JCP JOURNAL OF CONSUMER PSYCHOLOGY

SCP SOCIETY FOR CONSUMER PSYCHOLOGY

BANOVA: Bayesian Analysis of Experiments in Consumer Psychology

Michel Wedel
Robert H. Smith School of Business, University of Maryland

Chen Dong
AirBnB, Inc.

Accepted by Anirban Mukhopadhyay; Associate Editor, Joel Huber

This article introduces a Bayesian extension of ANOVA for the analysis of experimental data in consumer psychology. The approach, called BANOVA (Bayesian ANOVA), addresses some common challenges that consumer psychologists encounter in their experimental work, and is specifically suited for the analysis of repeated measures designs. There appears to be a recent surge in interest in those designs based on the recognition that they are sensitive to individual differences in response to experimental treatments and that they offer advantages for assessing causal mediating mechanisms, even at the individual level. BANOVA enables the analysis of repeated measures data derived from mixed within–between-subjects experiments with Normal and nonNormal-dependent variables and accommodates unobserved individual differences. It allows for the calculation of effect sizes, planned comparisons, simple effects, spotlight and floodlight analyses, and includes a wide range of mediation, moderation, and moderated mediation analyses. An R software package implements these analyses, and aims to provide a one-stop shop for the analysis of experiments in consumer psychology. The package is illustrated through applications to a number of data sets from previously published studies.

Keywords effect size; floodlight analysis; hierarchical generalized linear model; MCMC; mediation; moderation; R package; repeated measures design; simple effects

INTRODUCTION

This article introduces a Bayesian extension of ANOVA that provides an integral solution for the analysis of experimental data in consumer psychology. The approach, called BANOVA (Bayesian ANOVA; Dong & Wedel, 2017), addresses some common challenges that consumer psychologists encounter in their experimental work. Some of these are that traditional null hypothesis significance testing has led to misleading interpretations and misuse of *p*-values (Nuzzo, 2014; Wasserstein & Lazar, 2016), that dependent variables may not be Normally distributed and other distributions are preferable given the nature of the data (Micceri, 1989), that the data may exhibit outliers (Barnett & Lewis, 1994) or missing values (Rubin, 1987), that the observations are not independent but may follow multilevel data structures (Gelman & Hill, 2007), and that sample sizes are sometimes relatively small. Although existing techniques are available that address these challenges on a case-by-case basis, BANOVA addresses them in an integrated fashion. An accompanying R-software package provides a one-stop shop to implement these analyses, and produces outputs that are tailored to the needs of researchers in consumer psychology that are consistent across very different types of analyses, and can be interpreted in the same way and with relatively little effort regardless of the model specification. BANOVA thus eliminates the need to rely on multiple software packages, macros, websites, and so on, for the analysis of a single experiment. This offers not only an advantage of convenience, but it also avoids errors in stacking several analyses on top of each other. Floodlight, planned comparisons, simple effects, mediation, and moderated mediation analyses are accommodated for a wide range of statistical distributions of the measurements and experimental designs. BANOVA is especially useful for the analysis of repeated measures designs, but can be applied to between-subject designs as well.

Repeated Measures Designs

In consumer psychology repeated measures designs seem to be often been overlooked as

Received 4 January 2018; accepted 8 May 2019
Available online xx xx xxxx
Correspondence concerning this article should be addressed to Michel Wedel, Robert H. Smith School of Business, University of Maryland, 7621 Mowatt Ln, College Park, MD 20742, USA. Electronic mail may be sent to: mwedel@rhsmith.umd.edu

 1057-7408/2019/1532-7663
DOI: 10.1002/jcpy.1111

appropriate and preferable to more popular between-subjects designs. Yet, there appears to be a renewed interest in the application of repeated measures designs based on the recognition of some of their advantages in laboratory and field experiments (Morgan & Morgan, 2001; Smith, 2012). Single-case designs have been applied to assess the effectiveness of experimental treatments by repeatedly measuring the behavior of a participant over the course of an intervention (Kazdin, 2010; Smith, 2012). Despite their name, multiple participants may be included, in which case the designs are referred to as within-subject designs (see for examples, Graf, Mayer, & Landwehr, 2017; Sweldens, Van Osselaer, & Janiszewski, 2010). In these designs each participant serves as her own control, and multiple participants replicate the effect of the treatment (Horner & Spaulding, 2010). When some treatments are administered between-subjects this results in mixed designs.

We will broadly refer to all these types of designs as "repeated measures designs." Some of the foremost appeals of these designs are that they are sensitive to individual differences in response to experimental treatments, and these are, therefore, more effectively controlled for. In addition, repeated measures designs have greater power because treatment effects are tested against generally smaller within-subject variability, while the repeated measures afford additional data. These designs thus require smaller samples of participants (Cooper, Heron, & Heward, 2007). Furthermore, repeated measures designs are also advantageous for investigating the effect of mediating variables that are hypothesized to transmit the effect of the treatment onto the dependent variable. First, because they are cross-sectional, between-subject designs only accommodate mediators that have contemporaneous effects, while in addition they are often only measured after the dependent variable. On the contrary, repeated measures designs are preferable for assessing causal mechanisms when they allow the measurement of mediators to temporally precede that of the dependent variables (MacKinnon, Fairchild, & Fritz, 2007). Second, such longitudinal repeated measures designs are critical when the underlying mediating mechanism takes time to reveal itself and varies in strength as time passes (Selig & Preacher, 2009). Third, while between-subjects designs assume that the causal effect has the same strength for all participants, repeated measures designs allow one to assess whether experimental factors and mediators have effects that vary in magnitude across participants (Bullock, Green, & Ha, 2010; Hutchinson, Kamakura, & Lynch, 2000). Fourth, mechanisms that mediate treatment effects within-subjects may be substantively different from those that mediate the effect between-subjects (MacKinnon et al., 2007), and repeated measures designs allow one to disentangle these mechanisms.

Of course, downsides of repeated measures designs should not be overlooked, which primarily result from order (halo, demand, fatigue, learning) effects caused by exposing the participants to multiple treatments or stimuli in close temporal proximity. Order effects need to be mitigated by randomization, counterbalancing, or controls in the analysis, and may necessitate the development of additional stimulus material (Howitt & Cramer, 2011, p.179–181).

BANOVA

More advanced methods than ANOVA are needed to analyze repeated measures designs, which might be a barrier toward their application in consumer psychology. Challenges include the nonNormal distribution and missing values of the dependent variables (counts, binary variables, rating scales), smaller numbers of participants, and the multilevel structure of the data (Rindskopf, 2014; Smith, 2012). Clearly, these challenges make it inappropriate to analyze the data from such experiments with the standard ANOVA or regression analyses, and follow-up floodlight and mediation analyses (Baron & Kenny, 1986; Gelman & Hill, 2007; Krull & MacKinnon, 2001; Pieters, 2017; Spiller, Fitzsimons, Lynch, & McClelland, 2013). This article, therefore, proposes a general extension of ANOVA for repeated measures designs that provides a unified solution, which includes floodlight, simple effects, mediation, and moderated mediation analyses, for a wide range of experimental designs and statistical distributions of the measurements. It can be applied to simple between-subject designs as well, for which it will provide similar results as standard ANOVA if the measurements follow a Normal distribution. While many of the challenges can be addressed in the classical statistical framework, the BANOVA approach takes a Bayesian perspective on statistical inference, which offers several conceptual advantages (Kruschke, 2013). The R-software package developed by Dong and Wedel (2017) is extended in this article to implement the analyses, and produces outputs that are tailored to the needs of researchers in consumer psychology that can be interpreted by them with little effort.

We begin with a brief outline of some of the benefits of adopting the Bayesian framework for statistical inference and hypothesis testing in consumer research (Kruschke, 2013; Wagenmakers, 2007). We then provide a high-level overview of the BANOVA methodology for the analysis of repeated measures data, and describe how effect sizes, planned comparisons, simple effects, and floodlight analyses are calculated. Next, a Bayesian approach to mediation, moderation, and moderated mediation is discussed. We illustrate the approaches with several applications to previously published studies (the data sets are provided with the R package, all R commands are provided in the Methodological Details Appendix).

Bayesian Analysis of Experimental Data

The Bayesian statistical framework specifies how a researcher should learn about model parameters and hypotheses from data. It formalizes the researcher's uncertainty about model parameters via probability distributions (Gill, 2015, p.1–8). Specifically, before any data are collected, the prior uncertainty of a researcher about a parameter of a model, say a coefficient β in a regression model, is quantified via the prior probability distribution. It is briefly denoted here as $p(\beta)$. For example, if the researcher has little or no information about the value of the regression coefficient, the prior distribution would be uninformative (or diffuse) to reflect that all values of the coefficient are a priori (almost) equally likely. For a regression coefficient, the prior could be $\beta \sim \text{Normal}(0, s)$, a Normal distribution with a very large variance (say $s = 10{,}000$). In most cases this prior will have a negligible influence on inferences about the parameter, and throughout this article we will use such diffuse priors. Nevertheless, in some cases it may make sense to choose informative priors, for example if information is available from previous studies in the literature, or from earlier studies in a sequence of experiments (Gill, 2015, p. 53). Choosing informative priors may also increase statistical power in mediation analysis (Miočević, MacKinnon, & Levy, 2017). Nevertheless, although the choice of an informative prior should receive close scrutiny and be subjected to sensitivity analyses, we refrain from using them here, because informative priors have been criticized for their subjectivity from the point of view of classical statistics (Efron, 1978).

In classical statistics, the data are assumed to be a random sample from a population. The model parameter β is assumed to be fixed in the population by nature, and uncertainty about its value arises because rather than the entire population only a (random) sample of participants is included in the experiment. On the contrary, Bayesian statistics is based on the fundamentally different premise that the data are fixed by the experiment and that the uncertainty about the parameter β is characterized through a probability distribution. Once the data are collected in an experiment, Bayes' rule prescribes how the prior uncertainty about the parameter, $p(\beta)$, should be updated with the information in the data. The resulting posterior probability distribution, briefly denoted as $p(\beta|\text{data})$, encapsulates the uncertainty about the parameter value after the data are observed. This distribution is central in the Bayesian approach because it is the basis for inferences on hypotheses about the parameter. These inferences are predicated only upon the data that were collected in the experiment; unlike what is the case in classical statistics hypothetical repeated sampling of the data need not be invoked (Gill, 2015; p.57–61).

Although it would be wrong to conclude that Bayesian analysis will free the researcher from all problems surrounding the misuse of *p*-values (*p*-hacking) and the lack of replicability of experiments, it has several important properties that make it attractive for applications in consumer psychology:

1. Bayesian inferences are accurate even for small sample sizes (Gill, 2015; p.57–61). This is particularly advantageous for single-case designs, repeated measures designs with relatively few participants, or for making inferences on individual participants.
2. It has been proven that one may analyze the data before deciding whether or not to collect more data (Edwards, Lindeman, & Savage, 1963; Wagenmakers, 2007). In the Bayesian framework, this is equivalent to increasing the sample size. However, it is assumed that the experiment will eventually be published even if a null-result is obtained after more data are obtained, while it is good practice to report this "data-peeking" procedure.
3. The Credible Interval of a parameter is obtained from its posterior distribution, and it is the interval that contains the parameter with a certain (usually 95%) probability. This interpretation of the Credible Interval is intuitive and (unlike the classical confidence interval) does not invoke the concept of repeated sampling (Gill, 2015; p.57–61).

