Original Article

Integrating Informative Hypotheses Into the EffectLiteR Framework

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.*

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

Abstract

Using the EffectLiteR framework, researchers can test classical null hypotheses about effects of interest via Wald and F-tests, while taking into account the stochastic nature of group sizes. This paper aims at extending EffectLiteR to test informative hypotheses, assuming for example that the average effect of a new treatment is greater than the average effect of an old treatment, which in turn is greater than zero. We present a simulated data example to show two methodological novelties. First, we illustrate how to use the Fbar- and generalized linear Wald test to assess informative hypotheses. While the classical test did not reach significance, the informative test correctly rejected the null hypothesis, indicating the need to take into account the order of the treatment groups. Second, we demonstrate how to account for stochastic group sizes in informative hypotheses using the generalized non-linear Wald statistic. The paper concludes with a short data example.

Keywords

Empirical research across a broad range of disciplines aims at assessing the effectiveness of a treatment or an intervention. The interest might lie in average effects on a certain outcome or in conditional effects, given values of categorical or continuous covariates. For example, a researcher might be interested in the average effect of an extended day program on student achievement (Meyer & Van Klaveren, 2013) or the conditional effects of an anxiety intervention on symptoms and functioning measures of females versus

males (Grubbs et al., 2015). Currently, most researchers use linear regression or ANOVA (analysis of variance; Rutherford, 2001, Weiss, 2006) to analyze their data. However, this only yields distinct regression coefficients, whereas the effects of interest are often a more complicated function of these parameters. This makes interpretation of results cumbersome.

The EffectLiteR approach (Mayer & Dietzfelbinger, 2019; Mayer et al., 2016) provides a framework and an R package (R Core Team, 2020) for the estimation of average and conditional effects of a discrete treatment variable on a continuous outcome variable, conditioning on categorical and continuous covariates. This approach uses definitions of average and conditional effects from the causal inference literature (Imbens & Rubin, 2015; Steyer et al., 2014). In addition to the computation of these effects of interest and their standard errors, Wald or F-tests are used to test several hypotheses of interest, for example that (all) average effects are equal to zero, or (all) conditional effects are equal to zero in the population. A feature of the EffectLiteR approach are stochastic group weights, that have been embedded into the calculation of these effects (see, e.g., Mayer & Thoemmes, 2019). In the behavioral and social sciences, treatment group sizes often vary across samples. Stochastic group sizes should be considered whenever the exact group sizes are not determined before conducting the experiment. For an in-depth discussion on why it is important to include stochastic group sizes in general linear models and structural equation models, see Mayer and Thoemmes (2019).

A limitation of the current EffectLiteR approach is that it does not allow for testing informative hypotheses (Hoijtink, 2012; Silvapulle & Sen, 2005) yet. For example, assume there are three groups in a randomized experiment (without further covariates), where one is a control group ($X = 0$), one receives a standard treatment ($X = 1$) and one receives a novel treatment ($X = 2$) and we are interested in the group means of the dependent variable Y. The researcher expects that the group mean of the novel treatment, μ_2, is greater than the group mean of the standard treatment, μ_1, and that the group mean of the standard treatment is greater than the group mean of the control group, μ_0. Within the framework of classical null hypothesis testing, a two-step procedure has to be followed. First, we have to test $H_0: \mu_2 = \mu_1 = \mu_0$ against H_a: not H_0, that is, not all three means are equal. If we can reject H_0 in favor of H_a, a second step follows, in which we execute pairwise comparisons to determine which means are equal and which means are not equal. Thus, step one alone does not give us the answer we are looking for, and step two entails multiple testing, which brings along the risk of an inflated type I error rate. In an informative test, the a priori expectation of the order of the means is explicitly taken into account. For example, we can test the null hypothesis $H_0: \mu_2 = \mu_1 = \mu_0$ against the ordered hypothesis $H_a: \mu_2 > \mu_1 > \mu_0$ in a single step. The statistical approach to allow for these informative hypothesis tests is called constrained statistical inference (Hoijtink, 2012; Silvapulle & Sen, 2005). It is available in several R packages, including restriktor (Vanbrabant, 2020), ic.infer (Grömping, 2010) and bain (Gu et al., 2020).

PsychOpen GOLD

The goal of this paper is to integrate informative hypothesis testing into the Effect-LiteR approach. We will show how F- and Wald tests can be used in the regular as well as the informative case, where the quantities of interest can be expressed as a linear or non-linear function of the parameters. The paper is organized as follows. First, an introduction to the EffectLiteR framework will be given by means of a simulated data example along the lines of the original EffectLiteR paper (Mayer et al., 2016). It will be shown how (non-informative) hypotheses about effects of interest can currently be tested via standard F- or linear Wald tests. Second, it will be explained how to incorporate informative hypotheses into the EffectLiteR approach. For this purpose, the $\bar{F}$- and generalized linear Wald test as well as two procedures to calculate their p-values will be presented. Finally, stochastic group sizes will be included, which implies two further challenges. First, parameter estimation will be carried out by means of structural equation modeling (SEM) via the R package lavaan (Rosseel, 2012). And second, the generalized non-linear Wald test will be introduced. The paper concludes with a short artificial data example from the constrained statistical inference literature. Accompanying Supplementary Materials are provided and will be referenced throughout the paper.

Introduction Into the EffectLiteR Framework

In the following, the background of the EffectLiteR framework is explained. This will be illustrated by means of a non-randomized experiment, whose data of size $N = 1000$ are generated along the lines of Mayer et al. (2016). In the Supplementary Materials, the associated R script ("A-exampleI.R") as well as the data set itself ("B-exampleIData.rds") and detailed explanations regarding the R script ("C-explanation.pdf") can be found. Three treatment groups are considered, namely a group receiving innovative therapy ($X = 2$), a group receiving conventional therapy ($X = 1$) and a wait-list control group ($X = 0$). As covariates, the dichotomous variable gender with values male ($K = 0$) and female ($K = 1$) and the continuous covariate mental health pre-test (Z), are considered. The outcome variable is the post-test of mental health (Y). Mental health will be treated as a manifest variable. Table 1 shows the relative treatment group frequencies, the means of Z and the raw as well as the adjusted means of mental health Y.

