
library(car) # for Anova () function

# fit regression model with all 2-way interaction effects
lmod.treat <- lm(posttest ~ (group + treatexp + pretest.cent)^2, data = Data)
summary(lmod.treat)

# Type I ANOVA table
anova(lmod.treat)

# Type II ANOVA table
Anova(lmod.treat, type = 2)

# Type III ANOVA table
# attention: we must use an orthogonal or a sum-to-zero coding scheme
options(contrasts = c("contr.sum", "contr.poly"))
lmod.sum <- lm(posttest ~ (group + treatexp + pretest.cent)^2, data = Data)
summary(lmod.sum)
Anova(lmod.sum, type = 3)
```

The `anova( )` function is a function in base R, whereas the `Anova( )` function belongs to the car package (Fox et al., 2022). The former uses Type I SS, whereas the latter can handle Type II and III SS. Furthermore, R uses treatment coding per default, which has to be changed to a zero-to-sum coding scheme (for example sum coding) when using Type III SS. This prevents main and interaction effects from overlapping. For more information about coding schemes, see, for example, Cohen et al. (2003) and Hardy (2003). Note, however, that in some statistical packages (for example SAS), coding schemes are automatically taken care of. Table 2 shows the ANOVA results when using the three different SS. In Table 3, the results of the linear model using treatment coding can be seen. Appendix Table A1 shows the results of the linear model when using sum coding.

Table 2

ANOVA Results for Type I, Type II and Type III Sum of Squares (SS)

	Type I SS				Type II SS				Type III SS			
Source	***Df***	**SS**	***F*-value**	***p*-value**	***Df***	**SS**	***F*-value**	***p*-value**	***Df***	**SS**	***F*-value**	***p*-value**
Intercept									1	115.03	113.457	<0.001***
Group	2	85.85	42.340	<0.001***	2	64.14	31.632	<0.001***	2	66.02	32.558	<0.001***
Treatexp	1	24.29	23.956	<0.001***	2	24.02	23.694	<0.001***	1	12.26	12.097	<0.001***
Pretest	1	33.38	32.922	<0.001***	2	33.59	33.128	<0.001***	1	18.33	18.081	<0.001***
Group:treatexp	2	8.41	4.148	0.016*	2	8.31	4.100	0.017*	2	8.31	4.100	0.017*
Group:pretest	2	9.76	4.813	0.008**	2	9.05	4.461	0.012*	2	9.05	4.461	0.012*
Treatexp:pretest	1	0.02	0.016	0.899	1	0.02	0.016	0.899	1	0.02	0.016	0.899
Residuals	990	1003.72			990	1003.72			990	1003.72		

Note. In our computations, we used the same factor order as is shown in this table.

Table 3

Linear Regression Model Results Using Treatment Coding

Source		Coefficient	Estimate (*SE*)	*t*-value	*p*-value
Intercept		$\hat{\beta}_0$	0.007 (0.071)	0.091	0.928
	$I_{X=1}$	$\hat{\beta}_1$	0.380 (0.109)	3.490	< 0.001***
Group					
	$I_{X=2}$	$\hat{\beta}_2$	0.410 (0.101)	4.070	< 0.001***
Treatexp		$\hat{\beta}_3$	0.084 (0.123)	0.680	0.497
Pre-test		$\hat{\beta}_4$	0.073 (0.067)	1.096	0.273
	$I_{X=1}$	$\hat{\beta}_5$	0.113 (0.206)	0.548	0.584
Group:treatexp					
	$I_{X=2}$	$\hat{\beta}_6$	0.420 (0.154)	2.729	0.006**
	$I_{X=1}$	$\hat{\beta}_7$	0.031 (0.096)	0.327	0.743
Group:pre-test					
	$I_{X=2}$	$\hat{\beta}_8$	0.217 (0.078)	2.771	0.006**
Treatexp:pre-test		$\hat{\beta}_9$	-0.009 (0.070)	−0.127	0.899

Table 2 shows that the results differ depending on which SS was used. Notably, the results concerning the interaction terms are the same when using Type II and III SS. Generally, the results of the highest order terms (in our case the two-way interactions) are identical between Type II and III SS, but the results of the lower order terms differ. Furthermore, the result of the highest order term when using Type I SS (in our case the treatexp:pretest interaction) corresponds to the result of this term when using Type II and III SS. This term is shown in the last line before the residuals in the results tables and is the only term that has identical results between all three types of SS. Lastly, note that when using Type III SS, the p-values of the terms with $df = 1$ correspond to the p-values of these terms in the linear model when using sum coding (see Table 2 and Appendix Table A1). This is because $t^2 = F$ if $df = 1$.

Using treatment contrasts and assuming Z is mean centered, we can interpret the regression coefficients in Table 3. Recall that $X = 0$ is the control group. The intercept β_0 corresponds to the mean of Y in the control group given $K = 0$ and $Z = 0$ (see Appendix B for more details), while β_1 denotes the difference between the means of Y in the groups $X = 1$ and the control group, given $K = 0$ and $Z = 0$:

$$\beta_1 = E(Y \mid X = 1, K = 0, Z = 0) - E(Y \mid X = 0, K = 0, Z = 0)\,. \tag{2}$$

Similarly, β_2 denotes the difference between the means of Y in the groups $X = 2$ and the control group, given $K = 0$ and $Z = 0$:

$$\beta_2 = E(Y \mid X = 2, K = 0, Z = 0) - E(Y \mid X = 0, K = 0, Z = 0)\,. \tag{3}$$

The difference between the means of Y of $K = 0$ and $K = 1$ (given the control group $X = 0$ and $Z = 0$) corresponds to β_3:

$$\beta_3 = E(Y \mid X = 0, K = 1, Z = 0) - E(Y \mid X = 0, K = 0, Z = 0)\,. \tag{4}$$

And β_4 denotes the expected change in Y for a unit change in Z in the control group if $K = 0$:

$$\beta_4 = E(Y \mid X = 0, K = 0, Z = z + 1) - E(Y \mid X = 0, K = 0, Z = z)\,. \tag{5}$$

The change in the effect of $X = 1$ versus $X = 0$ (β_1) between $K = 1$ and $K = 0$ while keeping Z constant is denoted by β_5. Similarly, β_6 describes the change in the effect of $X = 2$ versus $X = 0$ (β_2) between $K = 1$ and $K = 0$ while keeping Z constant. The change in the effect of $X = 1$ versus $X = 0$ (β_1) for a unit change in Z when $K = 0$ is represented by β_7. Similarly, β_8 denotes the change in the effect of $X = 2$ versus $X = 0$ (β_2) for a unit change in Z when $K = 0$. Finally, β_9 describes the change in the effect of $K = 1$ versus $K = 0$ for a unit change in Z when $X = 0$.

Remember that in our example, the researcher assumed an ordering of adjusted means: $\mu_2^{adj} > \mu_1^{adj} > \mu_0^{adj}$. After obtaining a significant main effect of the group variable, the researcher will usually use contrasts to compare the means in depth. Note that in the NHST setting, if no hypothesis about the adjusted means has been specified right from the start, this is called post-hoc testing and should only be used in a descriptive or an exploratory manner. Furthermore, we have to control for familywise error rates (Keselman et al., 2011), which can be done by the emmeans package (Lenth et al., 2022). Concerning our simulated data example, we specify the contrasts in R as follows:

Listing 2

Specification of Contrasts

```
# this R code can be found under 02emmeans.R on the OSF project site
library(emmeans)

# post-hoc tests / contrasts
emmeans(lmod.treat, "group")
emmeans(lmod.treat, "group", weights = "proportional")

emmeans(lmod.treat, "group", contr = "trt.vs.ctrl")
emmeans(lmod.treat, "group", contr = "trt.vs.ctrl", weights = "proportional")