4. The probability that a certain hypothesis on the parameter value holds can be directly calculated from the posterior distribution. For example, one can calculate the probability of the hypotheses H_0:$\beta \leq 0$, which allows one to not only reject the null hypothesis (as is strictly the case in classical statistics), but also quantify the support for it (Wagenmakers, 2007).
5. This probability, $p(\beta \leq 0)$, is called the Bayesian p-value (Carlin & Louis, 1998; p.53). It has been shown that if the prior is symmetric around zero (for example the Normal distribution for β above) then the Bayesian p-value, unlike its classical counterpart, can be directly interpreted as a measure of the strength of evidence for the null hypothesis (Marsman & Wagenmakers, 2017).
6. The Bayesian approach allows one to calculate posterior distributions not only of parameters, but also of quantities derived from them. This is very useful when one wants to obtain the posterior distribution of contrasts defining a planned comparison, of indirect effects in mediation models (Zhang, Wedel, & Pieters, 2009), of effect sizes, or Johnson–Neyman points in floodlight analyses (Spiller et al., 2013). This will be shown later.
7. In data arising from repeated measures designs, one needs to allow for unobserved individual differences in the effect of the treatment, or of a mediator. Failure to do so may lead to biased inferences (Hutchinson et al., 2000). An analysis of such data with Hierarchical (Bayes) models naturally allows for this, which offers the benefit that each participant's estimate of the parameter borrows information from all other participants' data, and which is called "shrinkage." Because there is more shrinkage if participants are more similar or if there is little data on a participant (Gelman et al., 2013), this greatly helps the parameter estimates.

For many models, including BANOVA models for repeated measures designs, a challenge in Bayesian analyses is that the posterior distribution of the parameters, $p(\beta|\text{data})$, does not have a known form. Fortunately, the so called Markov Chain Monte Carlo (MCMC) algorithms allow one to iteratively draw samples from the posterior distribution to obtain an accurate approximation (the larger the number of draws, the better the approximation). The technical details are described in, for example, Gelman et al. (2013) and Gill (2015) and need not be of concern to the applied researcher, although it may take a little more computing time to produce results. An important point to keep in mind is that these iterative MCMC algorithms need to converge, and this needs to be checked in each application. If it has been established that the algorithm has converged after a certain number of draws, the draws before that point (called burn-in) are discarded. The remaining draws of the parameter values are used to calculate summaries of the posterior distribution: for example the average of the draws is an estimate of the parameter, their standard deviation is a measure of uncertainty about the parameter, and the fraction of draws that has a negative value is an estimate of the probability of H_0:$\beta \leq 0$.

While the implementation of MCMC estimation previously required a significant investment in programming, that burden has been eased by freely available software packages such as BUGS (Lunn, Best, & Spiegelhalter, 2000) and STAN (Carpenter et al., 2017). Nevertheless, for many researchers in consumer psychology the effort of writing programs in those languages to analyze their data on a case-by-case basis may still pose too much of a barrier or burden. The BANOVA R-package provides an interface with the STAN software, with simple syntax that makes it easy to specify models and do a range of follow-up analyses. For more entrepreneurial researchers, BANOVA can produce the STAN code as output for them to modify for more advanced analysis of their data.

BANOVA Tutorial

BANOVA is developed for the analysis of data collected in repeated measures experiments with a possibly nonNormal dependent variable, in which participants are exposed to between-subject and/or within-subject manipulations, and in which covariates and mediating variables may have been measured as well. We first list the most important measurement scales (continuous and categorical) for dependent variables measured in those experiments and the probability distributions that are associated with them. The choice of a probability distribution is most often uniquely determined by the properties of the measurements: for example, count data call for the use of a Poisson distribution and binary data for the use of a binomial distribution. The tight link between the properties of

measurements and statistical distributions greatly reduces ambiguity in selecting an appropriate statistical model for the analysis of experimental data in consumer psychology, and failure to recognize the correct distributional form may result in biases in the inferences from the experiment and errors in significance testing.

Statistical Distributions

Continuous responses

To analyze continuous data that are symmetrical, which arise for example when measuring participants' perceptions of processing fluency on a visual analog scale with a resolution of 100 increments, one uses the Normal distribution. To analyze data that are skewed to the right, for example, when response times are measured to study memory, one could use the log-Normal distribution (which is equivalent to taking the natural logarithm of the dependent variable and using a Normal distribution). To analyze data with outliers or fatter tails than the Normal distribution, one may use a t-distribution. Because the t-distribution has "fatter tails" than the Normal (the smaller the degrees of freedom, the fatter the tails), it may be used for robust inference, which is less sensitive to extreme observations (Bernardo & Giron, 1992). This may eliminate the need to subjectively remove outliers from the data.

Categorical variables

To model count data that can take on values 0, 1, 2, 3, etc. which may occur, for example, when one counts eye fixations on an area of interest in eye tracking studies, or jelly beans in studies of variety seeking, the Poisson distribution is used. To model data that take on the values 0 or 1, for example whether or not a participant recognizes an ad in an implicit memory task, a Bernoulli distribution is used. Alternatively, if the data represent the number of successes in K trials the outcome of which can be zero or one, for example, how many of $K = 20$ advertisements a participant correctly identifies, a Binomial distribution is used. For 0/1 data with more than two categories, for example when each participant chooses one of four brands, the Multinomial distribution is used. Categorical measurements that take on ordered values 1, 2, 3, . . ., arise, for example, when attitudes are measured on a 5-point rating scale. The ordered Multinomial distribution would be used in this case. BANOVA accommodates all of these distributions.

Missing values

Some values of the outcome or mediator variables measured in an experiment may be missing. For example, in eye movement data the eye tracker may fail to record the point of regard because of blinks or large head movements. Removing outliers creates a missing data record. The large literature on missing data imputation has shown that multiple imputation procedures, which involve repeatedly filling in each missing data point to reflect the uncertainty in its true value, are preferable to simply ignore the missing observations or replacing them with the sample average of that variable (Rubin, 1987). Bayesian methods allow one to repeatedly impute missing data at the same time the model is estimated. Enders, Fairchild, and MacKinnon (2013) demonstrate the benefits of such procedures in mediation analysis. In BANOVA, missing values of the dependent variable are imputed automatically along with the estimation of the model, using all information available in the data.

Within- And Between-Subjects Models

BANOVA models are so called Hierarchical Generalized Linear Models (Lee & Nelder, 1996; McCullagh & Nelder, 1989), which extend linear regression (and thus ANOVA) to situations where measurements of a dependent variable may follow, amongst others, a Normal, Poisson, Bernoulli, Binomial, Multinomial, or rank-Multinomial distribution. While these models do not need to be estimated in a Bayesian framework, MCMC estimation procedures make that particularly convenient. In these models, the expected values of the measurements of the dependent variable are linked to two submodels: a within-subjects and a between-subjects model (Dong & Wedel, 2017; MacKinnon, 2013, p.237; default canonical link functions are used, see for details McCullagh & Nelder, 1989). Many well-known models arise as special cases. For example, in the case of a Bernoulli-dependent variable, the BANOVA model is a (hierarchical or mixed) logistic regression model. We provide details of the within- and between-subject models next.

At the *within-subjects level*, there are (categorical) factors denoted as x_1, x_2, x_3, . . ., and (continuous) covariates and/or mediators denoted as z_1, z_2, z_3,

. . ., which vary within and across participants. We use lower case symbols for them. It is important to note that in BANOVA all continuous covariates z_1, z_2, z_3, . . ., are automatically mean-centered, and that effect-coding is automatically used for all factors, creating K-1 variables for a factor with K levels.

As an example, assume one 2-level within-subjects factor, say product category (bath towels and paper towels), effect-coded as x_1, and a continuous (mean-centered) within-subjects covariate z_1, say perceived quality, which may have a possible interactive effect on the dependent variable, that is, willingness to pay (Morales, 2005; study 1). We use the symbol θ for the parameters of the within-subjects model and omit subscripts that indicate participants and repeated measures for convenience. The within-subjects model is then:

$$y \sim \theta_0 + \theta_1 x_1 + \theta_2 z_1 + \theta_3 x_1 z_1 \qquad (1)$$

The "~" links the dependent variable, y, to the explanatory variables in the within-subjects model (more precisely, the ~ links a function of the expectation of y to the explanatory variables, for example the natural logarithm of the expectation in the case of a Poisson regression model; see for details McCullagh & Nelder, 1989). The within-subjects model applies to each individual participant separately, so that the values of the parameters θ_0, θ_1, θ_2, and θ_3 may be different for each participant. Model 1, therefore, accounts for unobserved differences between participants in the effect of the within-subjects variables, product category, and quality.

At the *between-subjects level* there are (categorical) factors denoted as X_1, X_2, X_3, . . ., and continuous covariates and/or mediators denoted as Z_1, Z_2, Z_3, . . ., which vary across participants only. We label all these between-subjects variables with upper case symbols and use β to indicate the corresponding parameters. It is important to note that all continuous covariates Z_1, Z_2, Z_3, . . ., are automatically mean-centered.

To continue the example, study 1 of Morales (2005) investigated whether people are willing to pay more for products that are sold by firms that exert more effort in marketing them, and if this effect disappears when they perceive the firm's motive as one of persuasion. Morales investigated this across the categories (x_1) and controlling for product quality (z_1). Effort (X_1; low vs. high) and motive (X_2; neutral vs. persuasion) are both manipulated at two levels between-subjects and effect-coded, with an interactive effect denoted by X_1X_2. In a hierarchical model, these two between-subjects variables may affect (moderate) each of the coefficients of the within-subject model in Equation 1: θ_0, θ_1, θ_2, and θ_3. In the example there are thus four between-subjects models:

$$\begin{aligned}
\theta_0 &= \beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \beta_{03}X_1X_2 + e_0 \\
\theta_1 &= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_1X_2 + e_1 \\
\theta_2 &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_1X_2 + e_2 \\
\theta_3 &= \beta_{30} + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_1X_2 + e_3
\end{aligned} \qquad (2)$$

Here, e_0, e_1, e_2, and e_3 are error terms that are assumed to follow a Normal distribution (currently in the BANOVA package they are assumed to be uncorrelated, which is somewhat restrictive). These error terms capture unobserved differences in the parameters θ_0, θ_1, θ_2, and θ_3 between participants. The first equation regresses the individual-specific intercepts (θ_0) on an overall intercept (β_{00}) and on the between-subjects terms. The parameters in question represent the main effects of X_1, effort (β_{01}), of X_2, motive (β_{02}), and of the X_1X_2 interaction (β_{03}). The second equation regresses the participant-specific parameters of the variable x_1, product category (θ_1), on an intercept (β_{10}) representing the main effect of product category, and on the between-subjects terms. The latter parameters represent the interactions of product category with marketing effort x_1X_1 (β_{11}) and with motive x_2X_2 (β_{12}), and the three-way interaction $x_1X_1X_2$ (β_{13}). The third equation captures the main effect of z_1, product quality (β_{20}), and its interactions with marketing effort z_1X_1 (β_{21}) and with the persuasion motive z_1X_1 (β_{22}), and the three-way interaction $z_1X_1X_2$ (β_{23}). The fourth equation captures the interaction between category and product quality x_1z_1 (β_{30}), the three-way interactions $z_1x_1X_1$ (β_{31}) and $z_1x_1X_2$ (β_{32}), and the four-way interaction $z_1x_1X_1X_2$ (β_{33}).