When all variables are observed, the EffectLiteR approach first estimates a linear regression model. Typically, this regression model contains not only the main effects of X, K and Z, but also the interaction effects of $X \cdot K$, $X \cdot Z$ and $K \cdot Z$ as well as the three-way interaction effect of $X \cdot K \cdot Z$. However, when defining the model via the `effectLite()` command, the researcher can specify any preferences regarding the interactions by means of the `interactions=` argument. In our case, we fit the regression model including all possible interactions:

Table 1

Raw Means of Mental Health Post-Test E(Y), Adjusted Means of Mental Health Post-Test AdjM(Y), Raw Means of Mental Health Pre-Test E(Z) and Relative Treatment Group Frequencies P(X = x, K = k)

Group	$X = 0$	$X = 1$	$X = 2$	$X = .$
$K = 0$	$E(Y \mid X = 0, K = 0) = 1.526$	$E(Y \mid X = 1, K = 0) = 1.630$	$E(Y \mid X = 2, K = 0) = 1.821$	$E(Y \mid K = 0) = 1.725$
	$AdjM(Y \mid X = 0, K = 0) = 1.483$	$AdjM(Y \mid X = 1, K = 0) = 1.613$	$AdjM(Y \mid X = 2, K = 0) = 1.819$	$AdjM(Y \mid K = 0) = 1.695$
	$E(Z \mid X = 0, K = 0) = 0.497$	$E(Z \mid X = 1, K = 0) = 0.703$	$E(Z \mid X = 2, K = 0) = 0.692$	$E(Z \mid K = 0) = 0.650$
	$P(X = 0, K = 0) = 0.104$	$P(X = 1, K = 0) = 0.074$	$P(X = 2, K = 0) = 0.291$	$P(K = 0) = 0.469$
$K = 1$	$E(Y \mid X = 0, K = 1) = 1.704$	$E(Y \mid X = 1, K = 1) = 1.744$	$E(Y \mid X = 2, K = 1) = 1.936$	$E(Y \mid K = 1) = 1.811$
	$AdjM(Y \mid X = 0, K = 1) = 1.706$	$AdjM(Y \mid X = 1, K = 1) = 1.744$	$AdjM(Y \mid X = 2, K = 1) = 1.924$	$AdjM(Y \mid K = 1) = 1.829$
	$E(Z \mid X = 0, K = 1) = 0.503$	$E(Z \mid X = 1, K = 1) = 0.644$	$E(Z \mid X = 2, K = 1) = 0.771$	$E(Z \mid K = 1) = 0.644$
	$P(X = 0, K = 1) = 0.206$	$P(X = 1, K = 1) = 0.097$	$P(X = 2, K = 1) = 0.228$	$P(K = 1) = 0.531$
$K = .$	$E(Y \mid X = 0) = 1.644$	$E(Y \mid X = 1) = 1.695$	$E(Y \mid X = 2) = 1.872$	
	$AdjM(Y \mid X = 0) = 1.601$	$AdjM(Y \mid X = 1) = 1.683$	$AdjM(Y \mid X = 2) = 1.874$	
	$E(Z \mid X = 0) = 0.501$	$E(Z \mid X = 1) = 0.700$	$E(Z \mid X = 2) = 0.726$	
	$P(X = 0) = 0.310$	$P(X = 1) = 0.171$	$P(X = 2) = 0.519$	

Note. The last column of the table shows the expectations and probabilities where we ignore X. The last row shows the expectations and probabilities where we ignore K.

$$E(Y|X, Z, K) = \left(\gamma_{000} + \gamma_{010} \cdot K + \gamma_{001} \cdot Z + \gamma_{011} \cdot K \cdot Z\right)$$
$$+ \left(\gamma_{100} + \gamma_{110} \cdot K + \gamma_{101} \cdot Z + \gamma_{111} \cdot K \cdot Z\right) \cdot I_{X=1} \qquad (1)$$
$$+ \left(\gamma_{200} + \gamma_{210} \cdot K + \gamma_{201} \cdot Z + \gamma_{211} \cdot K \cdot Z\right) \cdot I_{X=2},$$

where $I_{X=x}$ are dummy variables indicating the treatment group and the three-digit subscripts of the regression coefficients refer to the treatment group (first subscript), the gender K (second subscript) and the pre-test of mental health Z (third subscript). Based on this regression model, we obtain estimates for the twelve γs (see Table 2). While it is common to use βs in the regression literature, the EffectLiteR approach is based on the γ notation (Mayer et al., 2016).

Table 2

Estimated Regression Parameters (Simulated Data)

$\hat{\gamma}_{000}$	$\hat{\gamma}_{001}$	$\hat{\gamma}_{010}$	$\hat{\gamma}_{011}$	$\hat{\gamma}_{100}$	$\hat{\gamma}_{101}$	$\hat{\gamma}_{110}$	$\hat{\gamma}_{111}$	$\hat{\gamma}_{200}$	$\hat{\gamma}_{201}$	$\hat{\gamma}_{210}$	$\hat{\gamma}_{211}$
1.667	-0.285	0.029	0.300	-0.252	0.591	0.261	-0.546	0.128	0.322	0.032	-0.234

All the effects of interest in the EffectLiteR approach will be written as a function of the estimated regression parameters and the (conditional) expectations of the covariates. In the next section, possible hypotheses of interest are discussed.