emmeans(lmod.treat, "group", contr = "eff")
emmeans(lmod.treat, "group", contr = "eff", weights = "proportional")
```

The contrasts are based on the marginal means of Y, which are averaged over the levels of K at the mean of Z. Table 4 shows the marginal means of Y when using equal and when using proportional weights. Using proportional weights implies that the marginal means of Y are averaged over the marginal distribution of K at the mean of Z. This leads to slightly different results in our example compared to using equal weights.[4]

Table 4

Marginal Means (MMs) of the Depression Post-Test Y With Standard Errors in Parentheses and 95 % Confidence Intervals

	Equal Weights			Proportional Weights		
Group	***MM* (*SE*)**	**Lower CL**	**Upper CL**	***MM* (*SE*)**	**Lower CL**	**Upper CL**
$X = 0$	0.048 (0.062)	−0.073	0.169	0.044 (0.060)	−0.073	0.162
$X = 1$	0.485 (0.083)	0.323	0.646	0.475 (0.079)	0.320	0.629
$X = 2$	0.668 (0.046)	0.578	0.758	0.643 (0.047)	0.551	0.735

The first set of contrasts (see Table 5) compares the treatment groups with the control group, where the estimates of the differences in marginal means of Y in the levels of X are computed. The second set of contrasts (see Table 6) gives us effect contrasts, where the marginal means of Y are compared with the equally weighed cell means of Y for $Z = 0$. Both contrasts have also been computed using proportional weights.

4) Note that the differences would be much larger if the marginal distribution of K was more unbalanced and if there were strong interactions between X and K.

Table 5

Treatment Versus Control Contrasts Between Marginal Means of Y

		Equal Weights			Proportional Weights		
Contrast	**Estimate (*SE*)**	***Df***	***t*-ratio**	***p*-value**	**Estimate (*SE*)**	***t*-ratio**	***p*-value**
$MM_{X=1} - MM_{X=0}$	0.436 (0.103)	990	4.235	$< .001^{***}$	0.430 (0.099)	4.355	$< .001^{***}$
$MM_{X=2} - MM_{X=0}$	0.620 (0.077)	990	8.059	$< .001^{***}$	0.599 (0.076)	7.859	$< .001^{***}$

Table 6

Effect Contrasts Between Marginal Means of Y

		Equal Weights			Proportional Weights		
Contrast	**Estimate (*SE*)**	***Df***	***t*-ratio**	***p*-value**	**Estimate (*SE*)**	***t*-ratio**	***p*-value**
$MM_{X=0} - MM_{X=.}$	−0.352 (0.052)	990	−6.798	$< .001^{***}$	−0.343 (0.050)	−6.824	$< .001^{***}$
$MM_{X=1} - MM_{X=.}$	0.084 (0.061)	990	1.387	.166	0.087 (0.058)	1.500	.134
$MM_{X=2} - MM_{X=.}$	0.268 (0.046)	990	5.822	$< .001^{***}$	0.256 (0.046)	5.620	$< .001^{***}$

Considering the results, it would seem that the expectations of the researcher are satisfied in the data. That is, the marginal mean of group $X = 2$ is higher than the marginal mean of group $X = 1$ and the marginal mean of group $X = 1$ is higher than the marginal mean of group $X = 0$ (see Table 4). Therefore, it seems that there is a greater improvement in the depression post-test Y for CBT with the new drug compared to CBT with the old drug and that there is a greater improvement for CBT with the old drug compared to CBT only. The contrasts agree with this observation. Here, the treatment contrast between group $X = 2$ and group $X = 0$ is larger than the treatment contrast between group $X = 1$ and $X = 0$ (see Table 5). Furthermore, the effect contrast is largest for group $X = 0$ and smallest for group $X = 1$ (see Table 6).

EffectLiteR

Given the different choices for SS when using ANOVA, it becomes clear that main effects are not defined precisely and unambiguously when using ANOVA. However, they are defined precisely and unambiguously in the causal inference literature (see, e.g., Angrist et al., 1974; Neyman, 1990; Pearl, 2009; Rubin, 1974, 2005; Steyer et al., 2000). Common types of effects are the so-called average effects (for example, the effect of a treatment averaged over treatment experience), conditional effects (for example, the effect of a treatment for those without treatment experience only), effects on the treated, effects on the untreated, and so forth. Unfortunately, although the mathematical definition of these effects is well understood, they can be quite complicated and tedious to compute.

EffectLiteR is a framework and an R package that is built upon the clear definitions of effects in the causal inference literature (Imbens & Rubin, 2015; Steyer et al., 2014). Researchers can use EffectLiteR to estimate various effects of interest as well as their standard errors, when the treatment variable is categorical and the outcome variable is continuous. In addition, Wald or *F*-tests are used to test different hypotheses, for example that (all) average effects are equal to zero, or that (all) conditional effects are equal to zero in the population. For an in-depth introduction into EffectLiteR, see Mayer et al. (2016). In contrast to ANOVA, EffectLiteR directly provides the kind of effects that most applied researchers are interested in.

An average effect is defined as the unconditional expectation of the difference between expected outcomes under treatment and under control. It corresponds to the average causal effect if there are no further unobserved confounding variables and the regression of the post-test Y on the pre-test Z, given a combination of treatment experience K and treatment group X, is in fact linear. Furthermore, an average effect consists of the difference of adjusted means. Considering our running example and groups $X = 1$ and $X = 0$, this is (Mayer et al., 2016):

$$AE_{10} = E[E(Y \mid X = 1, K, Z) - E(Y \mid X = 0, K, Z)] \quad (06)$$

$$= E[E(Y \mid X = 1, K, Z)] - E[E(Y \mid X = 0, K, Z)] \quad (07)$$

$$= AdjM_1 - AdjM_0 . \quad (08)$$

After obtaining the descriptive statistics, the researcher can first fit the EffectLiteR model to compute the adjusted means and the effects of interest, but disregard the test statistics and p-values. For our running example, this can be done as follows in R:

Listing 3

Specification of EffectLiteR Model

```
# this R code can be found under 03EffectLite.R on the OSF project site

library(EffectLiteR)

elrmod <- effectLite(
    y = "posttest",
    x = "group",
    k = "treatexp",
    z = "pretest.cent",
    interactions = "2-way",
    method = "lm",
    data = Data
)

print(elrmod)
```

The full output is shown in Appendix C. The average effects and adjusted means can be found on lines 91–103 and are depicted in Table 7. Note that the average effects correspond to the differences in adjusted means and the subscripts refer to the groups which are compared:

$$AE_{10} = AdjM_1 - AdjM_0, \qquad (09)$$

$$AE_{20} = AdjM_2 - AdjM_0\,. \qquad (10)$$

Table 7

Average Effects and Adjusted Means, as Estimated by EffectLiteR

Group	Adjusted Mean (*SE*)	Average Effect (*SE*)
$X = 0$	0.044 (0.060)	
$X = 1$	0.475 (0.079)	0.430 (0.099)
$X = 2$	0.643 (0.047)	0.599 (0.076)

Furthermore, the adjusted means computed by EffectLiteR are very close, but not identical, to the marginal means with proportional weights computed by emmeans. This is because slightly different computations are used.[5]

While ANOVA tests whether main effects are significantly different from zero, EffectLiteR tests whether average effects are significantly different from zero. The number of average effects depends on the number of levels of the group variable. In our example, there is one control and two treatment groups, resulting in two average effects. In short, the EffectLiteR approach is more suitable than the ANOVA approach for testing research questions that focus on average effects. There are just a few rare cases, such as balanced designs, where the results are identical. For further applications of the EffectLiteR approach, see, for example, Flunger et al. (2019), Mayer et al. (2020), Mueller et al. (2015), Rek et al. (2022), and Sadovich (2020).

Informative Hypothesis Testing (in the EffectLiteR Framework)

As mentioned before, the researcher may have specific expectations about the data. In our example, the researcher expects that the average effect of CBT with the new drug is larger than the average effect of CBT with the old drug, which in turn is larger than zero:

5) The adjusted means are computed by averaging over *K* and *Z* simultaneously, which takes a possible dependency between them into account. The marginal means with proportional weights are computed by first averaging over *Z* and then averaging over *K*. This works as follows: In the first step, the mean of *Z* is plugged in to obtain a grid of estimated marginal means (a so-called EMM-grid). In the second step, their weighted row- or column-sums are computed.

$$AE_{20} > AE_{10} > 0. \tag{11}$$

After plugging in the definitions of AE_{20} and AE_{10} (see Equations 9 and 10), this corresponds to:

$$AdjM_2 - AdjM_0 > AdjM_1 - AdjM_0 > 0. \tag{12}$$

By adding $AdjM_0$ to each part of the equation, this can be simplified to:

$$AdjM_2 > AdjM_1 > AdjM_0\,. \tag{13}$$

In other words, the researcher expects a complete ordering of the adjusted means of the treatment groups: It is assumed that the adjusted mean of group $X = 2$ is larger than the adjusted mean of group $X = 1$ and that the adjusted mean of group $X = 1$ is larger than the adjusted mean of group $X = 0$.

Hypotheses that reflect the expectations of the researcher are known as informative hypotheses and both Bayesian and frequentist procedures for IHT are available (Barlow et al., 1972; Hoijtink, 2012; Robertson et al., 1988; Silvapulle & Sen, 2005; Vanbrabant, 2020). In this paper, we will focus on the frequentist approach. IHT is not widely adopted by applied researchers in the social sciences yet. This is unfortunate for multiple reasons. First, compared to classical NHST, IHT allows to formulate hypotheses in a way that can be closer to typical research questions. These typical research questions often contain specific directions or orders regarding regression coefficients, group means or effects of interest. Using NHST, it is not possible to directly test hypotheses about these orders or directions when there is more than one constraint, whereas IHT allows for it. Second, compared to NHST, IHT can lead to a substantial gain of power (up to 50%; see Vanbrabant et al., 2015). This is because the parameter space is restricted according to the directions and orders that are defined in the hypothesis.

If the researcher would like to implement IHT about adjusted means, a precise way to formulate the hypotheses of interest would be:

$$H_0\colon AdjM_2 = AdjM_1,\ AdjM_1 = AdjM_0, \tag{14}$$

and

$$H_1\colon AdjM_2 > AdjM_1,\ AdjM_1 > AdjM_0, \tag{15}$$

where at least one of the inequality constraints must be strictly true, whereas the other one may be an equality. Informative hypotheses can be tested via an informative Wald test, where constrained parameter estimates are used that have been obtained by means of quadratic programming. Appendix Section D explains the difference between a regular and an informative Wald test and Appendix Section E describes how to compute the

p-values for an informative Wald test. Informative test statistics (in the EffectLiteR framework) as well as the approaches for calculating their p-values are described with more technical details in Keck et al. (2021). Furthermore, simulation studies assessing the relevant practical properties can be found in Keck et al. (2022, 2023).

Regarding our running example, we use the following R syntax to compute the informative Wald test:

Listing 4

Informative Hypothesis Testing in EffectLiteR