It is important to note that in BANOVA, all inferences of interest (effect sizes, p-values and so on) are derived from the between-subjects model 2. The within-subjects model is estimated, but the parameter estimates of the within-subjects model are not used any further (with the exception of the calculation of individual-specific indirect effects as explained later).

Model Syntax

BANOVA models are specified in convenient shorthand notation (following standard R syntax). The full model in the example above is specified as ($y \sim x_1{*}z_1$, $\sim X_1{*}X_2$). The "~" separates the

dependent from the independent variables, and the "," separates the within- and between-subjects models. The "*" sign expands the term in question into all main effects and all interactions of its arguments: $X_1*X_2 = X_1 + X_2 + X_1{:}X_2$ (interactions are denoted as $X_1{:}X_2$). If, in the application, there had not been any between-subjects variables, the model would have been specified as ($y \sim x_1*z_1, \sim 1$). If it had been a between-subjects experiment without the within-subjects variables, the within-subjects model would simply be omitted ($y \sim X_1*X_2$). Note that in a between-subjects experiment, the parameters are the same for all participants (unobserved heterogeneity cannot be accommodated). Table 1 lists examples of the syntax for several commonly used models. If y follows a Normal distribution the model is a (hierarchical) linear regression, and is a logistic regression if y follows a Bernoulli distribution.

Credible Intervals and P-Values

Once a model such as the one in Equations 1 and 2 has been specified, the MCMC estimation machinery in the R package draws samples from the posterior distribution of each parameter, $p(\beta|$ data). Suppose that 10,000 draws, β^r, for $r = 1$ to 10,000 have been obtained. The vales of these tenthousand draws are then used to calculate an estimate of the parameter and to conduct statistical tests. The average of the draws is the parameter estimate (the "posterior mean"), and the standard deviation of the draws is a measure of uncertainty about the parameter (the "posterior standard deviation"). The Credible Interval (CI), is the interval that contains the parameter with a certain (usually 95%) probability, and is obtained as the 2.5 and 97.5 percentile points of the vales of the draws β^r.

The one-sided p-value is calculated as the fraction of draws β^r that has a negative (or positive) value. It is an estimate of the probability: $p(\beta \leq 0)$. This probability is often called the "Bayesian p-value," because if it is small, then $H_0{:}\beta \leq 0$ is unlikely (Carlin & Louis, 1998, p.53). The two-sided Bayesian p-value is calculated as two times $p(\beta \leq 0|$ data) or $p(\beta > 0|$data), whichever is smaller (it is the p-value reported by the BANOVA package).

Unfortunately, the classical p-value is often misinterpreted: it is not the probability that the null-hypothesis is true and does not quantify the evidence for the hypothesis. It only allows one to reject the null hypothesis. It has been argued that it should, therefore, be supplemented or replaced with statistical measures of evidence, in particular the likelihood ratio (Nuzzo, 2014; Wasserstein & Lazar, 2016). The likelihood ratio, alternatively called the Bayes Factor (BF), is the most widely accepted measure to quantify how much evidence a data set provides for a hypothesis (Edwards et al., 1963). When the prior is symmetric (for example, the diffuse Normal distribution for the coefficient β proposed above) and the hypothesis is directional, then the Bayesian p-value is a transformation of the Bayes Factor (logit (p)=log(BF); Marsman & Wagenmakers, 2017). This means that in those common cases not only is the Bayesian p-value the probability that the null-hypothesis is true, but that it can also be directly interpreted as the strength of evidence for the null-, or conversely, the alternative hypothesis. Jeffreys (1961) and Kass and Raftery (1995) provide guidelines for the interpretation of the Bayes factor. These guidelines roughly imply that below the commonly used cutoff of $p = .05$ the Bayesian p-value indicates strong evidence, below $p = .01$ it indicates very strong evidence, and below $p = .001$ decisive evidence for the effect.

Effect Sizes And Model Fit

We propose to calculate effect sizes for these hierarchical models building on Gelman and Pardoe (2006). The sums-of-squares (SS) for any effect in a between-subjects equation such as 2, is defined as the difference between the SS of the residuals

Table 1
Examples of BANOVA model specification using R modeling syntax

M	Within-subjects model	Between-subjects model	Terms included in the model
1	$y{\sim}x_1$	~ 1	x_1
2	$y{\sim}x_1{-}1$	~ 1	x_1, no intercept
3	–	$y{\sim}X_1$	X_1
4	–	$y{\sim}X_1{-}1$	X_1, no intercept
5	$y{\sim}x_1$	$\sim X_1$	x_1,X_1, $x_1{:}X_1$
6	$y{\sim}x_1{*}x_2{*}x_3$	~ 1	x_1,x_2,x_3, $x_1{:}x_2$, $x_1{:}x_3$, $x_2{:}x_3$,$x_1{:}x_2{:}x_3$
7	$y{\sim}x_1{*}x_2{*}x_3{-}x_1{:}x_2{:}x_3$	~ 1	x_1,x_2,x_3, $x_1{:}x_2$, $x_1{:}x_3$, $x_2{:}x_3$
8	$y{\sim}x_1$	$\sim X_1{*}X_2$	x_1, X_1,X_1, $x_1{:}X_1$, $x_1{:}X_2$, $X_1{:}X_2$, $x_1{:}X_1{:}X_2$
9	$y{\sim}x_1{+}z$	$\sim X_1$	x_1,z,X_1,$x_1{:}X_1$,$z{:}X_1$
10	$y{\sim}x_1$	$\sim X_1{+}Z$	x_1,X_1,Z,$x_1{:}X_1$,$x_1{:}Z$

Note. 1 denotes a constant associated with the intercept; *denotes a full expansion of the term in question into all main and interaction effects.
[a] x_1,x_2,x_3 are within-subjects factors, z a continuous between-subjects covariate; X_1 and X_2 are between-subjects factors, Z a continuous between-subjects covariate; ~ initiates the within- or between-subjects model; : separates the terms in an interaction.

obtained by setting the coefficient(s) for that effect to zero (SS_0) and the SS of the residuals of the full model (SS_e). Assume that in the example used above, we want to obtain the effect size of the three-way interaction: $x_1X_1X_2$. We first obtain SS_e, the SS of the residuals e_1 of the full model. We then set $\beta_{13} = 0$ and recalculate the residuals e_1 and their SS, which is denoted as SS_0. The SS for the effect is then: $SS_\beta = SS_0 - SS_e$. The proposed generalized partial eta-squared effect size measure is then calculated as follows: $\eta^2_P = \frac{SS_\beta}{s_y + SS_\beta + SS_e}$ Here, s_y is a "correction term" that accounts for the error variance of the within-subjects model in Equation 1 (Nakagawa & Schielzeth, 2013; p. 139; for example, for the binomial, the distribution-specific variance is $\pi^2/3$). We use partial eta-squared because it is not very sensitive to other terms included in the model and, therefore, generalizable across studies (Lakens, 2013). MCMC estimation automatically adjusts it to account for uncertainty in all parameters (Gelman & Pardoe, 2006; Lakens, 2013), and allows its Credible Interval to be calculated. These partial eta-squared effect sizes can be calculated for any model within the BANOVA family. The SS's thus obtained also allow for the calculation of an R^2 measure of fit for all BANOVA models (see for details Gelman & Pardoe, 2006; Nakagawa & Schielzeth, 2013).

Tables of Predictions

The draws of parameter values, β^r, obtained with MCMC can be used to calculate "tables of predictions." These tables contain the predicted values of the dependent variable,$\hat{y}$, at all possible combinations of the levels of the experimental factors: one-way, two-way, or three-way tables. In calculating these tables continuous covariates are set to their mean value (which is zero because of mean centering). In the example above, these would be the tables classified by x_1, by X_1, by X_2, by x_1 and X_1, by x_1 and X_2, by X_1 and X_2, and the three-way table classified by x_1 and X_1 and X_2 (while setting $z_1 = 0$). These tables are often much easier to interpret than the parameter estimates themselves, especially when there are multiple factors with interactions. Because the predicted value in each cell of the table is calculated for each draw of the parameters, its posterior distribution, $P(\hat{y}|\text{Data})$, can be obtained as well. The posterior distribution of the predictions is usually summarized via its mean and Credible Interval. Note that the overlap of two Credible Intervals does not necessarily imply a lack of evidence for a difference between predictions in the corresponding cells in the table (although absence of overlap indicates such evidence). Those comparisons are better done via planned comparisons, as explained below.

What Should the Researcher Report?

What statistics should the researcher report as evidence for a hypothesis? We propose the following (see also Pieters, 2017; Wasserstein & Lazar, 2016; Wilkinson, 1999):

1. The posterior mean of a parameter β, and its Credible Interval,
2. The (Bayesian) *p*-value, and/or the Bayes Factor,
3. The effect size, η^2_P, and its Credible Interval,
4. Tables of predictions, and their Credible Intervals.

Which tables of predictions are relevant will depend on both the hypotheses and results of the statistical tests in the experiment. For example, if a three-way interaction was hypothesized but the experiment did not show evidence for it, there is not much use in inspecting the corresponding three-way table of predictions. These statistics are part of the output of BANOVA.

Possible Follow-Up Analyses

After assessing the evidence for the effect of one or more experimental factors, a researcher often requires more detailed insight into their directions and magnitudes. Planned comparisons, simple effects, and spotlight and floodlight analyses enable these. The BANOVA package includes functions that perform the calculations in question (see the Methodological Details Appendix).