Formulation of Hypotheses

With a general linear hypothesis test, hypotheses concerning regression parameters like $\gamma_{100} = 0$ versus $\gamma_{100} \neq 0$ can be tested. In contrast, EffectLiteR allows for testing hypotheses concerning average and conditional effects like $AE_{10} = 0$ versus $AE_{10} \neq 0$. Here, AE_{10} is the average effect of treatment $X = 1$ compared to the control group $X = 0$. In this paper, we aim at combining EffectLiteR and informative hypotheses, which will allow for testing hypotheses like $AE_{20} = AE_{10}$ versus $AE_{20} > AE_{10}$. Regarding our illustrative example, one may be interested in the following hypotheses:

$$H_1 : AE_{20} = 0, AE_{10} = 0,$$
$$H_2 : AE_{20} \neq 0, AE_{10} \neq 0,$$
$$H_3 : AE_{20} > 0, AE_{10} > 0, \qquad (2)$$
$$H_4 : AE_{20} > AE_{10} > 0.$$

H_1 assumes that the average effect of innovative therapy (AE_{20}) and the average effect of conventional therapy (AE_{10}) equal zero. H_2 states that these average effects are not equal to zero. According to H_3, the average effects of innovative (AE_{20}) as well as conventional

therapy (AE_{10}) are greater than zero. Finally, H_4 postulates a complete ordering with the average effect of innovative therapy (AE_{20}) being greater than the average effect of conventional therapy (AE_{10}) and the average effect of conventional therapy being greater than zero. Note that H_3 represents a special case. Even though this is an informative hypothesis with two separate constraints, it can still be carried out within the framework of classical null hypothesis testing by using two one-sided tests (Casella & Berger, 2002). This is not possible anymore as soon as we impose two or more constraints at once as in H_4. The approach presented in this paper allows for testing all of the above mentioned hypotheses. This includes the calculation of the average effects by means of the model equations, which are presented in the next section.

EffectLiteR Model

The EffectLiteR model is based on intercept and effect functions (see also Steyer, 2021). Considering three treatment groups, a binary variable (K) with values 0 and 1 as well as a continuous covariate (Z), these can be formulated as follows:

$$
\begin{aligned}
E(Y \mid X, K, Z) &= g_0(K, Z) + g_1(K, Z) \cdot I_{X=1} + g_2(K, Z) \cdot I_{X=2}, \\
g_0(K, Z) &= \gamma_{000} + \gamma_{010} \cdot K + \gamma_{001} \cdot Z + \gamma_{011} \cdot K \cdot Z, \\
g_1(K, Z) &= \gamma_{100} + \gamma_{110} \cdot K + \gamma_{101} \cdot Z + \gamma_{111} \cdot K \cdot Z, \\
g_2(K, Z) &= \gamma_{200} + \gamma_{210} \cdot K + \gamma_{201} \cdot Z + \gamma_{211} \cdot K \cdot Z,
\end{aligned}
\tag{3}
$$

which corresponds to Equation 1. The intercept function $g_0(K, Z)$ represents the conditional regressive dependency of the outcome variable mental health Y on the covariates in the control group ($X = 0$). The values of the effect functions $g_1(K, Z)$ and $g_2(K, Z)$ represent the conditional treatment effects of conventional therapy ($X = 1$ versus $X = 0$) and innovative therapy ($X = 2$ versus $X = 0$) on the post-test Y, given values of the pre-test variable Z and the gender variable K. The average effects are defined as expectations of the effect functions $g_1(K, Z)$ and $g_2(K, Z)$. For example, the unconditional expectation of the $g_1(K, Z)$ effect function, $E[g_1(K, Z)]$, is the average effect of conventional therapy ($X = 1$ versus $X = 0$). The $g_1(K, Z)$ function represents the difference between the two regressions $E(Y \mid X = 1, K, Z)$ and $E(Y \mid X = 0, K, Z)$.

To be able to calculate the average effect, we need the unconditional expectation of Z, the unconditional expectation of K and the unconditional expectation of $K \cdot Z$. Note that for binary K, the unconditional expectation equals the probability of $K = 1$. The following calculations can also be found in the document "D-manualCalculations.pdf" in the Supplementary Materials, which contains a more exhaustive overview of the calculations done by EffectLiteR including all intermediate steps. The unconditional expectation of Z can be calculated as follows:

$$
\begin{aligned}
E(Z) = {} & E(Z\,|\,X = 0, K = 0) \cdot P(X = 0, K = 0) + E(Z\,|\,X = 0, K = 1) \cdot P(X = 0, K = 1) \\
& + E(Z\,|\,X = 1, K = 0) \cdot P(X = 1, K = 0) + E(Z\,|\,X = 1, K = 1) \cdot P(X = 1, K = 1) \quad (4) \\
& + E(Z\,|\,X = 2, K = 0) \cdot P(X = 2, K = 0) + E(Z\,|\,X = 2, K = 1) \cdot P(X = 2, K = 1),
\end{aligned}
$$

which yields 0.647 in our simulated data set. The unconditional probability of $K = 1$ can be obtained as:

$$
P(K = 1) = P(X = 0, K = 1) + P(X = 1, K = 1) + P(X = 2, K = 1), \tag{5}
$$

which is 0.531 in our example data set. Finally, $E(K \cdot Z)$ can be calculated as follows:

$$
\begin{aligned}
E(K \cdot Z) = {} & E(Z\,|\,X = 0, K = 1) \cdot P(X = 0, K = 1) \\
& + E(Z\,|\,X = 1, K = 1) \cdot P(X = 1, K = 1) \\
& + E(Z\,|\,X = 2, K = 1) \cdot P(X = 2, K = 1),
\end{aligned} \tag{6}
$$

which equals 0.342 in our simulated data set. We can obtain the average effect AE_{10} by taking the expectation of the effect function:

$$
\begin{aligned}
AE_{10} = {} & E\big[g_1(Z, K)\big] = E\big[\gamma_{100} + \gamma_{101} \cdot Z + \gamma_{110} \cdot K + \gamma_{111} \cdot Z \cdot K\big] \\
= {} & \gamma_{100} + \gamma_{101} \cdot E(Z) + \gamma_{110} \cdot P(K = 1) + \gamma_{111} \cdot E(K \cdot Z),
\end{aligned} \tag{7}
$$