```
# this R Code can be found under 04IHT.R on the OSF project site

elrmod <- effectLite(
    y = "posttest",
    x = "group",
    k = "treatexp",
    z = "pretest.cent",
    interactions = "2-way",
    method = "sem", # must be "sem" for effectLite_iht()
    fixed.cell = TRUE , fixed.z = TRUE,
    homoscedasticity = TRUE,
    data = Data
)

effectLite_iht(
    object = elrmod,
    constraints = "adjmean2 > adjmean1; adjmean1 > adjmean0"
)
# $test.stat
# [1] "Fbar"
# $Wald.info
# [1] 62.87448
# $pvalue
# [1] 1.054712e-14
```

For technical reasons, we must use the argument `method = "sem"` in order to use the `effectLite_iht()` function. In addition, we have specified the arguments `fixed.cell = TRUE, fixed.z = TRUE`, and `homoscedasticity = TRUE`, in order to obtain similar results as when using `method = "lm"`. The `effectLite_iht()` function contains a `constraints =` argument that can be used to specify the informative hypothesis. Here, the keyword `adjmean0` refers to the adjusted mean of the $X = 0$ group, while `adjmean1` and `adjmean2` refer to the adjusted means of the $X = 1$ and $X = 2$ group, respectively. The `constraints =` argument corresponds to the alternative hypothesis as presented in Equation (15). The function returns the value of the informative Wald test, and a p-value. For our simulated dataset,

we obtain $Wald_{info} = 62.87$, $p < .001$, which allows us to discard H_0 in favor of the ordered hypothesis H_1. This coincides indeed with the expectations of the researcher.

Instead of formulating the informative hypothesis in terms of the adjusted means, we could as well formulate the hypothesis in terms of the average effects, as in Equation (13):

Listing 5

Informative Hypothesis Testing Using Average Effects

```
# this R Code can be found under 04IHT.R on the OSF project site

effectLite_iht(object = elrmod, constraints = "Eg2 > Eg1; Eg1 > 0")
# $test.stat
# [1] "Fbar"
# $Wald.info
# [1] 62.87448
# $pvalue
# [1] 1.054712e-14
```

Here, the keyword `Eg1` represents the average effect of $X = 1$ compared to $X = 0$, denoted by AE_{10} in Equation (9). Similarly, the keyword `Eg2` represents the average effect of $X = 2$ compared to $X = 0$, denoted by AE_{20} in Equation (10). The result is identical: $Wald_{info} = 62.87$, $p < .001$.

Type A and Type B Hypotheses

In IHT, a distinction is often made between two types of hypotheses, which are called Type A and Type B hypotheses. The null and alternative hypotheses in our example are Type A hypotheses, which are usually of main interest. When testing Type A hypotheses, H_{0A} (see Equation 14) states that all restrictions are equality restrictions, whereas the alternative hypothesis H_{1A} (see Equation 15) states that at least one inequality restriction is strictly true. Here, the researcher would typically like to obtain a significant result, as this indicates that at least some of the constraints are not equality constraints and thus must be inequality constraints. In contrast, the Type B null hypothesis H_{0B} states that all inequality restrictions hold, whereas the Type B alternative hypothesis H_{1B} states that at least one inequality restriction is violated. When testing Type B hypotheses, the researcher would typically like to obtain a non significant result, because that would imply that we cannot reject the null hypothesis (and thus the expectations of the researcher) based on the data.

If the researcher observes that the expected constraints are satisfied in the data, testing Type A hypotheses suffices. However, if one or more of the assumed constraints are violated in the data, but only to a very small extent, which might be due to sampling

variability, the researcher should conduct a Type B hypothesis test before conducting a Type A hypothesis test and correct for multiple testing. We recommend to pre-register this approach.

In our example, Type B hypotheses should be tested if at least one of the two constraints, either $AdjM_2 > AdjM_1$ or $AdjM_1 > AdjM_0$, is violated to a small extent in the data. For example, instead of obtaining the estimates that satisfy the constraints ($\widehat{AdjM}_0 = 0.044$, $\widehat{AdjM}_1 = 0.475$ and $\widehat{AdjM}_2 = 0.643$, see Table 7), suppose we would obtain $\widehat{AdjM}_2 = 0.470$, meaning that the constraint $AdjM_2 > AdjM_1$ would be violated to a very small extent. In that case, the researcher should start with testing the Type B hypotheses:

$$H_{0B}: AdjM_2 > AdjM_1,\ AdjM_1 > AdjM_0, \tag{16}$$

and

$$H_{1B}: AdjM_2 \not> AdjM_1,\ AdjM_1 \not> AdjM_0\,. \tag{17}$$

In case the detected violation of a constraint is small, hypothesis test Type B might still be non significant, in which case the researcher can proceed to test hypothesis test Type A. If hypothesis test Type B is significant, then it is clear that the data is contradicting the hypothesis. Therefore, there is no need for testing hypothesis test Type A.

The following syntax illustrates how we can test this Type B hypothesis for our running example:

Listing 6

Type B Informative Hypothesis Test