Planned Comparisons

One may wish to further probe main effects of factors with more than two levels, using a Bayesian approach to planned comparisons. Contrast coding allows one to test specific hypotheses on differences between levels of a categorical variable. Rather than using a standard coding for the levels of the factor, such as dummy coding or effects coding, one specifies a unique comparison, or contrast, between specific levels of the factor. For instance, assume that in the marketing/persuasion example above, the within-subjects factor product category (x_1) has three levels: $x_1 = (1,2,3)$, paper towels, canned soup,

and bath towels. Then the contrast $c(-1,1,0)$ specifies a comparison between paper towels and canned soup, and $c(1,0,-1)$ specifies a comparison between paper towels and bath towels (the "c" specifies a vector in R). Based on the draws of the parameters, the CI of a contrast in the predicted values of y can be obtained. If the CI does not contain zero, there is evidence for a difference between the levels of the factor specified in the contrast. The interaction between two factors is calculated as the product of their two contrasts. For example, for a 3-level factor (x_1; product category) with contrast $c(1,0,-1)$ and a 2-level factor (X_1; marketing effort) with contrast $c(1,-1)$, the resulting interaction specifies a comparison of the difference between high and low marketing effort, between the first and third levels of the first factor (paper towels vs. bath towels). Planned comparisons can be made for within- and/or between-subjects factors. See for example Kerlinger and Pedhazur (1973; p. 128–140) and other textbooks for a detailed treatment of planned comparisons and contrast coding.

Simple Effects

Simple effects are often helpful for the interpretation of interaction effects. For instance, in the marketing/persuasion example with model ($y \sim x_1$, $\sim X_1{*}X_2$), to explore the moderating effect of the product category x_1 on marketing effort X_1, simple effects of X_1 can be calculated. These are the differences between low versus high marketing effort for each product category (levels of x_1). The simple effects can be obtained by specifying specific contrasts, with the moderator x_1 represented by dummy coding and the other factor(s) by effect coding. The simple effect of X_1 for the first product category, for example, is the main effect of X_1 when contrasts $c(1,-1)$ for X_1 and $c(0,1)$ for x_1 are specified. For simple effects at some other level of x_1, dummy coding is used with that specific level specified as the baseline. Thus, for three product categories, $x_1 = (1,2,3)$, paper towels, canned soup, and bath towels, specifying the two contrasts $c(1,0,0)$ and $c(0,0,1)$ for x_1 would result in the main effect of X_1 representing the simple effect of marketing effort for canned soup (the baseline level). When exploring three-way interactions (marketing effort × motive × category), the simple effects of a two-way interaction (marketing effort × motive) at a specific level of the third factor (product category), or the simple effect of one factor (marketing effort) at specific levels of the second (motive) and third (category) factors can be similarly specified through contrasts. Based on the draws of the parameters, the p-value, CI, and effect size of a contrast are obtained.

Spotlight Analysis

Spotlight analysis (Rogosa, 1980) involves planned comparisons to probe the interaction between a factor x_1 (product category) and a continuous variable z_1 (product quality). Specifically, one would like to know if there is a difference between the levels of the factor at a prespecified level of the continuous variable. If the factor x_1 (product category) has three levels, one could test the differences in willingness to pay (y) between the first two levels (paper towels vs. canned soup) via the contrast $c(-1,1,0)$ at a specific level of product quality, $z_1 = 1$, for example. Using the draws of the parameters, the p-value, CI, and effect size of that contrast can be calculated. Spotlight analyses can be done for within- and/or between-subjects factors and variables. It applies as well when both terms in the interaction are (categorical) factors.

Floodlight Analysis

Floodlight analysis (Bauer & Curran, 2005; Johnson & Neyman, 1936; Spiller et al., 2013), overcomes spotlight analysis' limitation of having to select one specific value of the continuous variable. It provides an estimate of the range of values at which there is evidence for a difference between the factor levels. The endpoints of this range are called the Johnson–Neyman (JN) points. To continue the example, we may wish to probe the interaction between x_1 (product category) and z_1 (product quality). Assuming x_1 has two levels (-1 and 1), the difference between them can be shown to be $2 \times (\beta_{10} + \beta_{30}z_1)$ (assuming all level-2 terms $X_1 = 0$, and $X_2 = 0$). Setting this expression to zero and solving for z yields: $z^0 = -\beta_{10}/\beta_{30}$. Or, we may probe the interaction between X_1 (marketing effort) and z_1 (product quality). The difference between the two levels of X_1 is $2 \times (\beta_{01} + \beta_{21}z_1)$ (assuming all other terms to be zero), which yields: $z^0 = -\beta_{01}/\beta_{21}$. Because z^0 is calculated at each of the draws of the parameters, its posterior distribution and CI can be obtained. The endpoints of the CI of z^0 are the JN points: for values of z inside the CI there is no evidence of a difference between the levels of x_1, only for values outside of the CI there is evidence for such a difference. If a JN point is inside the data range, the difference between the levels of x_1 outside of the JN point can be positive or negative,

which is indicated by the sign($\beta_{10} + \beta_{30}$JN). If one or both of the JN points are outside the observed range of z they can be ignored (Hayes, 2013; p.240).

To further continue the example, the difference between the levels of marketing effort, X_1 along levels of product quality, z_1 may also depend on the motive, X_2 because of their interactive effect. Now, the difference between the levels of X_1 can be shown to be $2\times(\beta_{01}+\beta_{03}X_2+\beta_{21}z_1+\beta_{23}X_2z_1)$. Setting this to zero and solving for z yields: $z^0 = -\frac{\beta_{01}+\beta_{03}X_2}{(\beta_{21}+\beta_{23}X_2)}$. Thus, there are two floodlight ranges, one for $X_2 = 1$: $z^0 = -\frac{\beta_{01}+\beta_{03}}{\beta_{21}+\beta_{23}}$ and one for $X_2 = -1 : z^0 = -\frac{\beta_{01}-\beta_{03}}{\beta_{21}-\beta_{23}}$. For each of these the CI of z^0 provides the JN points, which can be calculated from the draws of the parameters. Note that floodlight analyses can be conducted for repeated measures data with a nonNormal dependent variable. While in standard repeated measures models these JN points are only approximations (Bauer & Curran, 2005), this caveat does not hold when using the Bayesian approach proposed here.

Study 1: Gist Perception

We illustrate the application of BANOVA by reanalyzing data from a study on rapid perception of the gist of color and grayscale ads during brief and blurred exposures (Wedel & Pieters, 2015). The study involved 116 participants in a 5 (blur: normal, low, medium, high, very high) × 2 (color: full color, grayscale) × 2 (typicality: typical ads, atypical ads) mixed design. Participants were exposed to 32 full-page ads, of which 16 were typical and 16 were atypical for their category. Ad images were blurred with Gaussian blur filters and rendered in grayscales or full color. Participants were flashed an image for 100msec. and asked to identify whether it showed an ad or not. The dependent variable is the number of times, out of 16, that a participant identified the typical and atypical ads correctly. The key hypothesis was that the gist perception of typical versus atypical ads is better at higher levels of blur, and that color helps this even more (see for details Wedel & Pieters, 2015).

The BANOVA R package is used for the analysis (the Methodological Details Appendix provides the instructions on how to install it and the commands needed to run all analyses for this application). The dependent variable follows a Binomial distribution ($K = 16$). There is one within-subjects factor, typicality (typ), and two between-subjects factors, color (col) and blur (blr). The repeated measures BANOVA model is specified as: (*y*~typ, ~col*blr). About 100,000 draws were taken and one in 10 draws were retained. Convergence tests reported by BANOVA (Geweke, 1992; Heidelberg & Welch, 1983) indicate that the algorithm converged.

Table 2
Gist perception study: BANOVA results; for each term in the model the degrees of freedom (DF) sum-of-squares (SS), partial eta-squared (η_P^2) and Bayesian p*-value are provided*

Term	DF	SS	η_P^2	*p*-value
col	1	5.10	.006	.008
blr	4	29.39	.031	<.001
typ	1	21.81	.027	<.001
col:blr	4	1.19	.001	.160
col:typ	1	2.11	.003	.016
blr:typ	4	13.90	.017	<.001
col:blr:typ	4	1.93	.002	.032

The effect sizes and *p*-values in Table 2 show that there is decisive evidence for the typ × blr ($p < .001$; $\eta_P^2 = 0.017$), and strong evidence for the typ × col ($p = .016$; $\eta_P^2 = .003$) and typ × col × blr ($p = .032$, $\eta_P^2 = .002$) interactions, but with relatively small effect sizes. Table 3 and Figure 1 show the predicted number of correct ad identifications (out of 16) for all combinations of color, typicality, and blur. They reveal that typical ads are more often correctly identified than atypical ads. Typical color ads are better identified than typical grayscale ads, but that seems to be the case only when they are blurred. A planned comparison with typicality contrast $c(0,1)$ and blur contrast $c(0,1,1,1,1)$, which specifies the simple effect of color for normal typical ads, shows indeed that there is little evidence for a difference between normal typical color and grayscale ads ($\beta = -0.061$ with CI = (−0.341, 0.231), $p = .688$, $\eta_P^2 = .005$ with CI (−0.001,0.030)). Furthermore, Figure 1 shows that the lines for atypical color and atypical grayscale ads are very close, suggesting that there is no difference in identification between these ads for any level of blur. The figure also reveals a substantial negative trend in the identification of typical ads as blur increases, but no systematic effect of blur for atypical ads.

We use the Bayesian approach to floodlight analysis to further shed light on these effects. First, reanalysis of the data with blur as a continuous variable (the natural logarithm of the pixel radius of the blur filter, which ranges from 0 to 5.481) confirms these findings and again shows strong evidence for the typ × col × blr interaction ($\beta = .038$ CI = (0.007, 0.072), $p = 0.020$, $\eta_P^2 = .008$). Second, we explore the difference between color and grayscale

Table 3
Gist perception study: Predicted number of correct ad identifications out of 16, for combinations of color, typicality, and blur; posterior mean with 95% CI

	Color			Grayscale		
Blur	2.5%	Mean	97.5%	2.5%	Mean	97.5%
Typical						
1	11.88	13.00	13.90	12.19	13.30	14.10
2	12.11	13.35	14.16	9.70	11.16	12.47
3	10.12	11.53	12.63	9.09	10.51	11.85
4	10.11	11.54	12.83	7.39	8.87	10.27
5	7.60	9.16	10.58	4.64	6.12	7.84
Atypical						
1	8.87	10.43	11.71	8.47	10.05	11.47
2	6.40	8.11	9.86	5.43	6.95	8.51
3	7.14	8.93	10.48	7.35	8.83	10.23
4	7.68	9.13	10.46	7.47	9.21	10.59
5	6.80	8.30	9.81	5.75	7.33	8.96

ads at different levels of blur as moderated by ad typicality, using Bayesian floodlight analysis. It produces two sets of JN points: the 95% floodlight is (−20.22, 48.02) for atypical ads and (−9.24, 2.91) for typical ads. Both JN points are outside the data range for atypical ads, which confirms that there is no difference between color and grayscale for any level of blur for these ads (Figure 1). The right JN point is inside the data range for typical ads, which reveals that there is strong evidence of a difference between color and grayscale typical ads for levels of (log) blur higher than 2.91, which is a bit below blur level 2 (Table 3, Figure 1). Color thus mitigates the detrimental effect of blur on ad identification for typical, but not for atypical ads. The results of the BANOVA analyses of these data are very similar to those reported by Wedel and Pieters (2015), who used Bayesian multilevel logistic regression models for their analyses, but here we add the new planned comparisons, simple effects, and floodlight analyses.

