- Within-subject mediation analysis for experimental data in cognitive psychology and
- neuroscience
- Matti Vuorre<sup>1</sup> & Niall Bolger<sup>1</sup>
- <sup>1</sup> Department of Psychology, Columbia University

Author Note

- 6 Correspondence concerning this article should be addressed to Matti Vuorre, 406
- Schermerhorn Hall, 1190 Amsterdam Avenue MC 5501, New York, NY 10027. E-mail:
- mv2521@columbia.edu

5

Abstract

Statistical mediation allows researchers to investigate potential causal effects of experimental 10 manipulations through intervening variables. It is a powerful tool for assessing the presence 11 and strength of postulated causal mechanisms. Although mediation is used in certain areas 12 of psychology, it is rarely applied in cognitive psychology and neuroscience. One reason for 13 the scarcity of applications is that these areas of psychology commonly employ 14 within-subjects designs, and mediation models for within-subjects data is considerably more 15 complicated than for between-subjects data. Here, we draw attention to the importance and 16 ubiquity of mediational hypotheses in within-subjects designs, and we present a general and 17 flexible software package for conducting Bayesian within-subjects mediation analyses in the 18 R programming environment. We use experimental data from cognitive psychology to illustrate the benefits of within-subject mediation for theory testing and comparison. 20

*Keywords:* mediation; multilevel analysis; repeated measures; bayesian statistics; causal mechanism

Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience

Many important questions in psychology concern a causal chain of relationships

between an initial cause and its effect through an intermediary process. One common 26 method for investigating such causal models, statistical mediation, assesses to what extent 27 the effect of an independent variable (IV) on a dependent variable (DV) is mediated by an 28 intervening variable M. Mediation is suitable for answering many questions about causal processes in cognitive psychology and neuroscience, such as: "Do expectations alter brain activity, and thereby change how individuals respond to stimuli?" (Atlas, Bolger, Lindquist, 31 & Wager, 2010); and "Does cellphone use while driving cause traffic accidents by increasing drivers' attentional demands?" (e.g. Ishigami & Klein, 2009). Mediation models are valuable because they allow the evaluation of theoretical predictions about causal mechanisms, whether in testing a single theory or in theory comparison. Experiments in cognitive psychology, and related areas, often investigate moderational 36 hypotheses, by for example testing interaction effects in multi-way ANOVAs, but 37 investigations of mediational hypotheses (see Baron & Kenny, 1986 for a discussion on the distinction between mediating and moderating variables) in cognitive psychology are rare in comparison to many other branches of psychology (Table 1 in MacKinnon, Fairchild, & Fritz, 2007). One reason for the rarity of mediational hypotheses in this area may relate to the difficulties associated with appropriately analyzing mediation when the data consist of repeated measures over individuals in within-subject designs—as is often the case in cognitive experiments. One general, and increasingly common, strategy for analyzing repeated measures data and within-subject designs is multilevel modeling, which postulates that the data are nested within units, such as trial-level observations nested within individual participants. Over the past decade, multilevel analysis has been successfully implemented to assess mediation where data are repeatedly measured within individuals in different conditions, such as many experiments in cognitive psychology and neuroscience.

Here, we briefly introduce the classic mediation model of Baron and Kenny (1986) and the multilevel modeling approach to estimating mediation with repeated measures (Kenny, Kashy, & Bolger, 1998; Kenny, Korchmaros, & Bolger, 2003). We then introduce an easy to use software package for Bayesian estimation of multilevel mediation models in the R programming environment, and illustrate its use with an example.

### 55 Mediation

In what follows, we discuss a mediation model where X is the hypothesized causal variable, usually the IV in an experiment, Y is the measured outcome (the DV), and M a measured variable which hypothetically mediates X's effect on Y. We restrict our discussion to cases where X and Y are either binary or continuous and M is continuous. We focus on, and implement in the software presented below, this three variable mediation model because of its wide applicability and suitability for data from cognitive psychology and neuroscience experiments. This model is illustrated in Figure 1.

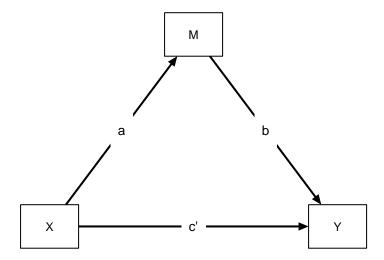


Figure 1. Diagram of the mediation model.