```
# this R Code can be found under 04IHT.R on the OSF project site

effectLite_iht(
    object = elrmod,
    constraints = "adjmean2 < adjmean1; adjmean1 < adjmean0"
)
# $test.stat
# [1] "Fbar"
# $Wald.info
# [1] 4.309257e-12
# $pvalue
# [1] 0.6304712
```

We formulated the Type B hypothesis in terms of the adjusted means, but we could as well have used the averaged effects. Unsurprisingly, when testing H_{0B} against H_{1B}, we obtain $Wald_{info} < 0.001$, $p = .630$. We are unable to reject the null hypothesis, and this

allows us to proceed and test the Type A hypothesis. Lastly, note that testing a Type B hypothesis before testing a Type A hypothesis rarely seems to alter the conclusions, as can be seen from the simulations conducted by Kuiper et al. (2015).

Further Types of Informative Hypotheses

It is also possible to formulate informative hypotheses using other types of constraints (see e.g. Hoijtink, 2012). For example, effect sizes can be incorporated as in:

$$H_1{:}\beta_2 - \beta_1 > d \cdot \sigma, \tag{18}$$

where d is an effect size according to Cohen (1988) and σ is the sample standard deviation of Y. Note that in this subsection, β_2 and β_1 are generic expressions for any type of effect and could either correspond to regression coefficients, or in the EffectLiteR setting, to average or conditional effects. The latter are defined as (Mayer et al., 2016):

$$CE_{10|K=0,Z=0} = E(Y|X=1,K=0,Z=0) - E(Y|X=0,K=0,Z=0), \tag{19}$$

$$CE_{20|K=0,Z=0} = E(Y|X=2,K=0,Z=0) - E(Y|X=0,K=0,Z=0). \tag{20}$$

Then, Equation 18 corresponds to:

$$H_1{:}CE_{20|K=0,Z=0} - CE_{10|K=0,Z=0} > d \cdot \sigma. \tag{21}$$

From a substantive point of view, this means that the researcher assumes that the difference between the conditional effect of receiving CBT together with the new drug ($X=2$), given $K=0, Z=0$, and the conditional effect of receiving CBT together with the old drug ($X=1$), given $K=0, Z=0$, is greater than d standard deviations. This may give some indication about the relevance of the difference between the two effects.

"About equality" constraints can be used to test informative hypotheses such as:

$$H_1{:}\,|\beta_2 - \beta_1| > d \cdot \sigma, \tag{22}$$

which corresponds to:

$$H_1{:}\beta_2 - \beta_1 > d \cdot \sigma,\ \beta_2 - \beta_1 < -d \cdot \sigma. \tag{23}$$

Finally, range constraints are a generalization of "about equality" constraints. They can be used to test informative hypotheses like:

$$H_1{:}\beta_2 - \beta_1 > \eta_1,\ \beta_2 - \beta_1 < \eta_2, \tag{24}$$

where the difference between β_2 and β_1 is supposed to lie in an interval with lower bound η_1 and upper bound η_2. Of course, combinations of all types of informative hypotheses are also possible.

Comparison With Equivalence Testing

Equivalence testing (Schuirmann, 1987; Seaman & Serlin, 1998; Wellek, 2010) is a special case of IHT. This is because hypotheses in equivalence testing are formulated using effect sizes, which is also an option in IHT. More specifically, hypotheses in equivalence testing are based on "smallest effect sizes of interest" (SESOIs), which are used to define a range of effect sizes that are of practical interest to the researcher. Equivalence testing became popular in reaction to the replication crisis and is often used in replication studies (see, e.g., Anderson & Maxwell, 2016; Lakens, 2017; Simonsohn, 2015). Here, researchers aim to show that an observed effect is small enough to conclude that its replication was unsuccessful. Generally, equivalence testing can be conducted within the framework of IHT, but IHT allows for a broader range of hypotheses that can be tested.

Regarding our running example, one may apply equivalence testing as follows. Let us assume that our running example is a replication study. We are interested to show that the difference between the raw means of Y in the groups $X = 1$ and $X = 2$ is small enough to conclude that it is not of practical relevance. In other words, we want to test whether the raw means of Y in the groups $X = 1$ and $X = 2$ are equivalent (hence the term equivalence testing).[6] Furthermore, let us assume that the original study had the same sample size as our running example.

For determining a SESOI, we can use one of multiple approaches (see e.g., Lakens et al., 2018). For a discussion on when to use which approach, see Baguley (2009). Here, we will use the popular small telescopes approach (Simonsohn, 2015). It defines the SESOI as the effect size that would give a certain power (say 33%) to the original study. Thus, it indicates the extent to which the replication results are consistent with an effect size large enough to have been detected in the original study (Simonsohn, 2015). We use the value of 33% power, which is typical for the approach. In our running example, this leads to equivalence bounds of -0.115 and 0.115.

We can then use the TOST (two one-sided tests) procedure (Goertzen & Cribbie, 2010; Lakens et al., 2018; Meyners, 2012; Quertemont, 2011; Rogers et al., 1993). It is implemented in the TOSTER package (Lakens & Caldwell, 2022). The procedure tests the effect estimate, in our case the difference between the raw means, against values at least as extreme as the lower and the upper equivalence bounds. The computations are implemented in R as follows:

6) It is also possible to use the adjusted instead of the raw means. However, in that case the function to obtain the equivalence bounds in Listing 7 would have to be adapted.

Listing 7

Equivalence Testing and IHT

```
# this R code can be found under 05TOST.R on the OSF project site
library(TOSTER)
library(pwr)

# frequency table for group
table(Data$group)
n1 <- 200 # sample size group '1'
n2 <- 500 # sample size group '2'

# determining the equivalence bounds via the small telescopes approach
d.33 <- (pwr.t.test(
    n = (n1 + n2) / 2,
    d = NULL,
    sig.level = 0.05,
    power = 0.33,
    type = "two.sample",
    alternative = "two.sided"
))$d
# d.33 = 0.1150074

m1 <- mean(Data$posttest[Data$group == 1])
# 0.438377
m2 <- sd(Data$posttest[Data$group == 1])
# 1.017934
sd1 <- mean(Data$posttest[Data$group == 2])
# 0.7120272
sd2 <- sd(Data$posttest[Data$group == 2])
# 1.047107
m2 - m1
# 0.5795566

# using the raw means and sd
tsum_TOST(
    m1 = m1, m2 = m2, sd1 = sd1, sd2 = sd2,
    n1 = n1, n2 = n2, eqb = d.33, alpha = 0.05, var.equal = FALSE
)
# partial output:
# TOST Results
#                    t       df     p.value
# t-test        -8.429    533.1     < 0.001
# TOST Lower    -6.756    533.1           1
# TOST Upper   -10.101    533.1     < 0.001
```

The difference between the raw means is (about) 0.58 and the equivalence bounds are set to $\Delta_L = -0.115$ and $\Delta_U = 0.115$. Using the TOST procedure includes two Welch t-tests. The test against the upper equivalence bound tests H_0:$0.58 - \Delta_U = 0$ against

$H_1: 0.58 - \Delta_U \neq 0$ and the test against the lower equivalence bound tests $H_0: 0.58 - \Delta_L = 0$ against $H_1: 0.58 - \Delta_L \neq 0$.

In our example, the test against the upper equivalence bound is significant, $t(533.1) = -10.101$, $p < .001$, whereas the test against the lower equivalence bound is non significant, $t(533.1) = -6.756$, $p = 1.0$. Since the conclusion of equivalence can only be drawn if both tests are significant, we cannot reject the presence of a difference in raw means between the groups $X = 1$ and $X = 2$. The classical test from NHST is significant, $t(533.1) = -8.429$, $p < .001$, indicating that the two groups $X = 1$ and $X = 2$ statistically differ with respect to their raw means.

In the following sections, we present two empirical data examples that serve to further demonstrate IHT in the EffectLiteR framework. The first example is in the context of linear regression and the second example is in the context of the generalized linear model.

Empirical Example on Linear Regression

The empirical example in this section is based on Keck et al. (2022). We used the "ACTG175" data set (Hammer et al., 1996), which comes with the R package speff2trial (Juraska, 2022) and originates from a randomized trial. For the sake of our illustration, we have changed the names of the variables with names that are more common in psychology. More precisely, let us assume that the treatment groups correspond to a group receiving an old, established vocational training program ($X = 0$) and a group receiving a promising, novel vocational training program ($X = 1$). The outcome variable Y is a measure of job satisfaction, and, being measured by a freehand continuous line scale, ranges from 0 till 787. As covariates, we consider a categorical variable indicating whether a subject has already completed vocational trainings in the past or not, and a continuous variable describing job satisfaction at baseline. We analyse a subset of the data from subjects currently holding a full-time job and exclude all other subjects as well as cases with missing data, which leads to a total sample size of $n = 236$. Note that our approach has not been fully tested to handle missing data, which is why we exclude the incomplete cases from the data set. The full R code for this example (including renaming the variable names) can be found on the OSF website (see Keck et al., 2024). Below, we only show the most relevant parts for the sake of illustration. To fit this model using EffectLiteR, we can use the following R code:

Listing 8

Fitting the Model Using EffectLiteR

```
# This R code can be found under 06ex1.R on the OSF project site

elrmod <- effectLite(
    y = "jobsatisfaction",
    x = "treatment",
    k = "past.training",
    z = "baseline",
    method = "sem",
    fixed.cell = TRUE, fixed.z = TRUE,
    homoscedasticity = TRUE,
    data = Data
)
elrmod@results@adjmeans
#           Estimate       SE  Est./SE
# adjmean0 259.4614 12.83423 20.21636
# adjmean1 299.8482 12.18761 24.60271
```

We again used `method = "sem"` as this is needed for the `effectLite_iht()` function. We can observe that the adjusted means for the control group ($X = 0$) and treatment group ($X = 1$) are $\widehat{AdjM}_0 = 259.46$ and $\widehat{AdjM}_1 = 299.85$ respectively.

Our first hypothesis of interest is that the adjusted mean of the treatment group ($X = 1$) is larger than the adjusted mean of the control group ($X = 0$). This is a Type A hypothesis and we have observed that our constraint is indeed satisfied in the data. Therefore, we can test our hypothesis of interest right away without testing a Type B hypothesis first. We test

$$H_0: AdjM_1 = AdjM_0 \tag{25}$$

against

$$H_1: AdjM_1 > AdjM_0 \tag{26}$$

using IHT and against

$$H_1: AdjM_1 \neq AdjM_0 \tag{27}$$

using NHST. To test this informative hypothesis, we can use the following R code:

Listing 9

Informative Hypothesis Test Using Adjusted Means

```
# This R code can be found under 06ex1.R on the OSF project site

effectLite_iht(object = elrmod, constraints = "adjmean1 > adjmean0")
# $test.stat
# [1] "Fbar"
#
# $Wald.info
# [1] 5.206926
#
# $pvalue
# [1] 0.01170985
```

The hypothesis is expressed in terms of adjusted means (for $X = 0$ and $X = 1$), but because the average effect (AE_{10}) is simply the difference between these two adjusted means, we can also formulate our hypothesis test as follows:

Listing 10

Informative Hypothesis Test Using the Average Effect