Analysis of Mediation and Moderation

In consumer psychology providing evidence that an experimental treatment has an effect is often not sufficient: for theory development: one needs to establish the mechanism by which the effect is produced, and investigate whether there are factors

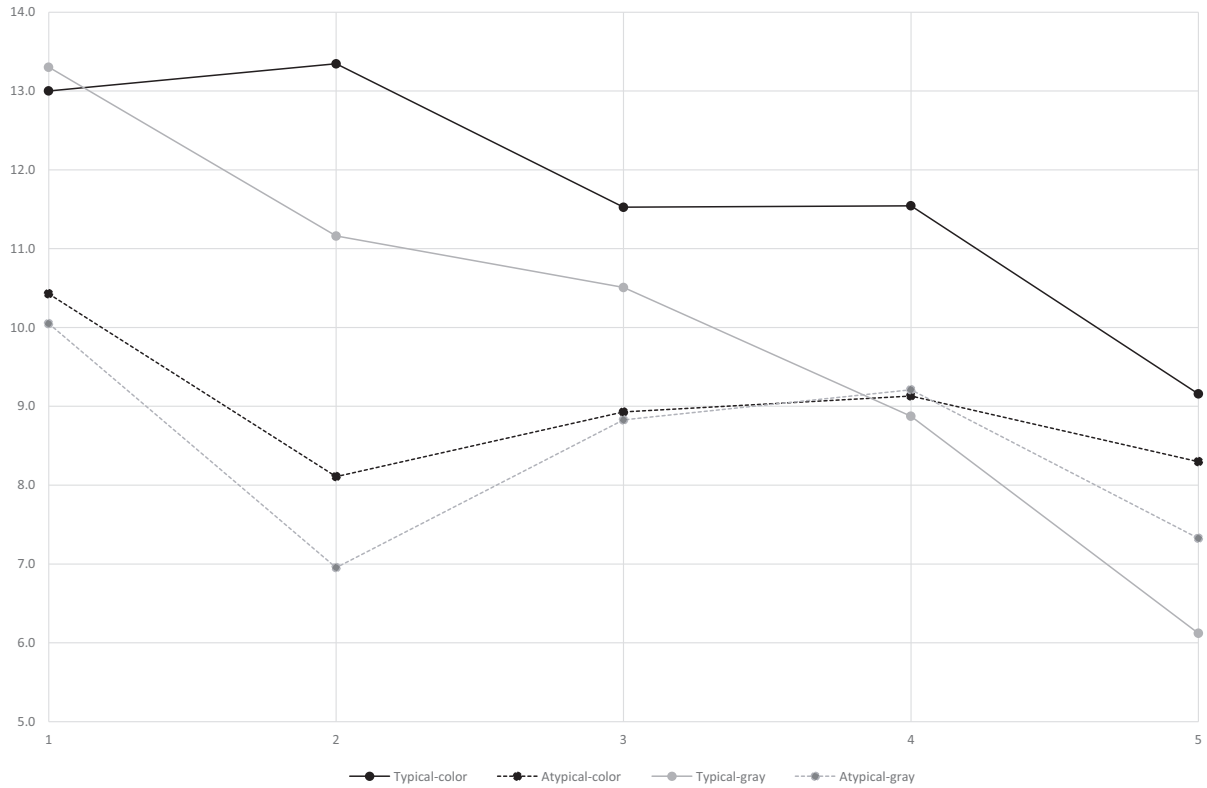

Figure 1. Gist perception study: Plot of predicted number of correct ad identifications against the level of blur for typical and atypical color and grayscale ads

that affect the magnitude of the treatment's effect or its mechanism. That is, one needs to test for, respectively, mediation, moderation, or moderated mediation (Bauer, Preacher, & Gil, 2006; Pieters, 2017; Preacher, Rucker, & Hayes, 2007; Yuan & McKinnon, 2009). Repeated measures experiments provide several advantages for those purposes (Judd, Kenny, & McClelland, 2001).

Mediation In Repeated Measures Experiments

First, in between-subjects experiments not only is the mediator measured at the same time as (or even after) the dependent variable, but participants are randomly assigned to levels of the experimental factor and not to levels of the mediating variable. Therefore, these designs only provide correlational evidence of mediating mechanisms (Bullock et al., 2010; Pirlott & MacKinnon, 2016). On the other hand, repeated measures designs allow for temporal precedence of the measurement of the mediator, and, therefore, may provide a higher degree of confidence in the causality of the effects under study. Second, because popular between-subjects designs are cross-sectional, they only accommodate mediators that transmit the effect of the experimental treatment contemporaneously. On the contrary, longitudinal repeated measures designs allow the measurement of the mediator to temporally precede that of the dependent variable (MacKinnon et al., 2007), and are thus particularly useful when the underlying mediating mechanism takes time to reveal itself or varies in strength as time passes (Selig & Preacher, 2009). Third, some mechanisms that mediate treatment effects operate between-subjects, for example via psychological traits such as need-for-cognition or openness-to-change, while other mechanisms operate within-subjects, for example via psychological states such as attention or emotions. In some cases the same mediator may operate both within- and between-subjects. Properly designed repeated measures experiments allow one to disentangle these different mechanisms (MacKinnon et al., 2007). Fourth, a between-subjects design only allows for the estimation of average mediation effects, and, hence, does not allow one to evaluate the extent of unobserved heterogeneity in causal relations (Bullock et al., 2010). If dependent variables are nonNormal and there is heterogeneity of causality, averaging of participant-specific causal effects may even result in biased inferences. Repeated measures designs are required to assess if participants differ in the effect of the experimental treatment on the dependent variable or on the mediator, and in the indirect effect of treatment transmitted via the mediator on the dependent variable (Pirott & MacKinnon, 2016).

Banova Models For Mediation And Moderation

The framework and model syntax provided in Table 3 accommodate the specification of a plethora of mediation, moderation, and moderated mediation relationships and thereby allow for an appropriate analysis of a wide variety of assumed causal mechanisms using repeated measures data. We provide 11 examples of those models in Table 4, which cover many common cases. The models in the Table involve a single mediator, a single moderator, and no other covariates, but extensions are straightforward and accommodated within the framework. The table indicates at which level the experimental factor, the mediator, and the moderator are measured or manipulated, and it provides the within- and between-subjects models for the dependent variable (y) and the mediator (m or M). Most cases presented in the Table involve repeated measures designs, but several between-subjects designs are accommodated as well.

The most common cases are models 1–5. Model 1 is for a repeated measures design with a factor manipulated within-subjects and a within-subjects mediator variable. Model 2 is for a between-subjects design with a between-subjects manipulation and a between-subjects mediator variable. Model 3 is for a repeated measures design with two within-subjects factors and their interaction (moderation). Model 4 is for a between-subjects design with two factors and their interaction. Finally, model 5 is for a repeated measures design with one within-subjects factor and one between-subjects factor, where the between-subjects factor moderates the effect of the within-subjects factor.

The models in Table 4 are presented in syntax that can directly be used to run the BANOVA models in R. All models allow for various distributions of the dependent variable (Normal, Poisson, Binomial, Multinomial, etc.), but the mediator should be continuous and Normally distributed for the calculation of indirect effects to be valid. Figure 2 illustrates four models that accommodate moderated mediation.

This framework in Table 4 generalizes various models for repeated measurements that have been proposed in the (consumer) psychology literatures. For example, Judd et al. (2001) case 3 is model 3 in Table 4, and their case 2 is model 5. Krull and MacKinnon's (2001, p.254) 1-1-1 model (numbers

Table 4
Examples of mediation and moderated mediation models for repeated measures experiments; the terms in the within-and between-subjects model are provided, as well as the equations for the dependent and mediator variables[a,b]

	Model	Within-subjects	Between-subjects	(x.1) *Y*-equation (x.2) *M*-equation
1	Within-ss mediation	x_1,m	–	(1.1) y~x_1+m, ~1 (1.2) m~x_1, ~1
2	Between-ss mediation (*between-subjects design*)	–	X_1, M	(2.1) y~X_1+m (2.2) M~X_1
3	Within-ss moderation	x_1,x_2	–	(4.1) y~x_1*x_2, ~1
4	Between-ss moderation(*between-subjects design)*	-	X_1,X_2	(5.1) y~X_1*X_2
5	Between-ss moderation of within-ss treatment	x_1	X_1	(6.1) y~x_1,~X_1
6	Between-ss mediation(*repeated measures design*)	-	X_1,M	(3.1) y~X_1+M,~1 (3.2) M~X_1
7	Within-ss mediation of between-ss. treatment, with moderated mediation.	m	X_1	(7.1) y~m,~X_1 (7.2) m~1,~X_1
8	Within-ss moderated mediation	x_1,x_2,m	–	(8.1) y~x_1+x_2*m,~1 (8.2) m~x_1+x_2,~1
9	Between-ss moderated mediation	–	X_1,X_2,M	(9.1) y~X_1*X_2+X_2*M (9.2) M~X_1*X_2
10	Within-ss mediation of within-ss treatment, between-ss moderation	x_1,m	X_1	(10.1) y~x_1+m, ~X_1 (10.2) m~x_1, ~X_1
11	Within-ss mediation of between-ss treatment, between-ss moderation	M	X_1,X_2	(11.1) y~m, ~X_1+X_2 (11.2) m~1, ~X_1+X_2

[a]Lowercase indicates within-subjects variables, uppercase indicates between-subjects variables.
[b]The symbol y indicates the dependent variable, X_1 (or x_1) indicates the experimental factor, M (m) stands for the mediator, and X_2 (x_2) for a moderator. The experimental factor is categorical with two or more levels, mediators are continuous, and moderators can be continuous or categorical with two or more levels.

indicate the measurement level of *X*-*M*-*Y*) is model 1 and their 2-2-1 model is model 6 in Table 4. Model 7 is a generalization of their 2-1-1 model, where in addition the effect of m on y is moderated by X_1. As another example, model 10 could be used to analyze a moderation-of-process design (Pirott & MacKinnon, 2016, p.11), with a within-subjects treatment x_1, a within-subjects (measured) mediator m, and a between-subjects treatment X_1 that blocks or enhances the effect of the mediator m, for stronger evidence of causality of the mechanism.

Indirect Effects in Bayesian Models For Mediation and Moderated Mediation

The indirect, or causal, effect of an experimental treatment is calculated as the product of two coefficients that capture the effect of the treatment on the mediator (α), respectively the effect of the mediator on the dependent variable (β). It has long been recognized that the indirect effect, $\gamma = \alpha \times \beta$, does not have a known distribution, because the product of two Normal distributions does not follow a known distributional form. This poses a problem for statistical testing of the underlying mechanism that has seen several solutions, such as the Bootstrap (Bollen & Stine, 1990) and Bayes Credible Intervals (Zhang et al., 2009). In addition, standard deviations and tests of the indirect effect in classical multilevel mediation and moderated mediation analyses are based on approximations only (Krull & MacKinnon, 2001). Using a Bayesian approach, however, the posterior distribution and Credible Intervals of the indirect effect can be calculated accurately in all cases (Zhang et al., 2009).