which yields 0.082 in our simulated data set. The computation of the average effect AE_{20} can proceed analogously:

$$
\begin{aligned}
AE_{20} = {} & E\big[g_2(K, Z)\big] = E\big[\gamma_{200} + \gamma_{201} \cdot Z + \gamma_{210} \cdot K + \gamma_{211} \cdot Z \cdot K\big] \\
= {} & \gamma_{200} + \gamma_{201} \cdot E(Z) + \gamma_{210} \cdot P(K = 1) + \gamma_{211} \cdot E(K \cdot Z),
\end{aligned} \tag{8}
$$

which is 0.274 in our example data set. Furthermore, the adjusted means can be computed as follows:

$$
\begin{aligned}
AdjM(Y\,|\,X = 0) = {} & E\big[E(Y\,|\,X = 0, K, Z)\big] = E\big[g_0(K, Z)\big] \\
& = \gamma_{000} + \gamma_{001} \cdot E(Z) + \gamma_{010} \cdot P(K = 1) + \gamma_{011} \cdot E(K \cdot Z), \\
AdjM(Y\,|\,X = 1) = {} & E\big[E(Y\,|\,X = 1, K, Z)\big] = E\big[g_0(K, Z) + g_1(K, Z)\big] \\
& = \gamma_{000} + \gamma_{001} \cdot E(Z) + \gamma_{010} \cdot P(K = 1) + \gamma_{011} \cdot E(K \cdot Z) \\
& \quad + \gamma_{100} + \gamma_{101} \cdot E(Z) + \gamma_{110} \cdot P(K = 1) + \gamma_{111} \cdot E(K \cdot Z), \\
AdjM(Y\,|\,X = 2) = {} & E\big[E(Y\,|\,X = 2, K, Z)\big] = E\big[g_0(K, Z) + g_2(K, Z)\big] \\
& = \gamma_{000} + \gamma_{001} \cdot E(Z) + \gamma_{010} \cdot P(K = 1) + \gamma_{011} \cdot E(K \cdot Z) \\
& \quad + \gamma_{200} + \gamma_{201} \cdot E(Z) + \gamma_{210} \cdot P(K = 1) + \gamma_{211} \cdot E(K \cdot Z).
\end{aligned} \tag{9}
$$

PsychOpen GOLD

In our simulated data set, the adjusted means equal 1.601, 1.683 and 1.875. As can be seen from the equations, the average effect of conventional therapy equals the difference of the adjusted means of the groups $X = 1$ and $X = 0$, while the average effect of innovative therapy equals the difference of the adjusted means of the groups $X = 2$ and $X = 0$. The columns of Table 3 titled "Fixed group sizes" show the EffectLiteR estimates of all adjusted means and all average effects, including their standard errors. The columns titled "Stochastic group sizes" show the lavaan estimates when considering stochastic group sizes, which will be addressed later.

Table 3

Adjusted Means and Average Effects Estimates With Standard Errors in Parentheses (Simulated Data)

	Fixed group sizes		**Stochastic group sizes**	
Group	**Adjusted mean**	**Average effect**	**Adjusted mean**	**Average effect**
Control ($X = 0$)	1.601 (0.099)		1.601 (0.099)	
Conventional therapy ($X = 1$)	1.683 (0.127)	0.082 (0.161)	1.683 (0.128)	0.082 (0.162)
Innovative therapy ($X = 2$)	1.875 (0.074)	0.274 (0.124)	1.875 (0.075)	0.274 (0.124)

Note. The average effect of conventional therapy ($X = 1$ versus $X = 0$) as well as the average effect of innovative therapy ($X = 2$ versus $X = 0$) equal the differences of the respective adjusted means: $1.683 - 1.601 = 0.082$ and $1.875 - 1.601 = 0.274$.

To test the hypotheses of interest, EffectLiteR makes use of the F-statistic or the linear Wald test. Both will be explained below.

Hypothesis Testing

The F-test can be calculated as follows (Seber & Lee, 2012, p. 100):

$$F = \frac{n}{h}(R\widehat{\gamma})'(R\widehat{I}_1^{-1}R')^{-1}(R\widehat{\gamma}) \xrightarrow{D} F, \tag{10}$$

where n is the number of observations, h is the row rank of R, which is the constraint matrix specifying the linear combinations of regression coefficients expressing the hypothesis of interest, $\widehat{\gamma}$ is the vector of estimated regression coefficients and $\widehat{I}_1$ their unit information matrix. Under the null hypothesis, the F-statistic follows an F distribution with $df_1 = h, df_2 = n - p$, where p is the column rank of X, the design matrix of the regression model.

Using the same notation, the linear Wald test can be defined as (Fahrmeir et al., 2013, p. 663):

$$Wald = n(R\widehat{\gamma})'(R\widehat{I}_1^{-1}R')^{-1}(R\widehat{\gamma}) \xrightarrow{D} \chi^2. \tag{11}$$

Under the null hypothesis, the linear Wald test is χ^2 distributed with $df = h$. We can construct two versions of the linear Wald statistic, a regular and a corrected one (Seber & Lee, 2012, p. 100). The difference between them lies in the calculation of the unit information matrix. In the regular linear Wald statistic, the unit information matrix can be expressed as:

$$\widehat{I}_1 = \frac{1}{n} \cdot X'X \cdot \frac{1}{S^2}, \tag{12}$$

where $X'X$ is the cross-product of the design matrix and the mean squared error S^2 is calculated as follows:

$$S^2 = \frac{RSS(\widehat{\gamma})}{n}. \tag{13}$$

$RSS(\widehat{\gamma})$ is the residual sum of squares and can be obtained as:

$$RSS(\widehat{\gamma}) = \sum_{i=1}^{n} e_i^2, \tag{14}$$

where e_i are the deviations of the observed from the predicted data based on the regression model using the unrestricted estimates $\widehat{\gamma}$. To obtain the corrected linear Wald statistic, S^2 in Equation 12 is substituted by $S^2_{corrected}$, the degrees of freedom corrected mean squared error. It is defined as:

$$S^2_{corrected} = \frac{RSS(\widehat{\gamma})}{n-p}. \tag{15}$$

If we use $S^2_{corrected}$, $Wald/h$ will be identical to the F-statistic (Fahrmeir et al., 2013, p. 131). If we use S^2, $Wald/h$ and F values will be similar, but not identical in small samples. EffectLiteR uses $S^2_{corrected}$ to calculate the Wald statistic, but it is important to keep this subtle difference in mind when using other software programs, especially if the sample size is small.