```
# This R code can be found under 06ex1.R on the OSF project site

effectLite_iht(object = elrmod, constraints = "Eg1 > 0")
# $test.stat
# [1] "Fbar"
#
# $Wald.info
# [1] 5.206926
#
# $pvalue
# [1] 0.01170985
```

Here, as before, the keyword `Eg1` represents the average effect AE_{10}. In both cases, we obtain $Wald_{info} = 5.21$, $p = .012$, allowing us to reject the null hypothesis in favor of the alternative. Note that if we would ignore the order, the resulting (non-informative) Wald statistic would still be 5.21 (because the constraints are satisfied in the data). But the p-value would be twice as large (in this case). This demonstrates the greater power which is typically obtained when using IHT compared to NHST.

The second hypothesis of interest is that the difference in adjusted means between the treatment ($X = 1$) and control group ($X = 0$) for subjects who have already completed vocational trainings in the past ($K = 1$) is larger than zero. Again, this is a Type A

hypothesis, but this time regarding a conditional effect, which is defined as (Mayer et al., 2016):

$$CE_{10|K=1} = E[E(Y|X=1,K=1,Z)|K=1] - [E(Y|X=0,K=1,Z)|K=1] \quad (28)$$

$$= AdjM_{1|K=1} - AdjM_{0|K=1}. \quad (29)$$

To obtain the adjusted means for the combinations of the levels of X and K, the following R code can be used:

Listing 11

Fitting the Model Using EffectLiteR

```
# This R code can be found under 06ex1.R on the OSF project site

elrmod@results@adjmeansgk
#              Estimate      SE   Est./SE
# adjmean0gk0 210.3477 45.02987  4.671293
# adjmean1gk0 287.6858 34.86036  8.252521
# adjmean0gk1 265.0214 13.34676 19.856606
# adjmean1gk1 301.2251 12.98068 23.205644
```

We can observe that our constraint is satisfied in the data: $\widehat{AdjM}_{1|K=1} = 301.23$ is larger than $\widehat{AdjM}_{0|K=1} = 265.02$. Therefore, we do not need to test a Type B hypothesis first, before testing our hypothesis of interest. We test

$$H_0: AdjM_{1|K=1} = AdjM_{0|K=1} \quad (30)$$

against

$$H_1: AdjM_{1|K=1} > AdjM_{0|K=1} \quad (31)$$

using IHT and against

$$H_1: AdjM_{1|K=1} \neq AdjM_{0|K=1} \quad (32)$$

using NHST:

Listing 12

Informative Hypothesis Test Regarding a Conditional Effect

```
# This R code can be found under 06ex1.R on the OSF project site

# in terms of adjusted means
effectLite_iht( object = elrmod, constraints = "adjmean1gk1 > adjmean0gk1")
# $test.stat
# [1] "Fbar"
#
# $Wald.info
# [1] 3.781235
#
# $pvalue
# [1] 0.02653014

# in terms of conditional effects
effectLite_iht(object = elrmod, constraints = "Eg1gk1 > 0")
# $test.stat
# [1] "Fbar"
#
# $Wald.info
# [1] 3.781235
#
# $pvalue
# [1] 0.02653014
```

Using both IHT and NHST, we obtain $Wald_{info} = Wald_{reg} = 3.78$. Furthermore, we obtain $p = .026$ when using IHT and $p = .052$ when using NHST. In this case, the greater power of IHT compared to NHST does make a difference concerning the significance of the results.

Empirical Example on the Generalized Linear Model

The empirical example in this section is based on Keck et al. (2023). We used the "ProblemDrinking" data set, which is available on the OSF project site ("problemDrinking.sav") (see Keck et al., 2024). It stems from a randomized study investigating the effectiveness of mobile messaging interventions on problematic drinking behavior (Muench et al., 2017). We consider three groups, namely a control group ($X = 0$), which receives weekly self-tracking texts, a group obtaining static tailored texts ($X = 1$) and a group obtaining adaptive, that is individually tailored texts ($X = 2$). The outcome variable Y is the reduction of the sum of weekly drinks. We treat Y as a count variable. As covariates, we consider the sum of weekly drinks at baseline, age and gender.

The full R script of this example, "07ex2.R", can be found in the project's OSF repository (see Keck et al., 2024). Note that the computation of the average and conditional effects in this section is based on Poisson regression and thus differs from the computation in linear regression. At the time of writing, the EffectLiteR package does not include (out of the box) support for Poisson outcome variables yet. Instead, we provided a script "effectLite_pois.R" that will take care of the computations for this particular example, and is used in the "07ex2.R" script.

We start with using `glm()` to fit the Poisson model:

Listing 13

Using glm() to Fit a Poisson Model

```
# This R code can be found under 07ex2.R on the OSF project site

fit.glm <- glm(formula = drinksum_post ~ treat + drinksum_pre + age + gender +
                treat:drinksum_pre + treat:age + treat:gender +
                drinksum_pre:age + drinksum_pre:gender + age:gender,
                family = "poisson"(link = "log"), data = Data)
```

Our first hypothesis of interest is that the average effect of receiving adaptive tailored texts ($X = 2$) is larger than the average effect of receiving static tailored texts ($X = 1$). We first compute the adjusted means:

Listing 14

Adjusted Means for the Three Groups

```
# This R code can be found under 07ex2.R on the OSF project site

get_adjmeans(fit.glm)
# adjmean0 adjmean1 adjmean2
# 22.02197 17.24442 15.26413
```

From these adjusted means, we can compute the average effects. For $X = 1$, the average effect (in terms of reduction) is $\widehat{AE}_{10} = 4.78$ (the difference between `adjmean0` and `adjmean1`), while for $X = 2$, the average effect is $\widehat{AE}_{10} = 6.76$ (the difference between `adjmean0` and `adjmean2`). This is in line with our expectations, and we can proceed with a Type A hypothesis test. We test

$$H_0: AE_{20} = AE_{10} \tag{33}$$

against

$$H_1: AE_{20} > AE_{10} \tag{34}$$

using IHT and against

$$H_1: AE_{20} \neq AE_{10} \tag{35}$$

using NHST.

Listing 15

Informative Wald Statistic for This Hypothesis

```
# This R code can be found under 07ex2.R on the OSF project site

Wald.reg.ave <- getStat(fit.glm, type = "regular", effect = "average")
Wald.reg.ave[1]
# 27.93786

# p-value regular Wald
1 - pchisq (Wald.reg.ave[1], df = 1)
# 1.252745e-07

Wald.info.ave <- getStat(fit.glm, type = "informative", effect = "average")
Wald.info.ave[1]
# 27.93786

# informative p-value (warning: takes about 14-18 hours)
# pvalue <- get_informative_pvalue(object = fit.glm, data = Data, R = 1000,
#                                  effect = "average",
#                                  Wald.orig = Wald.info.ave [1])
# pvalue
# 0
```

The informative Wald statistic equals $Wald_{info} = 27.94$. Because the constraint is satisfied in the data, the regular (non-informative) Wald statistic is the same ($Wald_{reg} = 27.94$). The p-value for the regular Wald test is easy to compute, and is very small ($p < .001$). The computation of the p-value for the informative test takes a long time (about 14–18 hours), but results again in a very small p-value ($p < .001$).

The second hypothesis of interest is that the difference in adjusted means between the group receiving individually tailored texts ($X = 2$) and the control group ($X = 0$) is larger for females ($K = 0$) than for males ($K = 1$). This is a Type A hypothesis concerning conditional effects, which are defined as (Mayer et al., 2016):

$$CE_{20|K=0} = E[E(Y|X=2,K=0,Z)|K=0] - E[E(Y|X=0,K=0,Z)|K=0] \quad (36)$$

$$= AdjM_{2|K=0} - AdjM_{0|K=0}, \quad (37)$$

$$CE_{20|K=1} = E[E(Y|X=2,K=1,Z)|K=1] - E[E(Y|X=0,K=1,Z)|K=1] \quad (38)$$

$$= AdjM_{2|K=1} - AdjM_{0|K=1}. \quad (39)$$

We can compute the adjusted means for the various combinations of X and K as follows:

Listing 16

Adjusted Means for the Different X and K Levels

```
# This R code can be found under 07ex2.R on the OSF project site

get_adjmeansgk(fit.glm)
# adjmean0gk0  adjmean1gk0  adjmean2gk0  adjmean0gk1  adjmean1gk1  adjmean2gk1
#    22.78305     17.89236     14.79266     22.21254     17.24841     18.87397
```

We observe that our constraint is again satisfied in the data, since $\widehat{CE}_{20|K=0} = 7.99$ (`adjmean2gk0 - adjmean0gk0`) is larger than $\widehat{CE}_{20|K=1} = 3.34$ (`adjmean0gk1 - adjmean2gk1`). Therefore, we directly test:

$$H_0: CE_{20|K=0} = CE_{20|K=1} \quad (40)$$

against

$$H_1: CE_{20|K=0} > CE_{20|K=1} \quad (41)$$

using IHT and against

$$H_1: CE_{20|K=0} \neq CE_{20|K=1} \quad (42)$$

using NHST. To compute the informative and regular Wald statistics, we can use the following code:

Listing 17

Informative and Regular Wald Test for Conditional Effects