As an example, consider a between-subjects experiment with one 2-level factor and a continuous between-subjects mediator (model 2 in Table 4). The syntax is: (y~X_1 + M), and (M~X_1). The coefficient α_1 captures the effect of X_1 on M; β_1 captures the effect of M on y, the indirect effect of X_1 is thus $\gamma = \alpha_1 \times \beta_1$. The Bayesian estimation algorithm produces draws (indexed by $r = 1,\ldots,R$) of the posterior distributions of the coefficients α_1 and β_1. If we have, for example, R = 10,000 draws (values) of α_1^r and β_1^r then the indirect effect can be calculated for each of these 10,000 draws as: $\gamma^r = \alpha_1^r \times \beta_1^r$. This results in 10,000 values of the indirect effect that are draws from its posterior distribution: $P(\gamma|\text{Data})$. They can be used to calculate an estimate, CI, and p-value to assess the evidence for the indirect effect (Zhang et al., 2009; Yuan & McKinnon, 2009).

As an example for a repeated measures experiment, assume one 2-level within-subjects factor, a continuous within-subjects mediator, and a 2-level between-subjects moderator. This example corresponds to model 9 in Table 4, the syntax for which is (y~x_1+m,~X_1), and (m~x_1,~X_1). This model allows

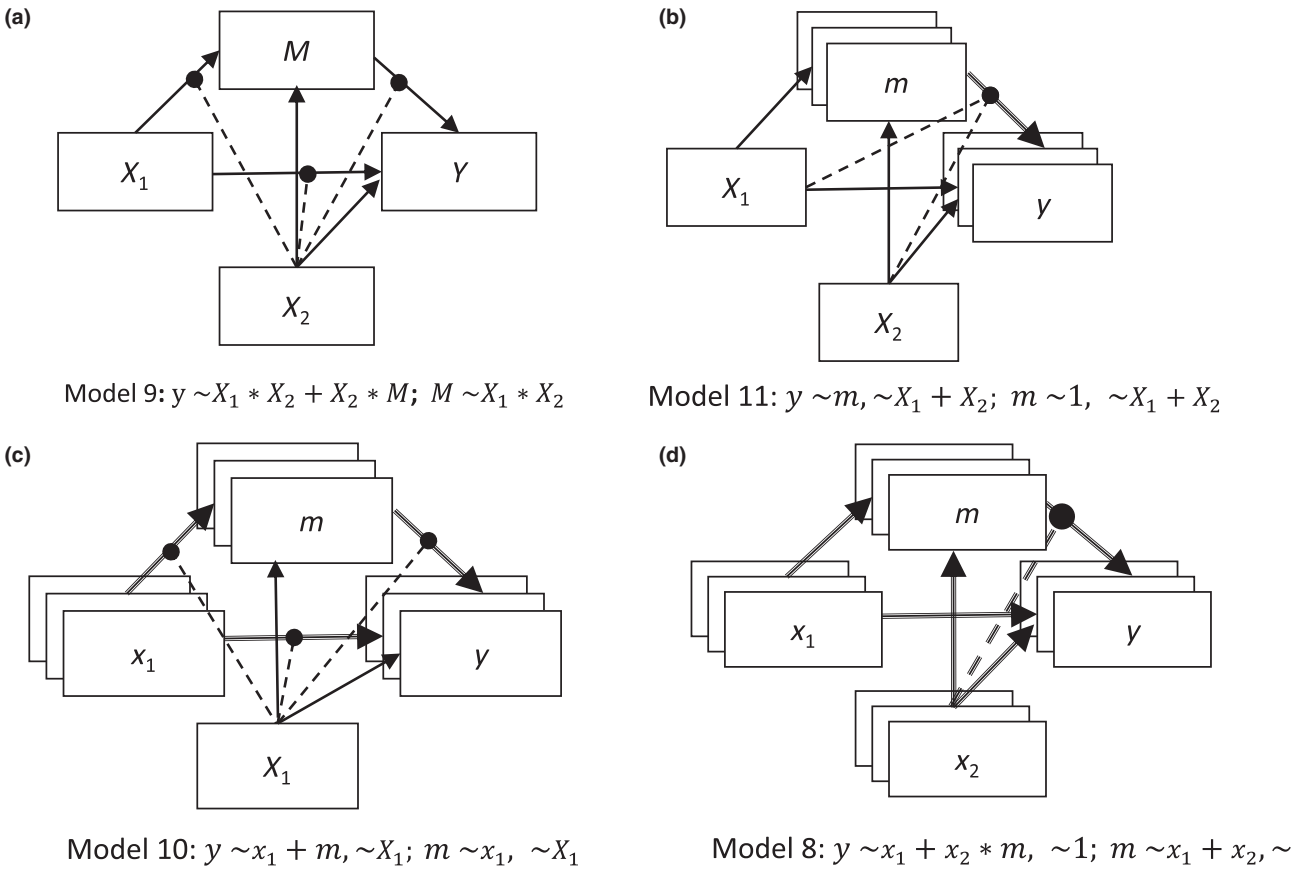

Figure 2. Diagrams for four moderated mediation models selected from Table 4. Notes: Solid arrows indicate direct effects; dotted lines indicate moderating effects. Single box (upper case symbol) indicates variable measured/manipulated between-subjects; multiple boxes (lower case symbol) indicate variable measured/manipulated within-subjects. Compound arrows indicate heterogeneous individual-level and average effects. Model numbers and equations correspond to those in Table 4

one to test for mediation and moderated mediation: X_1 moderates the direct effects of x_1 and m on y, as well as the effect of x_1 on m. The direct effects of x_1 on y are $\beta_{10} + \beta_{11}X_1$ (see Equation 2); note that there are two direct effects because of moderation by X_1: one for $X_1 = 1$ ($\beta_{10} + \beta_{11}$), and one for $X_1 = -1$: ($\beta_{10} - \beta_{11}$). There are thus also two indirect effects of x_1: $\gamma = (\beta_{20} + \beta_{21}X_1) \times (\alpha_{10} + \alpha_{11}X_1)$, because of the moderating effect of X_1: one indirect effect obtained by substituting $X_1 = 1$ and one by substituting $X_1 = -1$. For this model, and any other model specified within the general framework the posterior distributions of all indirect effects can be obtained from the draws of the parameters, and thus their Credible Intervals and p-values can be obtained.

p-values and Effect Sizes of Indirect Effects

With much of the scientific interest in consumer psychology resting on indirect effects, reporting Credible Intervals, p-values, and effect sizes for these effects is "needed to convey the most complete meaning of the results" (APA, 2010, p.33). Once the draws of an indirect effect γ^r are obtained as explained above, the Credible Interval can be calculated as the 2.5 and 97.5 percentile points of these draws. The one-sided p-value for the hypothesis $H_0 : \gamma \leq 0$ is the fraction of draws γ^r that has a negative value. The two-sided p-value is calculated as twice the fraction of draws that has a negative or positive value, whichever is smaller. These quantities provide measures of the amount of evidence for the indirect effect.

For the effect size of the indirect effect, we propose to calculate a (new) generalized partial eta-squared measure. Several measures have been previously proposed to capture the size of the indirect effect in the context of mediation models. They fall into the following broad categories: 1. standardized estimates of indirect effects, 2. ratios of the relative magnitude of the indirect relative to the direct effect, 3. indices of explained variance based on

mediation model residuals (Kelly & Preacher, 2012; Lachowicz, Preacher, & Kelly, 2011; Preacher & Kelly, 2018). Desiderata for effect size indices, as well as advantages and disadvantages of the existing measures are discussed by Preacher and Kelly (2018). They also point to the desirability of interval estimates for these quantities.

The generalized partial-eta-squared (η_P^2) measure proposed here falls in the category of measures of explained variance. It offers the advantage of an interpretation as a proportion, and of being comparable across studies (Lakens, 2013). Furthermore, it is calculated in a way that is conceptually similar to the other effect sizes in BANOVA (see above), and is available regardless of the distribution of the data or the model formulation. It accommodates the uncertainty about the parameters (because it is averaged across the draws of all parameters), and its' Credible Intervals can be calculated as well. We define the SS for the indirect effect as the difference between the SS of the residuals obtained by setting the indirect effect to zero and the SS of the residuals of the full model. We first obtain the residuals of the full model and their sums-of-squares: SS_e. Then we subtract the effect of the experimental factor on the mediator from the observations of the mediator, which yields M_0 (or m_0). We substitute M_0 (or m_0) for M (or m) in the model for y and recalculate the residuals and their sums-of-squares, which is indicated as SS_0. The SS for the indirect effect is then: $SS_\gamma = SS_0 - SS_e$, and the generalized partial eta-squared effect size is: $\eta_P^2 = \frac{SS_\gamma}{s_y + SS_\gamma + SS_e}$~. Here, s_y is a "correction term" that accounts for the error variance of the within-subjects model (see Nakagawa & Schielzeth, 2013).

These calculations extend to the general case with within- and/or between-subjects factors, mediators, and moderators. In those cases, the SS for the indirect effects is obtained by calculating the direct effects of the experimental factor on the mediator M (or m) for all levels of relevant moderating factors, subtracting that from M (or m), and recalculating the SS of the error. The draws from the posterior distribution of the effect sizes of the indirect effects enable the calculation of their Credible Intervals. It is important to note that both η_P^2 and the limits of its CI are not necessarily constrained to be positive and may take on negative values if the effect sizes are very small.

Study 2: Direct and Indirect Conditioning

We apply BANOVA to data from a study by Sweldens et al. (2010) in order to illustrate an application to a repeated measures design with mediation and moderation. Sweldens and coauthors investigated evaluative conditioning, by testing if brand attitudes can be influenced by showing the brands together with pleasing pictures. Attitude change via conditioning can result either from a direct transfer of affect from the picture to the brand, or from an indirect association of the brand and the picture in memory. In Sweldens' et al. (2010) experiment 1, indirect conditioning was implemented by presenting a brand sequentially with the same picture, and direct conditioning was implemented by presenting it simultaneously with different pictures. The pictures were either neutral or positive; three brands received the neutral and three the positive conditioning.

We reanalyze part of the data according to a mixed design with one within-subjects factor (*cond* = neutral, positive) and one between-subjects factor (*type* = indirect, direct). We dropped the "reevaluation" level of the within-subjects factor, and investigate whether the conditioning effect is mediated by the attitudes toward the pictures (*pict*). We also test if the effect of the mediator is moderated by the between-subjects factor (*type*). Because the authors used a two-stage experimental design, the number of missing values for the mediator is larger than what one would want to impute and we use only data without missing values (n = 148, N = 888). The R package "BANOVA" is used for the analysis (the Methodological Details Appendix provides the commands needed to run the analyses). The Shapiro–Wilk Normality test reveals that dependent variable *att*, an average of three 7-point scales, is approximately Normally distributed. Model 10 in Table 4 and Figure 2c: (*att*~*cond+pict*, ~*type*), is used for the analysis. Estimation details are the same as in study 1, the convergence tests indicate that the algorithm converged. We perform a Bayesian mediation analysis, which allows us to calculate the direct and indirect effects, effect sizes, their Credible Intervals, and p-values. In addition, we report the percentage of participants with a positive individual-level indirect effect.