To test H_1 against H_2 in the EffectLiteR framework, R has to be defined as follows. AE_{10} can be written as $r_1'\widehat{\gamma}$, where r_1' equals:

$$r_1' = (0\ 0\ 0\ 0\ 1\ E(Z)\ P(K=1)\ E(K \cdot Z)\ 0\ 0\ 0\ 0). \tag{16}$$

Similarly, AE_{20} can be written as $r_2'\widehat{\gamma}$, where r_2' equals:

$$r_2' = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ E(Z)\ P(K = 1)\ E(K \cdot Z)). \tag{17}$$

Combining r_1' and r_2' gives the full constraint matrix:

$$R = \begin{pmatrix} 0\ 0\ 0\ 0\ 1\ E(Z)\ P(K = 1)\ E(K \cdot Z)\ 0 & 0 & 0 & 0 \\ 0\ 0\ 0\ 0\ 0\ \ \ 0\ \ \ \ \ 0\ \ \ \ \ \ \ \ \ \ 0\ \ \ \ 1\ E(Z)\ P(K = 1)\ E(K \cdot Z) \end{pmatrix}. \tag{18}$$

This hypothesis of no average effects is routinely calculated by EffectLiteR and yields $F(2, 988) = 2.66$, $p = .071$. Thus, we cannot reject $H_1: AE_{20} = 0$, $AE_{10} = 0$ in favor of $H_2: AE_{20} \neq 0$, $AE_{10} \neq 0$, using an α-level of .05. Furthermore, the current EffectLiteR approach is not equipped to test H_1 against the fully ordered informative hypothesis H_4. In the following section, it will be shown how to extend the EffectLiteR approach to incorporate informative hypotheses.

Integrating Informative Hypotheses into the EffectLiteR Framework

Informative hypotheses reflect prior order expectations regarding means, regression coefficients or combinations thereof. These hypotheses can be constructed by means of different constraints (see, e.g., Hoijtink, 2012). The most important constraints are equality constraints like $\mu_1 = \mu_2$ and inequality constraints like $\mu_1 < \mu_2$ or $\mu_2 - \mu_1 > 0.5$. Both large-sample test statistics (generalized Wald, Score or likelihood ratio) and small-sample test statistics ($\bar{F}$, $\bar{E}$) have been developed (Barlow et al., 1972; Robertson et al., 1988; Silvapulle & Sen, 2005). The $\bar{F}$-test was first introduced by Kudô (1963) and generalized by Wolak (1987). The $\bar{F}$-test and the generalized linear Wald test will be explained in more detail below.

Fbar-Statistic

The $\bar{F}$-statistic can be defined as follows (Silvapulle & Sen, 2005, p. 29):

$$\bar{F} = n(R\widetilde{\gamma})'(R\tilde{I}_1^{-1}R')^{-1}(R\widetilde{\gamma}) \xrightarrow{D} \bar{F}. \tag{19}$$

Note that R in Equation 19 is the same R as in Equation 10. However, $\widehat{\gamma}$, the vector of estimated unrestricted regression coefficients in Equation 10, is now replaced by $\widetilde{\gamma}$, the vector of estimated restricted regression coefficients. $\tilde{I}_1$ is their unit information matrix, which can also be exchanged by $\widehat{I}_1$, the unit information matrix of the unrestricted fit. According to Silvapulle and Sen (2005, p. 29), including the constant $\frac{1}{h}$ from the regular F-statistic in the $\bar{F}$-statistic is not necessary, as it does not affect the results. Under the null hypothesis, the $\bar{F}$-statistic follows an $\bar{F}$ distribution, which is a weighted mixture of F distributions.

If the constraints are linear, $\widetilde{\gamma}$ can be found using quadratic programming, for example via the subroutine `solve.QP()` in the R package quadprog (Turlach & Weingessel, 2019). In a simple regression example with three coefficients, the unrestricted estimates $\widehat{\gamma}$ may be $(-0.1, 0.2, 0.5)$. However, if $H_a{:}\beta_3 > \beta_2 > \beta_1 > 0$, β_1 will be constrained to be greater than zero in our estimation procedure, leading to the restricted estimates $\widetilde{\gamma}$ that may look like $(0.001, 0.19, 0.48)$. Note that the estimates of β_2 and β_3 also change slightly, even though they satisfied the constraints right from the beginning. Furthermore, the restricted estimation will lead to different residuals that are used in the informative test statistic.

Generalized Linear Wald Test

The generalized linear Wald statistic is a generalization of the regular Wald test and can be found in Silvapulle and Sen (2005, p. 154):

$$Wald = n(R\widetilde{\gamma})'(R\tilde{I}_1^{-1}R')^{-1}(R\widetilde{\gamma}) \xrightarrow{D} \bar{\chi}^2. \tag{20}$$

Under the null hypothesis, the generalized linear Wald test is $\bar{\chi}^2$ distributed, which is a weighted mixture of χ^2 distributions. Again, we can construct two versions of the generalized linear Wald statistic, depending on whether S^2 or $S^2_{corrected}$ is used to calculate $\hat{I}_1$. Note that to obtain $RSS(\widetilde{\gamma})$ under the restricted hypothesis, e_i are the deviations of the observed from the predicted data based on the regression model using the restricted estimates $\widetilde{\gamma}$. If we use $S^2_{corrected}$, the generalized linear Wald statistic will be identical to the $\bar{F}$-statistic.