```
# This R code can be found under 07ex2.R on the OSF project site

Wald.reg.cond <- getStat(fit.glm, type = "regular", effect = "conditional")
Wald.reg.cond[1]
# 1.75101

# p-value regular Wald
1 - pchisq(Wald.reg.cond[1], df = 1)
# 0.1857499

Wald.info.cond <- getStat(fit.glm, type = "informative", effect = "conditional")
Wald.info.cond[1]
# 1.75101

# informative p-value (warning: takes about 4-5 hours)
# pvalue <- get_informative_pvalue(object = fit.glm, data = Data, R = 1000,
#                                  effect = "conditional",
#                                  Wald.orig = Wald.info.cond[1])
# pvalue
# 0.104
```

The Wald statistics are $Wald_{reg} = Wald_{info} = 1.75$. The p-value for the regular Wald statistic is $p = .186$. The p-value for the informative Wald statistic is $p = .104$. Again, computing the latter p-values takes a long time. In both cases, we cannot reject the null hypothesis.

Discussion

This paper provided a condensed outline of the theoretical motivation for using IHT in the EffectLiteR framework as well as practical instructions on how to apply this method in the context of linear regression and the generalized linear model. We hope that this paper will stimulate researchers to question the common practice of using ANOVA in combination with NHST to compare groups. Our critique of this procedure is mainly focused on two aspects: The first point of criticism is focused on the unclear definitions of effects due to the different possible choices of sum of squares (SS) in ANOVA. In contrast, when using our proposed method, effects of interest are defined in a precise and unambiguous way. The second point of criticism refers to the expected order of the effects that is ignored when using NHST. In contrast, when using our proposed method, the order of the effects can be considered directly in the hypotheses.

Snippets of R code were shown in the various code listings included in the paper to illustrate how the EffectLiteR package can be used to test informative hypothesis about

adjusted means, average, and conditional effects. The full R code for all examples is available on the OSF project site (see Keck et al., 2024). Only for the generalized linear model example did we provide custom R code that needs to be adapted by the user. In future work, we plan to create easy to use functions (within EffectLiteR) that can handle IHT in the context of generalized linear models.

Together with our past work (Keck et al., 2021, 2022, 2023), we have provided thorough technical explanations as well as useful practical information and instructions for applied researchers who wish to use IHT in the EffectLiteR framework. We have built a solid foundation of our method when using regression models and would like to expand our method to Structural Equation Modeling (SEM) in the future. Some of the ground work for this has already been done in Keck et al. (2021), where we used SEM for parameter estimation when considering stochastic group weights (Mayer & Thoemmes, 2019). Further implementing our method in SEM will be especially useful since most variables of interest in the social and behavioral sciences, such as "quality of life" or "socio-economic status", are latent in nature and should not be treated as manifest.

Another potential area for further development is extending the presented approach to a Bayesian framework. In this manuscript, we focused on the frequentist approach. Here, the EffectLiteR model is estimated using either OLS or ML in the example with the continuous dependent variable, and IWLS in the example with the count dependent variable. Furthermore, informative test statistics are used. Both aspects have Bayesian counterparts: The regression models used can be estimated using Bayesian techniques. For an example of a Bayesian EffectLiteR application using blavaan (Merkle et al., 2021), see Mayer et al. (2017). Furthermore, informative hypotheses can be considered in a Bayesian framework using Bayes factors (e.g., Hoijtink, 2012; Van Lissa et al., 2020). Combining Bayesian EffectLiteR and Bayesian informative hypothesis testing is promising and may provide even more flexibility, in particular when more and specific prior information is available that can be incorporated in the analysis.

Funding: This work has been supported by the Research Foundation Flanders (FWO, Grant G020115N to Yves Rosseel and Axel Mayer).

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist. An earlier version of this paper was included as Chapter 5 in the unpublished PhD dissertation of Caroline Keck.

Data Availability: The simulated data set that has been used for the running example as well as the empirical data set that has been used in the generalized linear regression example are available at Keck et al. (2024) under the names "runningExampleData.csv" and "problemDrinking.sav". The empirical data set that has been used in the linear regression example can be obtained via the R package speff2trial.

Supplementary Materials

For this article, the following Supplementary Materials are available:

- R codes. (Keck et al., 2024)
- Empirical data set. (Keck et al., 2024)
- Simulated data set. (Keck et al., 2024)

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1–12. https://doi.org/10.1037/met0000051

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1974). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444–455. https://doi.org/10.2307/2291629

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions*. Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression / correlation analysis in the behavioral sciences*. Erlbaum. https://doi.org/10.4324/9780203774441

Edwards, L. (Ed.). (1993). *Applied analysis of variance in behavioral science, Vol. 137*. CRC.

Flunger, B., Mayer, A., & Umbach, N. (2019). Beneficial for some or for everyone? Exploring the effects of an autonomy-supportive intervention in the real-life classroom. *Journal of Educational Psychology, 111*(2), 210–234. https://doi.org/10.1037/edu0000284

Fox, J. (2016). *Applied regression analysis and generalized linear models*. SAGE.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., . . . Zeileis, A. (2022). *Car: Companion to applied regression* [R package version 3.1-1]. https://CRAN.R-project.org/package=car

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*(3), 527–537. https://doi.org/10.1348/000711009X475853

Graefe, L., Hahn, S., & Mayer, A. (2022). On the relationship between ANOVA main effects and average treatment effects. *Multivariate Behavioral Research, 58*(3), 467–485. https://doi.org/10.1080/00273171.2022.2068122

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of Statistical Software, 33*(10), 1–31. https://doi.org/10.18637/jss.v033.i10

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., & Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine, 335*(15), 1081–1090. https://doi.org/10.1056/NEJM199610103351501

Hardy, M. A. (2003). *Regression with dummy variables*. SAGE. https://doi.org/10.4135/9781412985628.

Hector, A., von Felten, S., & Schmid, B. (2010). Analysis of variance with unbalanced data: An update for ecology & evolution. *Journal of Animal Ecology, 79*(2), 308–316. https://doi.org/10.1111/j.1365-2656.2009.01634.x

Herr, D. G., & Gaebelein, J. W. (1978). Nonorthogonal two-way analysis of variance. *Psychological Bulletin, 85*(1), 207–216. https://doi.org/10.1037/0033-2909.85.1.207

Hoijtink, H. (2012). *Informative hypotheses. Theory and practice for behavioral and social scientists*. Chapman & Hall/CRC.

IBM Corporation. (2020). *IBM SPSS statistics for Windows (version 27.0)*. IBM Corporation.

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

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Juraska, M. (2022). *Speff2trial: Semiparametric estimation for a two-sample treatment effect* [R package Version 1.0.5]. https://CRAN.R-project.org/package=speff2trial

Keck, C., Mayer, A., & Rosseel, Y. (2021). Integrating informative hypotheses into the EffectLiteR framework. *Methodology, 17*(4), 307–325. https://doi.org/10.5964/meth.7379