Table 5 shows the effect sizes and the p-values. The first column refers to the main effects of the within-subjects variables (intercept, cond, pict), the second column refers to the main effect (intercept) of the factor *type* and its moderating effects (of cond and pict). There is decisive evidence for an effect of the mediator (*pict*) with a medium effect size (β = 0.269 with 95%CI = (0.182, 0.354), $p < .0001$; η_P^2 = .06 with 95%CI = (0.04,0.09)). There is little evidence for an effect of the conditioning

Table 5
Conditioning study: Results of the BANOVA of brand attitudes; (a) provides effect sizes (with 95%CIs); (b) provides Bayesian p*-values*

	Intercept	type
(a) Effect sizes (95% Credible Interval)		
Intercept	0.9272 (0.922, 0.932)	0.0211 (0.005, 0.039)
cond	0.0090 (0.001, 0.029)	0.0052 (−0.001, 0.260)
pict	0.0644 (0.040, 0.090)	0.0012 (−0.001, 0.007)
(b) *p*-values		
Intercept	<.0001	.004
cond	.328	.368
pict	<.0001	.692

[a]The columns labeled "intercept" refer to the (average) effects of the within-subjects variables (rows: *cond*, *pict*), the columns labeled "*type*" refer to the average effects of the between-subjects variable *type* (row: intercept), and its interactions with the within-subjects variables (rows: *cond* and *pict*).

manipulation after the mediator is accounted for ($p = .328$; $\eta_P^2 = .009$ with 95%CI = (0.001, 0.029)). There is very strong evidence for a main effect of *type* of conditioning with a small effect size ($p = .004$; $\eta_P^2 = .02$ with 95%CI = (0.005,0.039)). There is little evidence for the *cond* × *type* interaction ($p = .368$; $\eta_P^2 = .005$ with 95%CI = (−0.001, 0.025); note that the 95%CI of the effect size of an effect for which $p < .05$ may contain zero, but not necessarily needs to do so), or for the *pict* × *type* interaction ($p = .692$; $\eta_P^2 = .001$ with 95% CI = (−0.001,0.007)). Table 6 contains the predictions classified by *cond* and/or *type*.

The mediator (*pict*) is approximately Normally distributed and model 10 in Table 4 is used for its analysis: (*pict~cond,~type*). Estimation details are the same as above. Table 7 shows the effect sizes and *p*-values, which reveals decisive evidence for the impact of the conditioning (*cond*) manipulation on the mediator, with a very large effect size ($p < .0001$; $\eta_P^2 = .480$ with 95%CI = (0.448,0.513)). And although there is strong evidence for the differences between *types* of conditioning ($p = .004$; $\eta_P^2 = .01$ with 95%CI = (0.001,0.03)), there is little evidence that *type* moderates the effect of cond ($p = .336$; $\eta_P^2 = .001$ with 95%CI = (−0.007,0.01)).

There is decisive evidence for the indirect effect of *cond* with a medium-to-large effect size ($p < .0001$; $\eta_P^2 = .134$ with 95%CI = (0.078, 0.200)). For *type* = "simultaneous", the *indirect effect* of *cond* = "pos" is 0.34 (95%CI = (0.12, 0.58)). For *type* = "sequential" the indirect effect of *cond* = "pos" is 0.42 (95% CI = (0.30, 0.55)). None of the Credible Intervals cover zero, and while the latter is somewhat higher there is considerable overlap in the Credible Intervals. For *type* = "sequential", 98%, and for *type* = "simultaneous", 99% of the participants have a positive indirect effect of *cond* = "pos". This reveals that almost all participants in the study show evidence of an indirect effect of the (positive) conditioning manipulation (*p*-values are calculated as well but significance tests at the individual level have very little power because of the small number of observations per participant).

The results of the analyses of Sweldens et al. (2010) are not directly comparable to those reported

Table 6
Conditioning study: Predictions of brand attitudes and 95% CI; (a) and (b) provide the main effects of both factors (cond, type); (c) provides their joint (interactive) effect

Grand mean	2.50%	97.50%	
4.133	4.002	4.256	
Cond	Mean	2.50%	97.50%
(a)			
pos	4.070	3.906	4.265
neu	4.196	3.993	4.393
Type	Mean	2.50%	97.50%
(b)			
simultaneous	3.942	3.705	4.164
sequential	4.323	4.159	4.466

Cond	Type	Mean	2.50%	97.50%
(c)				
pos	simultaneous	3.939	3.635	4.258
pos	sequential	4.201	3.985	4.387
neu	simultaneous	3.946	3.605	4.285
neu	sequential	4.446	4.238	4.664

Table 7
Conditioning study: Results for the BANOVA of the mediator (picture attitudes); (a) provides effect sizes (with 95%CIs); (b) provides Bayesian p*-values*

	Intercept	Type
(a) Effect sizes (95% Credible Interval)		
Intercept	0.937 (0.933, 0.941)	0.013 (0.001, 0.030)
Cond	0.480 (0.448, 0.513)	0.001 (−0.007, 0.01)
(b) *p*-values		
Intercept	<.0001	.004
Cond	<.0001	.336

[a]The columns labeled "intercept" refer to the (average) effects of the within-subjects variable (rows: *cond*); The columns labeled "type" refer to the average effects of the between-subjects variable type (row: intercept) and its interaction with the within-subjects variable (rows: *cond*).

here, because we utilize only part of their data and omit some of the factors and levels. Yet, because the variables (*att*, *pict*) are approximately Normally distributed and they used a multilevel model similar to the BANOVA model used here, the main conclusions from both analyses are similar. But in addition, our analyses reveal that there is no moderating effect of the type of conditioning nor a moderated mediation, that the effect size of the indirect effect of conditioning is medium-to-large, and that virtually all participants have a positive indirect effect.

Study 3: Mediation Analysis of Processing Fluency Scale

We next analyze five data sets from study 2 in Graf et al. (2017) to further illustrate applications to repeated measures designs with mediation. The five data sets were used to test fluency effects, manipulated within-subjects at two levels. The data sets pertain to the effects of: a. readability of statements on truth judgments (6-point scale); b. car design typicality on liking (100-point scale); c. art pictures' symmetry on liking (100-point scale); d. the number of exposures (0 or 8) to Kanji characters on liking (100-point scale); and e. ease of pronunciation of food additives on risk perception (7-point scale). One purpose of the study was to test whether the most frequently used single semantic differential scale (100-point: difficult - easy) mediates all these manipulated fluency effects. The sample for this part of the study consisted of 254 respondents.

We use BANOVA to reanalyze these data sets (the Methodological Details Appendix provides the commands needed to run the analyses). The dependent variables in data sets a and e failed the Shapiro–Wilk Normality test. While Graf et al. (2017) assumed a Normal distribution, we thus use an ordinal Multinomial model with 6, respectively 7 categories, for the analyses of data sets a and e. The (standardized) dependent variable in the other data sets is analyzed assuming a Normal distribution. Model 1 in Table 4 ($y \sim m + x$, ~ 1) is used, with the subjective fluency mediator (m) and the fluency manipulation (x) as within-subjects variables. The model ($m \sim x$, ~ 1) is used for the mediators (the multilevel model in Graf et al. (2017) included a random effect for stimuli, but because of the relatively small number of stimuli we include a fixed effect, where that effect is identifiable). Estimation details are the same as in study 1, the convergence tests indicate that the algorithm converged. We perform Bayesian mediation analysis to calculate the indirect effects and associated statistics, and we also report the percentage of participants with a positive indirect effect. Table 8 contains the results.

There is decisive evidence that all five fluency manipulations caused participants to experience a higher level of subjective fluency ($p < .0001$), and that subjective fluency positively affected participants' judgments of statement truth, car liking, art liking, Kanji liking, and food risk ($p < .0001$). For statement truth ($\eta_P^2 = .270$) and food risk perception ($\eta_P^2 = .339$), the sizes of the effect of the manipulation on subjective fluency are large. Interestingly, unlike in the original research, the larger effect sizes occur for the dependent variables analyzed with an ordinal Multinomial model. The effect sizes for the art liking data are medium, $\eta_P^2 = .065$, while for the remaining two data sets (Kanji and car liking) they are small. Note that even though the distribution of the dependent variable differs between the five data sets, in all cases the η_P^2 effect size was calculated, with Credible Intervals. For all five data sets, the 95% CI of the indirect effect does not cover zero. The Bayesian p-values indicate that each of the data sets presents decisive evidence ($p < .0001$) for the indirect effects of the manipulations via subjective processing fluency (Table 8). For statement truth ($\eta_P^2 = .062$) and food risk perception ($\eta_P^2 = .046$), the sizes of the indirect effects are medium, but for the other studies they are small, with $\eta_P^2 = .01$ or lower. Across data sets, the percentage of participants that show a positive indirect effect ranges from 63.4% (car design) to 90.6% (statement truth).

The main findings from the BANOVA analyses are largely similar to those reported by Graf et al. (2017). Nevertheless, the BANOVA analyses uncover stronger effects of the mediator in data sets c and e, and a stronger effect of the fluency manipulation on the dependent variable in data set e. But, except for data set e, the indirect effects of the fluency manipulations were found to be smaller. The present analyses reveal the additional insights afforded by BANOVA, in providing effect sizes for (direct and) indirect effects and their Credible Intervals, p-values of the indirect effects, and participant-specific estimates of indirect effects.