In our example, the corrected Wald test as well as the $\bar{F}$-test yield the same results, as expected: $\bar{\chi}^2 = 5.316$ and $\bar{F} = 5.316$. In contrast, the uncorrected Wald test produces the slightly different result $\bar{\chi}^2 = 5.381$. In the following, the calculation of p-values for the generalized linear Wald and $\bar{F}$-statistic is explained.

p-Values

To calculate the p-values of the $\bar{F}$- and generalized linear Wald test, two approaches can be followed, which should yield similar results. These procedures are described in detail in the document "E-pValues.pdf" that can be found in the Supplementary Materials. Essentially, the first approach calculates the p-value by means of simulation (Silvapulle & Sen, 2005, p. 98). By generating a large amount of normal (or non-normal) random data under the null hypothesis and calculating their test statistics, it is possible to determine the proportion of times, the newly obtained test statistic exceeds the originally observed test statistic as an estimate of the p-value.

The second approach is more economical, as the mixing weights are estimated first, also by means of simulation, and directly used to calculate the p-value afterwards

(Silvapulle & Sen, 2005, p. 79). Here, generating a large amount of (normal) random data and calculating the parameters of interest allows to determine the proportions of times that a different number of constraints are satisfied as an estimate of the weights.

Estimating the p-value for the obtained $\bar{F}$-value yields $p = .019$ (simulation approach) and $p = .020$ (weights approach). This allows us to reject hypothesis $H_1 \colon AE_{20} = 0, AE_{10} = 0$ in favor of hypothesis $H_4 \colon AE_{20} > AE_{10} > 0$. A complete ordering of average effects of the treatment groups therefore seems to be in accordance with our data. This is an illustration of how informative hypothesis testing has greater power compared to classical null hypothesis testing, since the former could not reject $H_1 \colon AE_{20} = 0, AE_{10} = 0$ in favor of $H_2 \colon AE_{20} \neq 0, AE_{10} \neq 0$.

Intermediate Summary

It can be concluded that integrating informative hypotheses into the EffectLiteR framework enriches testing hypotheses regarding effects of interest. This is because it allows to directly test hypotheses that correspond to the researcher's prior order expectations of the treatment groups. Using this approach, it is not necessary anymore to follow a cumbersome two-step procedure with potentially increased type I error rates, as is needed in classical null hypothesis testing. In fact, it is even possible to detect significant results that would not have been detected via classical null hypothesis testing, as was the case in our motivating example. Going one step further, the next section illustrates how to account for stochastic group sizes while testing informative hypotheses in the EffectLiteR framework.

Stochastic Group Sizes

The treatment groups ($X = x$) and the binary covariate ($K = k$) divide the sample in several groups by means of combinations of their different levels. The group size is the number of observations in a particular group. The group proportions are the ratios of group sizes divided by the total sample size. The corresponding model parameters are called group weights.

So far, we have treated the group sizes as fixed. Theoretically, this implies that each time we would replicate the study, the group sizes would remain the same. However, in practice, this is not always the case, and the group sizes may vary from sample to sample. In this case, it is more correct to treat the group sizes as stochastic, and this implies that the group weights become free parameters in our model. Mayer and Thoemmes (2019) showed that failing to take the stochastic nature of group sizes into account may lead to increased type I error rates even in randomized experiments. If we want to account for stochastic group weights, our vector of parameters will not only include the regression coefficients but also the sample proportions:

$$\theta \ = (\gamma', p'). \tag{21}$$

In our simulated data example, $\widehat{\theta}$ contains:

$$\begin{aligned}
\widehat{\gamma}' &= (1.667 \quad -0.285 \quad 0.029 \quad 0.300 \quad -0.252 \quad 0.591 \quad 0.261 \quad -0.546 \quad 0.128 \\
&\qquad 0.322 \quad 0.032 \quad -0.234). \\
\widehat{p}' &= (0.104 \quad 0.206 \quad 0.074 \quad 0.097 \quad 0.291 \quad 0.228).
\end{aligned} \tag{22}$$

Note that since Kiefer and Mayer (2019) proved that γ and p are independent, the variance-covariance matrix is block-diagonal:

$$Var(\theta) \ = \begin{pmatrix} Var(\gamma) & 0 \\ 0 & Var(p) \end{pmatrix}. \tag{23}$$

To apply informative hypothesis testing in the EffectLiteR framework while taking into account stochastic group sizes, two changes are needed. First, a new method is needed to obtain $\tilde{\theta}$, the vector of restricted regression coefficients and sample proportions, and second, the generalized non-linear Wald statistic has to be used. This is because the effects of interest are no longer a linear combination of the model parameters and quadratic programming cannot be used to find $\tilde{\theta}$.

Estimation of $\tilde{\theta}$ by Means of SEM

To obtain $\tilde{\theta}$, several steps should be followed. The model is specified as a SEM, which is estimated via the R package elavaan (Rosseel, 2012). SEM software is used because it can handle constraints on functions of parameters. The hypotheses are included in the model definition as quantities of interest, but the constraints imposed on them are defined separately. This allows us to first fit an unconstrained SEM. If all quantities of interest already satisfy the constraints, the generalized non-linear Wald test is calculated based on the results of the unconstrained fit. If at least one quantity of interest does not satisfy the constraints, that is, at least one constraint is active, a constrained SEM is fitted and the generalized non-linear Wald test is calculated based on the results of the constrained fit. This procedure is implemented in the R script "A-exampleI.R" and is explained in the document "C-explanation.pdf", which are both part of the Supplementary Materials. The resulting parameter vector $\tilde{\theta}$ is used in the generalized non-linear Wald test, which is explained in the following section.