Keck, C., Mayer, A., & Rosseel, Y. (2022). Overview and evaluation of various frequentist test statistics using constrained statistical inference in the context of linear regression. *Frontiers in Psychology, 13*, Article 899165. https://doi.org/10.3389/fpsyg.2022.899165

Keck, C., Mayer, A., & Rosseel, Y. (2023). Evaluation of frequentist test statistics using constrained statistical inference in the context of the generalized linear model. *Health Psychology and Behavioral Medicine, 11*(1), Article 2222164. https://doi.org/10.1080/21642850.2023.2222164

Keck, C., Rosseel, Y., & Mayer, A. (2024). *Informative hypothesis testing in the EffectLiteR framework: A tutorial* [OSF project page containing R codes, empirical data set, and simulated data set]. OSF. https://osf.io/qk9hu/

Keselman, H. J., Miller, C. W., & Holland, B. (2011). Many tests of significance: New methods for controlling Type I errors. *Psychological Methods, 16*(4), 420–431. https://doi.org/10.1037/a0025810

Kuiper, R., Nederhoff, T., & Klugkist, I. (2015). Properties of hypothesis testing techniques and (Bayesian) model selection for exploration-based and theory-based (order-restricted) hypotheses. *British Journal of Mathematical and Statistical Psychology, 68*(2), 220–245. https://doi.org/10.1111/bmsp.12041

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., & Caldwell, A. (2022). *TOSTER: Two one-sided tests (TOST) equivalence testing* [R package Version 0.6.0]. https://CRAN.R-project.org/package=toster

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lenth, R. V., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2022). *Emmeans: Estimated marginal means, aka least-squares means* [R package Version 1.8.2]. https://CRAN.R-project.org/package=emmeans

Macnaughton, D. B. (1998). *Which sum of squares are best in an unbalanced analysis of variance?* MatStat Research Consulting. https://pdfs.semanticscholar.org/a90a/ddae5a32e69fd174b38e1c9d32ec417a38e4.pdf

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

Mayer, A., Zimmermann, J., Hoyer, J., Salzer, S., Wiltink, J., Leibing, E., & Leichsenring, F. (2020). Interindividual differences in treatment effects based on Structural Equation Models with latent variables: An effectliter tutorial. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(5), 798–816. https://doi.org/10.1080/10705511.2019.1671196

Mayer, A., & Dietzfelbinger, L. (2019). *EffectLiteR: Average and conditional effects* [R package Version 0.4-4]. https://CRAN.R-project.org/package=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. https://doi.org/10.1080/00273171.2016.1151334

Mayer, A., & Thoemmes, F. (2019). Analysis of variance models with stochastic group weights. *Multivariate Behavioral Research, 54*(4), 542–554. https://doi.org/10.1080/00273171.2018.1548960

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

Merkle, E., Rosseel, Y., & Goodrich, B. (2021). *blavaan: Bayesian latent variable analysis* [R package Version 0.5-5]. https://CRAN.R-project.org/package=blavaan

Meyners, M. (2012). Equivalence tests - a review. *Food Quality and Preference, 26*(2), 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003

Mueller, B., Mayer, A., Richter, T., Križan, A., Hecht, T., & Ennemoser, M. (2015). Differential effects of reading trainings on reading processes: A comparison in Grade 2. *Zeitschrift fuer Erziehungswissenschaft, 18*, 489–512. https://doi.org/10.1007/s11618-015-0648-0

Muench, F., van Stolk-Cooke, K., Kuerbis, A., Stadler, G., Baumel, A., Shao, S., McKay, J. R., & Morgenstern, J. (2017). A randomized controlled pilot trial of different mobile messaging

interventions for problem drinking compared to weekly drink tracking. *PLoS ONE, 12*(2), Article e0167900. https://doi.org/10.1371/journal.pone.0167900.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science, 5*(4), 465–472. https://doi.org/10.1214/ss/1177012031

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96–146. https://doi.org/10.1214/09-SS057

Quertemont, E. (2011). How to statistically show the absence of an effect. *Psychologica Belgica, 51*(2), 109–127. https://doi.org/10.5334/pb-51-2-109

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rek, K., Kappelmann, N., Zimmermann, J., Rein, M., Egli, S., & Kopf-Beck, J. (2022). Evaluating the role of maladaptive personality traits in schema therapy and cognitive behavioral therapy for depression. *Psychological Medicine, 53*(10), 1–10. https://doi.org/10.1017/S0033291722001209

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553–565. https://doi.org/10.1037/0033-2909.113.3.553

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

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistician, 100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA*. SAGE Publications.

Sadovich, V. (2020). *Fehlertolerante Anzeigengestaltung für Augmented Reality Head-up-Displays*. Springer.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science, 16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657–680. https://doi.org/10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*(4), 403–411. https://doi.org/10.1037/1082-989X.3.4.403

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions*. Wiley.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online, 5*(2), 39–71.

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. 606–630). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_295

Van Lissa, C., Gu, X., Mulder, J., Rosseel, Y., Van Zundert, C., & Hoijtink, H. (2020). Techer's corner: Evaluating informative hypotheses using the Bayes factor in structural equation models. *Structural Equation Modeling: A Multidiciplinary Journal, 28*(2), 292–301. https://doi.org/10.1080/10705511.2020.1745644

Vanbrabant, L. (2020). *Restriktor: Constrained statistical inference* [R package Version 0.2-800]. https://CRAN.R-project.org/package=restriktor

Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: Sample-size tables for ANOVA and regression. *Frontiers in Psychology, 5*, Article 1565. https://doi.org/10.3389/fpsyg.2014.01565.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press.

Yong, E. (2012). Bad copy: In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature, 485*, 298–300. https://doi.org/10.1038/485298a

Appendices

Appendix A

Results of the Linear Model Using Effect Coding

Table A1

Linear Regression Model Results Using Sum Coding

Source		Coefficient	Estimate (*SE*)	*t*-value	*p*-value
Intercept		$\hat{\beta}_0$	0.400 (0.038)	10.652	< .001***
	$I_{X=1}$	$\hat{\beta}_1$	−0.352 (0.052)	−6.798	< .001***
Group					
	$I_{X=2}$	$\hat{\beta}_2$	0.084 (0.061)	1.387	.166
Treatexp		$\hat{\beta}_3$	−0.131 (0.038)	−3.478	< .001***
Pre-test		$\hat{\beta}_4$	0.152 (0.036)	4.252	< .001***
	$I_{X=1}$	$\hat{\beta}_5$	0.089 (0.052)	1.715	.087
Group:treatexp					
	$I_{X=2}$	$\hat{\beta}_6$	0.032 (0.061)	−0.533	.594
	$I_{X=1}$	$\hat{\beta}_7$	−0.083 (0.050)	−1.660	.097
Group:pre-test					
	$I_{X=2}$	$\hat{\beta}_8$	−0.051 (0.056)	−0.912	.362
Treatexp:pre-test		$\hat{\beta}_9$	0.004 (0.035)	0.127	.899

Appendix B

Interpretation of Regression Coefficients

In general, the expected value of Y given X, K and Z is defined as:

$$\begin{aligned} E(Y \mid X, K, Z) = \beta_0 + \beta_1 \cdot X_{X=1} + \beta_2 \cdot X_{X=2} + \beta_3 \cdot K_{K=1} + \beta_4 \cdot Z + \\ \beta_5 \cdot X_{X=1} \cdot K_{K=1} + \beta_6 \cdot X_{X=2} \cdot K_{K=1} + \\ \beta_7 \cdot X_{X=1} \cdot Z + \beta_8 \cdot X_{X=2} \cdot Z + \beta_9 \cdot K_{K=1} \cdot Z\,. \end{aligned} \tag{B1}$$

Furthermore, β_0 is generally defined as follows:

$$\beta_0 = E(Y \mid X = 0, K = 0, Z = 0)\,. \tag{B2}$$

In our running example, Z is mean-centered, which implies that $E(Z) = 0$. Moreover, $E(Y \mid X = 0, K = 0)$ is defined as:

$$E(Y \mid X = 0, K = 0) = E[E(Y \mid X = 0, K = 0, Z) \mid X = 0, K = 0] \tag{B3}$$

$$= E[\beta_0 + \beta_4 \cdot Z \,|\, X = 0, K = 0] \tag{B4}$$

$$= \beta_0 + \beta_4 \cdot E(Z \,|\, X = 0, K = 0)\,. \tag{B5}$$

In our running example, $E(Z) = E(Z \,|\, X = 0, K = 0) = 0$. As a consequence, $E(Y \,|\, X = 0, K = 0) = \beta_0$.

Appendix C

EffectLiteR Output

---------- Variables ----------

Outcome variable Y: posttest