Discussion

This article introduced a Bayesian framework for ANOVA of repeated measures, mixed within–between-subjects experiments, as well as for standard

Table 8

Processing Fluency studies: columns 3–5 show the effect of x *on the mediator, columns 6–8 show the effects of* x *and* m *on the y-variable; columns 9–12 show the indirect effects, all with effect sizes,* p*-values (and 95% CIs); the last column shows the percentage of participants with a positive indirect effect. Differences in findings with Graf et al. (2017) are described in the text*

		Effect on mediator			Effect on *y*			Indirect effect			
Study	IV	*a*	η^2	*p*-value	*b*	η^2	*p*-value	*c*	η^2	*p*-value	%
a	*x*	0.453 (0.404, 0.499)	0.270 (0.253, 0.289)	<.0001	−0.146 (−0.230,−0.053)	0.012 (0.003, 0.023)	<.0001	0.329 (0.258, 0.385)	0.062 (0.039, 0.083)	<.0001	0.906
	m				0.726 (0.594, 0.871)	0.201 (0.164, 0.235)	<.0001				
b	*x*	0.094 (0.037, 0.126)	0.026 (0.012, 0.040)	<.0001	0.196 (0.140, 0.244)	0.088 (0.060, 0.123)	<.0001	0.012 (0.003, 0.021)	0.000 (0.000, 0.001)	<.0001	0.634
	m				0.122 (0.044, 0.206)	0.034 (0.008, 0.075)	<.0001				
c	*x*	0.186 (0.130, 0.235)	0.065 (0.043, 0.088)	<.0001	0.0230 (−0.008, 0.069)	0.003 (0.000, 0.008)	.140	0.077 (0.050, 0.103)	0.014 (0.006, 0.025)	<.0001	0.791
	m				0.417 (0.355, 0.491)	0.236 (0.189, 0.290)	<.0001				
d	*x*	0.126 (0.087, 0.175)	0.023 (0.013, 0.034)	<.0001	0.024 (−0.007, 0.056)	0.002 (0.000, 0.004)	.240	0.052 (0.034, 0.072)	0.005 (0.002, 0.008)	<.0001	0.827
	m				0.411 (0.351, 0.467)	0.184 (0.155, 0.216)	<.0001				
e	*x*	0.443 (0.389, 0.492)	0.339 (0.309, 0.363)	<.0001	0.286 (0.159, 0.408)	0.046 (0.016, 0.083)	<.0001	0.277 (0.179, 0.365)	0.046 (0.019, 0.075)	<.0001	0.850
	m				0.626 (0.399, 0.849)	0.141 (0.081, 0.208)	<.0001				

between-subjects experiments, and illustrated it with several applications to previously published data. The applications illustrated the additional insights afforded by the BANOVA analyses. BANOVA is implemented in an easy-to-use (free) R-package that interfaces with the STAN software. Outputs are easy to interpret because of their similarity with those of standard ANOVA, and include a table with effect sizes and *p*-values, a table with parameter estimates, and a table with predictions of the dependent variable for all possible experimental conditions (and their Credible Intervals). Planned comparisons, simple effects, and floodlights, estimates of aggregate and individual-level indirect effects, and Credible Intervals of these estimates are also provided. BANOVA thus provides a one-stop shop for the analysis of experiments in consumer psychology that obviates the need to stack analyses from multiple packages, which is convenient and avoids errors. The underlying STAN code can be printed for more adventurous users to modify and run independently, for models that have more complexity than the ones accommodated in the package, or when informative priors need to be specified. An additional advantage of working within R is that it is easy to load data, and many other functions to process the data or the output of the analyses are available, such as for calculating new variables, merging data sets, limiting a data set to records with full data, or to participants with a certain characteristic and for testing distributional assumptions.

For between-subjects designs, when the Normal distribution holds and/or sample sizes and effect sizes are relatively large, standard ANOVA will yield inferences that are similar to BANOVA. In most other cases when the design involves repeated measures and/or nonNormal dependent variables it may not, and standard floodlight and mediation analyses are only approximate. BANOVA was primarily developed for those cases. Of course, statistical software such as Stata, SAS, SPSS, Mplus, and R, can also provide piecemeal solutions in many of those situations, but the researcher needs to program the models and interpret their results on a case-by-case basis, for which often extensive post-processing of the output is required. For example, to analyze binary data collected in a repeated measures design, a researcher could estimate a hierarchical logistic regression model using one these software packages (for example, with the command *xtmelogit* in Stata, the procedure *glimmix* in SAS, or the function *glmer* in R), using effect-coded variables for the experimental factors and their

interactions. However, the interpretation of the estimates is often cumbersome, significance tests may be unreliable for small samples, the effect sizes are not comparable to those for regression models, and external software is needed to conduct floodlight or mediation analyses. In contrast, BANOVA provides a unified solution, regardless of the experimental design and distribution of the measurements and allows for floodlights, mediation, and moderation as an integral part of the analysis. In addition, the conceptual advantages of using a Bayesian approach to ANOVA were highlighted, and it was advocated that three measures of evidence for a hypothesis should be reported: estimates, *p*-values, effect sizes, and Credible Intervals for these quantities.

The BANOVA approach explicated in the present article thus provides a comprehensive, coherent, and intuitive framework for theory testing in consumer psychology, and its flexibility allows one to account for the key features of most measurements and experimental designs. We hope that this article convinces consumer psychologists to adopt this approach to further improve the quality of their statistical inferences, in a way that requires minimal effort.

ACKNOWLEDGMENTS

The authors are grateful to Laura Graf, Stijn Van Osselaer, Chris Janiszewski, Rik Pieters, and Steven Sweldens for making their data available for this research, to Shehjar Razdan for conducting the content analyses, and to Rik Pieters, Peter Lenk, and to the editor, area editor, and three anonymous reviewers for very useful comments on the manuscript.

REFERENCES

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). New York, NY: Wiley.

Baron, R. M., & Kenny, D. A. (1986). Moderator mediator variables distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400. https://doi.org/10.1207/s15327906mbr4003_5

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*, 142–163. https://doi.org/10.1037/1082-989X.11.2.142

Bernardo, J. M., & Giron, F. J. (1992). Robust sequential prediction from nonrandom samples: The election night forecasting case. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (4th ed., pp. 61–77). Oxford, UK: Oxford University Press.

Bollen, K. A., & Stine, R. A. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*, 550–558.

Carlin, B. P., & Louis, T. A. (1998). *Bayes and empirical Bayes methods for data analysis*. New York, NY: Chapman & Hall/CRC.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1~32.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). New York, NY: Prentice Hall.

Dong, C., & Wedel, M. (2017). BANOVA: An R-Package for hierarchical Bayesian ANOVA. *Journal of Statistical Software*, *81*, 1–46.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242. https://doi.org/10.1037/h0044139

Efron, B. (1978). Controversies in the foundations of statistics. *The American Mathematical Monthly*, *85*, 231–246. https://doi.org/10.1080/00029890.1978.11994566

Enders, C. K., Fairchild, A. J., & MacKinnon, D. P. (2013). A Bayesian approach for estimating mediation effects with missing data. *Multivariate Behavioral Research*, *48*, 340–369. https://doi.org/10.1080/00273171.2013.784862

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. R. (2013). *Bayesian data analysis* (3rdd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/ hierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*, 241–251. https://doi.org/10.1198/004017005000000517

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169–194). Oxford, UK: Clarendon Press.

Gill, J. (2015). *Bayesian methods for the social and behavioral sciences* (3rd ed.). London, UK: Chapman and Hall.

Graf, L. K., Mayer, S., & Landwehr, J. R. (2017). Measuring processing fluency: One versus five items. *Journal of Consumer Psychology*, *28*, 393–411.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: The Guilford Press.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144. https://doi.org/10.1287/opre.31.6.1109

Horner, R. H., & Spaulding, S. (2010) Single-case research designs. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1386–1394). Thousand Oaks, CA: Sage Publications.

Howitt, D., & Cramer, D. (2011). *Introduction to research methods in psychology* (3rd ed.). Harlow, Essex: Pearson Education Limited.

Hutchinson, J. W., Kamakura, W. A., & Lynch, J. G. (2000). Unobserved heterogeneity as an alternative explanation for "reversal" effects in behavioral research. *Journal of Consumer Research*, *27*, 324–344. https://doi.org/10.1086/317588

Jeffreys, H. (1961). *The theory of probability*. Oxford, UK: Oxford University Press.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, *6*, 115–134. https://doi.org/10.1037/1082-989X.6.2.115

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*, 791. https://doi.org/10.1080/01621459.1995.10476572

Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152. https://doi.org/10.1037/a0028086

Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York, NY: Holt, Rinehart andWinston.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*, 249–277. https://doi.org/10.1207/S15327906MBR3602_06

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*, 573–603. https://doi.org/10.1037/a0029146

Lachowicz, M. J., Preacher, K. J., & Kelly, K. (2011). A novel measure of effect size for mediation analysis. *Psychological Methods*, *23*, 244–261.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1-11.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, *58*, 619–678.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs: A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

MacKinnon, D. P. (2013). *Introduction to statistical mediation analysis*. New York, NY: Routledge.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593. https://doi.org/10.1146/annurev.psych.58.110405.085542

Marsman, M., & Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one-sided p-value. *Educational and Psychological Measurement*, *77*, 529–539.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York, NY: Chapman and Hall.

Micceri, T. H. (1989). The Unicorn, the Normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Miočević, M., MacKinnon, D. P., & Levy, R. (2017). Power in Bayesian mediation analysis for small sample research. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 666–683. https://doi.org/10.1080/10705511.2017.1312407

Morales, A. E. (2005). Giving firms an "e" for effort: Consumer responses to high-effort firms. *Journal of Consumer Research*, *31*, 806–812. https://doi.org/10.1086/426615

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist*, *56*, 119–127.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142.

Nuzzo, R. (2014). Scientific Method: Statistical errors. *Nature*, *506*, 150–152. https://doi.org/10.1038/506150a

Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research*, *44*, 692–716. https://doi.org/10.1093/jcr/ucx081

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, *66*, 29–38. https://doi.org/10.1016/j.jesp.2015.09.012

Preacher, K. J., & Kelly, K. (2018). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115.

Preacher, K. J., Rucker, D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227. https://doi.org/10.1080/00273170701341316

Rindskopf, D. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology*, *52*, 179–189. https://doi.org/10.1016/j.jsp.2013.12.003

Rogosa, D. R. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*, 307–321. https://doi.org/10.1037/0033-2909.88.2.307

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Selig, J. P., & Preacher, K. J. (2009). Mediation models for longitudinal data in developmental research. *Research*

in Human Development, *6*, 144–164. https://doi.org/10.1080/15427600902911247

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*, 510–550. https://doi.org/10.1037/a0029312

Spiller, S. A., Fitzsimons, G. J., Lynch, J. G. Jr, & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, *50*, 277–288. https://doi.org/10.1509/jmr.12.0420

Sweldens, S., Van Osselaer, S. M. J., & Janiszewski, C. (2010). Evaluative conditioning procedures and the resilience of conditioned brand attitudes. *Journal of Consumer Research*, *37*, 473–489. https://doi.org/10.1086/653656

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, *14*, 779–804. https://doi.org/10.3758/BF03194105

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wedel, M., & Pieters, F. G. M. (2015). The buffer effect: The role of color when advertising exposures are brief and blurred. *Marketing Science*, *34*, 134–143. https://doi.org/10.1287/mksc.2014.0882

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322.

Zhang, J., Wedel, M., & Pieters, F. G. M. (2009). Sales effects of attention to feature advertisements: A Bayesian mediation analysis. *Journal of Marketing Research*, *46*, 669–681. https://doi.org/10.1509/jmkr.46.5.669

[Corrections added on July 04, 2019, after first online publication: Online supporting information files have been updated with minor corrections to the technical appendix to reflect a change made to the software used to program the package.]

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. Installing the R and the BANOVA package

Appendix S2. The R Commands for the Analyses of the Gist Perception Study (Study 1)

Appendix S3. The R Commands for the Analyses of the Conditioning Study (Study 2)

Appendix S4. The R Commands for the Analyses of the Fluency Study (Study 3)