Generalized Non-linear Wald Test

The generalized non-linear Wald statistic can be found in Silvapulle and Sen (2005, p. 166) and is defined as follows:

$$Wald \ = nc(\widetilde{\theta})'[C(\widetilde{\theta})\widetilde{I}_1^{-1}C(\widetilde{\theta})']^{-1}c(\widetilde{\theta}) \xrightarrow{D} \overline{\chi}^2, \tag{24}$$

where c is a function of $\widetilde{\theta}$ that returns a vector, C is the Jacobian matrix of c and $\widetilde{I}_1$ the unit information matrix. Note that if $c(\widetilde{\theta})$ was linear, it would be equal to R in the generalized linear Wald statistic in Equation 20.

In our simulated data example, we obtain the following results. The columns of Table 3 titled "Stochastic group sizes" show the estimates of all adjusted means and all average effects, including their standard errors. It can be observed that the standard errors get slightly larger when stochastic group sizes are considered, compared to when group sizes are assumed to be fixed, which reflects the increased uncertainty. However, when the sample size is smaller, the differences between the two sets of standard errors become larger. This is illustrated in the R script "F-exampleISmallN.R" in the Supplementary Materials. The data set with $N = 50$ is available under the name "G-exampleISmallNData.rds" and the results are presented in the document "H-exampleISmallNResults.pdf". As for the Wald statistic in our running example, we obtain $\overline{\chi}^2 = 5.306$, $p = .019/.020$, depending on whether the simulation or weights approach is used to calculate the p-value. The test statistic is slightly different than the one we obtained without considering stochastic group sizes ($\overline{\chi}^2 = 5.381$), while the p-values are nearly equal.

Intermediate Summary

We showed how to account for stochastic group sizes when testing informative hypotheses within the EffectLiteR framework. If group weights are considered as free parameters in the model, uncertainty increases, which manifests itself in increased standard errors of parameters estimates. This seems to be more evident in small samples than it is in large samples. This observation is important especially in the social and behavioral sciences, where treatment group sizes often vary across samples and the exact group sizes are not determined before conducting the experiment. However, at this point, considering stochastic group sizes comes with an increased computational cost, as parameter estimation has to be carried out in a two-step procedure by means of lavaan.

Data Example

In this section, the illustrated methods will be applied to an artificial data example from the constrained statistical inference literature (Hoijtink, 2012; Vanbrabant, 2018). The accompanying R script can be found in the Supplementary Materials and is named "I-exampleII.R". We recommend this R script as a starting point for all researchers who would like to apply our method. The Anger Management data set comes with the restriktor package (Vanbrabant, 2018, 2020). It contains 40 observations on decrease in aggression level over the course of eight weeks and considers the continuous covariate

age. Subjects are assigned to either one of the following groups: No training ($X = 0$), physical training ($X = 1$), behavioral therapy ($X = 2$) and a combination of physical exercise and behavioral therapy ($X = 3$). All groups have the same size of $n = 10$. Since this data set is artificial and group sizes are fixed to 10, we do not consider stochastic group sizes here. Table 4 shows the raw and adjusted means of decrease in aggression level Y as well as the relative treatment group frequencies.

Table 4

Raw as Well as Adjusted Means of Decrease in Aggression Level $E(Y)$, $AdjM(Y)$ and Relative Treatment Group Frequencies $P(X = x)$

Group	Raw/adjusted Means and relative treatment group frequency
$X = 0$	$E(Y \mid X = 0) = -0.200$
	$AdjM(Y \mid X = 0) = -0.920$
	$P(X = 0) = 0.250$
$X = 1$	$E(Y \mid X = 1) = 0.800$
	$AdjM(Y \mid X = 1) = 0.922$
	$P(X = 1) = 0.250$
$X = 2$	$E(Y \mid X = 2) = 3.100$
	$AdjM(Y \mid X = 2) = 3.341$
	$P(X = 2) = 0.250$
$X = 3$	$E(Y \mid X = 3) = 4.100$
	$AdjM(Y \mid X = 3) = 4.237$
	$P(X = 3) = 0.250$

Possible hypotheses regarding the effects of interest are:

$$
\begin{aligned}
&H_1 : AE_{30} = 0, \; AE_{20} = 0, \; AE_{10} = 0, \\
&H_2 : AE_{30} \neq 0, \; AE_{20} \neq 0, \; AE_{10} \neq 0, \\
&H_3 : AE_{30} > 0, \; AE_{20} > 0, \; AE_{10} > 0, \\
&H_4 : AE_{30} > AE_{20} > AE_{10} > 0,
\end{aligned}
\tag{25}
$$

where AE again refers to an average effect. H_1 assumes that all average effects equal zero, that is, the average effect of the combination of physical exercise and behavioral therapy (AE_{30}), the average effect of behavioral therapy (AE_{20}) and the average effect of physical training (AE_{10}) are assumed to be zero. H_2 states that these average effects are not equal to zero. According to H_3, the average effects are greater than zero. Finally, H_4 postulates a complete ordering with the average effect of physical exercise and behavioral therapy (AE_{30}) being greater than the average effect of behavioral therapy

(AE_{20}), the average effect of behavioral therapy (AE_{20}) being greater than the average effect of physical training (AE_{10}) and the average effect of physical training (AE_{10}) being greater than zero.

The adjusted means and average effects estimates are presented in Table 5.

Table 5

Adjusted Means and Average Effects Estimates With Standard Errors in Parentheses (Anger Management Data)

Group	Adjusted mean	Average effect
No training ($X = 0$)	-0.920 (0.777)	
Physical training ($X = 1$)	0.922 (0.744)	1.842 (1.076)
Behavioral therapy ($X = 2$)	3.341 (0.706)	4.261 (1.049)
Physical training & behavioral therapy ($X = 3$)	4.237 (0.801)	5.157 (1.116)

Testing Hypothesis H_1 against H_4, $\bar{F} = 27.59$, $p < .001$ (using either one of the two approaches to calculate the p-value) is obtained. Furthermore, the regular F-test of H_1 against H_2 yields $F = 9.20$, $p < .001$. Again, we observe that the test statistic of the informative hypothesis test is larger than the test statistic of the classical null hypothesis test, even though both p-values are highly significant.