Treatment variable X: group (Reference group: 0)
Categorical covariates K: treatexp
Continuous covariates in Z = (Z1): Z1 = pretest.cent

Levels of Treatment Variable X

X	group (original)	Indicator
0	0	I_X=0
1	1	I_X=1
2	2	I_X=2

Levels of Unfolded Categorical Covariate K

K	treatexp	Indicator
0	0	I_K=0
1	1	I_K=1

Cells

	group (original)	K	Cell
1	0	0	00
2	0	1	01
3	1	0	10
4	1	1	11
5	2	0	20
6	2	1	21

---------- Regression Model ----------

E(Y|X,K,Z) = g0(K,Z) + g1(K,Z)*I_X=1 + g2(K,Z)*I_X=2
g0(K,Z) = g000 + g001 * Z1 + g010 * I_K=1 + g011 * I_K=1 * Z1
g1(K,Z) = g100 + g101 * Z1 + g110 * I_K=1 + g111 * I_K=1 * Z1
g2(K,Z) = g200 + g201 * Z1 + g210 * I_K=1 + g211 * I_K=1 * Z1

Intercept Function g0(K,Z) [Reference group: 0]

Coefficient	Estimate	SE	Est./SE	p-value
g000	0.006	0.071	0.091	0.928
g001	0.073	0.067	1.096	0.273
g010	0.084	0.123	0.680	0.497
g011	-0.009	0.070	-0.127	0.899

Effect Function g1(K,Z) [group: 1 vs. 0]

Coefficient	Estimate	SE	Est./SE	p-value
g100	0.380	0.109	3.490	0.001
g101	0.031	0.096	0.327	0.743
g110	0.113	0.206	0.548	0.584
g111	0.000	NA	NA	NA

Effect Function g2(K,Z) [group: 2 vs. 0]

Coefficient	Estimate	SE	Est./SE	p-value
g200	0.410	0.101	4.070	0.000
g201	0.217	0.078	2.771	0.006
g210	0.420	0.154	2.729	0.006
g211	0.000	NA	NA	NA

---------- Cell Counts ----------

Cell Counts

This table shows cell counts including missings.
See also output under lavaan results for number of observations
actually used in the analysis.

treatexp	0	1
group		
0	200	100
1	150	50
2	200	300

---------- Main Hypotheses ----------

H0: No average effects: E[g1(K,Z)] = E[g2(K,Z)] = 0
H0: No covariate effects in control group: g0(K,Z) = constant
H0: No treatment*covariate interaction: g1(K,Z), g2(K,Z) = constant
H0: No treatment effects: g1(K,Z) = g2(K,Z) = 0

	F value	df1	df2	p-value
No average effects	31.123	2	990	7.79e-14
No covariate effects in control group	0.589	3	990	6.22e-01
No treatment*covariate interaction	4.334	4	990	1.77e-03
No treatment effects	13.433	6	990	1.09e-14

---------- Adjusted Means ----------

	Estimate	SE	Est./SE
Adj.Mean0	0.0442	0.0599	0.738
Adj.Mean1	0.4746	0.0786	6.037
Adj.Mean2	0.6430	0.0471	13.651

---------- Average Effects ----------

	Estimate	SE	Est./SE	p-value	Effect Size

E[g1(K,Z)]	0.430	0.0988	4.35	1.47e-05	0.412
E[g2(K,Z)]	0.599	0.0762	7.86	1.02e-14	0.573

---------- Effects given a Treatment Condition ----------

	Estimate	SE	Est./SE	p-value	Effect Size
E[g1(K,Z)| X=0]	0.418	0.0930	4.49	7.85e-06	0.400
E[g2(K,Z)| X=0]	0.554	0.0775	7.14	1.82e-12	0.530
E[g1(K,Z)| X=1]	0.409	0.0925	4.42	1.10e-05	0.391
E[g2(K,Z)| X=1]	0.522	0.0810	6.45	1.80e-10	0.500
E[g1(K,Z)| X=2]	0.447	0.1137	3.93	9.15e-05	0.428
E[g2(K,Z)| X=2]	0.657	0.0806	8.15	1.11e-15	0.629

---------- Effects given K = k ----------

	Estimate	SE	Est./SE	p-value	Effect Size
E[g1(K,Z)| K=0]	0.380	0.109	3.49	5.04e-04	0.363
E[g2(K,Z)| K=0]	0.410	0.101	4.07	5.12e-05	0.392
E[g1(K,Z)| K=1]	0.493	0.175	2.82	4.94e-03	0.472
E[g2(K,Z)| K=1]	0.830	0.116	7.14	1.83e-12	0.795

---------- Effects given X = x, K = k ----------

	Estimate	SE	Est./SE	p-value	Effect Size

E[g1(K,Z)|X = 0, K = 0]	0.380	0.109	3.50	4.93e-04	0.364
E[g2(K,Z)|X = 0, K = 0]	0.414	0.101	4.11	4.21e-05	0.397
E[g1(K,Z)| X=1, K=0]	0.380	0.109	3.49	5.05e-04	0.363
E[g2(K,Z)| X=1, K=0]	0.409	0.101	4.06	5.19e-05	0.392
E[g1(K,Z)| X=2, K=0]	0.379	0.109	3.48	5.17e-04	0.363
E[g2(K,Z)| X=2, K=0]	0.405	0.101	4.02	6.17e-05	0.388
E[g1(K,Z)| X=0, K=1]	0.493	0.175	2.82	4.91e-03	0.472
E[g2(K,Z)| X=0, K=1]	0.832	0.116	7.15	1.65e-12	0.796
E[g1(K,Z)| X=1, K=1]	0.497	0.175	2.84	4.54e-03	0.476
E[g2(K,Z)| X=1, K=1]	0.860	0.117	7.36	3.82e-13	0.823
E[g1(K,Z)| X=2, K=1]	0.492	0.175	2.81	5.04e-03	0.471
E[g2(K,Z)| X=2, K=1]	0.825	0.116	7.09	2.55e-12	0.789

---------- Hypotheses given K = k ----------

H0: No average effects given K=0: E[g1(K,Z)|K=0] = E[g2(K,Z)|K=0] = 0
H0: No average effects given K=1: E[g1(K,Z)|K=1] = E[g2(K,Z)|K=1] = 0

	F value	df1	df2	p-value
No average effects given K=0	9.92	2	990	5.44e-05
No average effects given K=1	25.84	2	990	1.15e-11

Appendix D

Test Statistics

To test hypotheses of interest, EffectLiteR uses the regular Wald test used in NHST:

$$Wald_{reg} = n(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}), \tag{D1}$$

where n is the sample size, $\boldsymbol{R}$ is the constraint matrix specifying the hypothesis of interest, $\hat{\boldsymbol{I}}_1$ is the unit information matrix and $\hat{\boldsymbol{\beta}}$ is the vector of unconstrained regression parameters that has been obtained using maximum likelihood estimation.

When adopting IHT, we have to use the informative Wald test:

$$Wald_{info} = n(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}). \tag{D2}$$

Note that in contrast to $Wald_{reg}$, $\hat{\boldsymbol{\beta}}$ has been replaced by $\tilde{\boldsymbol{\beta}}$, the vector of constrained regression parameters satisfying $\boldsymbol{R}\boldsymbol{\beta} \geq \boldsymbol{0}$ that has typically been obtained using quadratic programming. Under the null hypothesis, $Wald_{info}$ asymptotically follows a $\bar{\chi}^2$ distribution, which is a mixture of χ^2 distributions. The constraint matrix $\boldsymbol{R}$ looks the same in NHST and IHT.

Appendix E

Calculation of p-Values

The p-values in IHT can be computed as follows (Silvapulle & Sen, 2005, pp. 86):

$$\Pr(\bar{\chi}^2 \geq \bar{\chi}^2_{obs}) = \sum_{i=0}^{q} w_i(H_0, H_1)\Pr[(h-q+i)\chi^2_{h-q+i} \geq \bar{\chi}^2_{obs}]. \tag{E1}$$

The weight w_i is some non-negative value denoting the probability that $\tilde{\boldsymbol{\beta}}$ has exactly i elements for which the constraints are non-active. The sum of the weights from 0 to q is one. In the linear regression case, q is the rank of the design matrix $\boldsymbol{X}$ under the null hypothesis.

There are two ways to obtain the p-value. The weights approach, where the mixing weights w_i are obtained in a first step and then used in Equation E1, is more economical than the simulation approach. Both will be explained in the following.

Weights Approach

In case the residuals are normally distributed, the weights can be computed by using the multivariate normal probability distribution function. For this purpose, the ic.weight() function of the R package ic.infer (Grömping, 2010) can be used. If the residuals are not normally distributed, the weights can be obtained as follows (Silvapulle & Sen, 2005, p. 79):

1. Transform the sample data set such that it is under the null hypothesis.
2. Calculate $\tilde{\boldsymbol{\beta}}$, subject to a constraint, for example $\boldsymbol{\beta} \geq 0$.
3. Count the number of elements of the vector $\tilde{\boldsymbol{\beta}}$ that satisfy the constraint specified in Step two.
4. Repeat the previous three steps, for example $B = 100,000$ times.

5. Estimate w_i by the proportion of times $\widetilde{\boldsymbol{\beta}}$ has exactly i elements for which the constraints are non-active, for example i positive elements, with $i = 0, \ldots, q$.

Transforming the sample data set to be under the null hypothesis can be accomplished by exchanging the values of the outcome variable of interest by randomly generated values from the corresponding distribution.

Simulation Approach

To calculate the p-value for an existing sample value of the $\bar{F}$-statistic or the (generalized) informative Wald statistic, the following four steps have to be taken (Silvapulle & Sen, 2005, p. 98):

1. Transform the sample data set such that it is under the null hypothesis.
2. Calculate the informative test statistic.
3. Repeat the previous steps, for example $B = 100,000$ times.
4. Estimate the p-value by means of M/B, with M being the number of times the test statistic in the second step exceeded the sample value of interest.

Again, transforming the sample data set to be under the null hypothesis can be accomplished by exchanging the values of the outcome variable of interest by randomly generated values from the corresponding distribution. Using this approach, the advantage that any error distribution may be used for computing the p-value should be carefully weighed with the disadvantage of an increased computational cost.