Discussion

This paper demonstrated how to integrate informative hypotheses into the EffectLiteR framework, while taking into account stochastic group sizes. We provided R scripts and explanatory documents in the Supplementary Materials. The method described allows to directly test hypotheses that represent the researcher's prior order expectations about the effects of interest in the treatment groups. In contrast to classical null hypothesis testing, it is not necessary anymore to follow a cumbersome two-step procedure with potentially increased type I error rates due to multiple testing. Furthermore, we were able to show that testing informative hypotheses about effects of interest has greater power compared to classical null hypothesis testing. Vanbrabant (2018) already showed this in the simpler context of comparing group means instead of effects of interest.

Considering stochastic group sizes in the approach described in this paper does not seem to impact the results substantially if the sample size is large. However, when dealing with small sample sizes, standard errors might be underestimated when erroneously treating stochastic group sizes as fixed. This is because of adequately accounting for the increased uncertainty that stems from considering group weights as free instead of as fixed parameters in the model.

Two approaches to estimate the *p*-values when testing informative hypotheses about effects of interest were introduced. When the residuals are normally distributed, the weights approach is preferred due to its efficiency. Furthermore, when stochastic group sizes are involved, we also prefer the weights approach, albeit for practical reasons: In particular when some of the group sizes are small, many simulation iterations may fail and must be replaced, making the simulation approach very time-consuming.

The limitations of this paper and the outlook on future research are the following. First, we only considered manifest variables. In the future, the presented methods should be extended to be able to deal with latent variables. Furthermore, the small-sample properties of the generalized non-linear Wald test are unknown. Thus, testing informative hypotheses in the EffectLiteR framework while accounting for stochastic group sizes should be examined further in future research by means of simulation studies. Moreover, we did not examine the consequences when assumptions of the general linear model like homoscedasticity are not fulfilled. When considering stochastic group sizes, it has already been shown that in the two-group context, erroneously assuming equal variances is only critical regarding standard error estimation of effects when group sizes are unequal (see, e.g., Berry, 1993; Mayer & Thoemmes, 2019). This can lead to increased type I error rates. Future research should thus examine the effect of violations of model assumptions when testing informative hypotheses in the EffectLiteR framework taking into account stochastic group sizes.

Finally, the two-step procedure presented to test informative hypotheses about effects of interest while considering stochastic group sizes needs a lot of manual tuning at this point. To make this approach more efficient and accessible for applied researchers, it might be possible to adapt it using the distance statistic (Silvapulle & Sen, 2005, p. 154). This would avoid the need for estimating the (order-)constrained model using SEM, and may result in a more efficient testing procedure.

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

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: The data for this article are freely available (see the Supplementary Materials section).

Supplementary Materials

For this article, the following Supplementary Materials are available via PsychArchives repository (for access see Index of Supplementary Materials below):

PsychOpen GOLD

- R scripts.
- R data sets.
- Explanatory documents.

Index of Supplementary Materials

Keck, C., Mayer, A., & Rosseel, Y. (2021). *Supplementary materials to: Integrating informative hypotheses into the EffectLiteR framework* [Scripts, data sets, and additional information]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5299

References

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

Berry, W. D. (1993). *Understanding regression assumptions.* Sage.

Casella, G., & Berger, R. L. (2002). *Statistical inference.* Duxbury Press.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications.* Springer.

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

Grubbs, K. M., Cheney, A. M., Fortney, J. C., Edlund, C., Han, X., Dubbert, P., Sherbourne, C. D., Craske, M. G., Stein, M. B., Roy-Byrne, P. P., & Sullivan, J. G. (2015). The role of gender in moderating treatment outcome in collaborative care for anxiety. *Psychiatric Services, 66*(3), 265-271. https://doi.org/10.1176/appi.ps.201400049

Gu, X., Hoijtink, H., Mulder, J., Van Lissa, C. J., Van Zundert, C., Jones, J., & Waller, N. (2020). Bain: Bayes factors for informative hypotheses (R package version 0.2.4) [Computer software manual]. https://CRAN.R-project.org/package=bain

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

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

Kiefer, C., & Mayer, A. (2019). Average effects based on regressions with a logarithmic link function: A new approach with stochastic covariates. *Psychometrika, 84*(2), 422-446. https://doi.org/10.1007/s11336-018-09654-1

Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika, 50*(3/4), 403-418. https://doi.org/10.2307/2333909

Mayer, A., & Dietzfelbinger, L. (2019). EffectLiteR: Average and conditional effects (R package version 0.4-4) [Computer software manual]. 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

Meyer, E., & Van Klaveren, C. (2013). The effectiveness of extended day programs: Evidence from a randomized field experiment in the Netherlands. *Economics of Education Review, 36*, 1-11. https://doi.org/10.1016/j.econedurev.2013.04.002

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. https://www.R-project.org/

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

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://www.jstatsoft.org/v48/i02/

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

Seber, G. A. F., & Lee, A. J. (2012). *Linear regression analysis.* Wiley.

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

Steyer, R. (2021). *Introduction to probability and conditional expectation* (unpublished). https://doi.org/10.13140/RG.2.2.18977.33125

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

Turlach, B.A., & Weingessel, A. (2019). Quadprog: Functions to solve quadratic programming problems (R package version 1.5-8) [Computer software manual]. https://CRAN.R-project.org/package=quadprog

Vanbrabant, L. (2018). *Reduction in sample size by order restrictions* [Doctoral dissertation, Ghent University, Belgium].

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

Weiss, D. J. (2006). *Analysis of variance and functional measurement: A practical guide.* Oxford University Press.

Wolak, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association, 82*(399), 782-793. https://doi.org/10.1080/01621459.1987.10478499