The three causal paths in Figure 1—a, b, and c', corresponding to X's effect on M, M's effect on Y, and X's effect on Y having taken M into account, respectively—correspond to parameters from two regression models, one in which M is the outcome and X the predictor,

and one in which Y is the outcome and X and M the simultaneous predictors. From these parameters, we can compute the mediation effect (the product ab; also known as the indirect effect), and the total effect of X on Y,

Thus, the total causal effect of X, which is captured by the parameter c, can be

$$c = c' + ab. (1)$$

decomposed precisely into two components, a direct effect c' and an indirect effect ab (the 70 product of the a and b paths). There is evidence of mediation when the uncertainty interval 71 (we later define this interval in more detail and distinguish Confidence and Credible Intervals) for ab is sufficiently small that one can rule out zero as a likely population value. There is evidence of complete mediation if the uncertainty interval for the direct component c' is narrow around zero. As in all conclusions from data, assertions of mediation are probabilistic (and we caution users not to interpret the uncertainty intervals as hypothesis 76 tests). Furtheremore, the distinction between complete and partial mediation may not always be useful, and researchers may instead want to focus on the magnitude of the 78 mediation effect (Shrout & Bolger, 2002). 79 To continue with the cellphone and traffic accident example, talking on a cellphone (X) 80 may increase the driver's attentional or cognitive load (M; path a), which in turn may lead to traffic accidents (Y; path b) (e.g. see Ishigami & Klein, 2009). If attentional or cognitive 82 demands completely explained the cellphone use  $\rightarrow$  traffic accidents relationship, then the mediation effect (ab) would be close in size to c, whereas the direct effect (c') would be close to zero. If attentional demands did not completely explain the relationship, c' and its associated uncertainty interval would also allow excluding zero as a plausible value. Viewed in this light, the mediation model consists of a pair of regression models, and inference is performed by interpreting the model's estimated parameters and their transformations (ab, c', c; Baron & Kenny, 1986; Shrout & Bolger, 2002). Importantly, this logic can be extended to multilevel regression models to analyze data with repeated measures (Kenny et al., 2003), 91 a task we turn to next.

114

115

116

## Multilevel mediation model for repeated measures

Multilevel modeling (sometimes called hierarchical modeling, or linear mixed modeling) 93 is a general approach for treating non-independent observations, such as repeated measures 94 within individuals in psychological experiments. The key assumption in a multilevel model is that the lower or trial level observations are nested within upper level units (individual participants), and the general approach consists of estimating regression models where parameters at each level are estimated simultaneously. In educational research, the upper level units can be schools, and the lower level observations can be students within those schools. In cognitive experiments, the upper level units are persons, and the lower-level units 100 are measured repeatedly over trials. Such data structures characterize many—if not 101 most—within-subject experiments in cognitive psychology, where each subject is repeatedly 102 exposed to each level of the treatment variable. For example, in the Stroop task (Stroop, 103 1935), subjects usually observe (multiple instances of) both congruently and incongruently 104 colored letters. 105

Within-subject designs and repeated measures have traditionally been analyzed with 106 methods such as repeated measures AN(C)OVA. However, multilevel models have many 107 benefits over these methods, such as the ability to naturally account for unbalanced data 108 (unequal number of observations across individuals, groups, and conditions), the ability to 109 incorporate continuous and categorical variables, estimation of the variation in effects across 110 individuals, and the extent to which the effects covary in the population of individuals (see 111 e.g. Bolger & Laurenceau, 2013; Gelman & Hill, 2007; Jaeger, 2008; McElreath, 2016). For 112 these and other reasons, multilevel models have grown in popularity at a rapid pace. 113

Importantly, multilevel modeling is also applicable in mediation analysis when the data consist of multiple measurements within individuals, allowing either some or all of the paths to vary between individuals in the study. As in other types of data where measurements are

correlated (within-individuals, for example), ignoring this structure of the data can lead to 117 inaccurate standard errors and thereby lead to over- or underconfidence in one's findings. 118 When each of the X, M, and Y variables are repeatedly measured within individuals, the 119 mediation model is commonly known as  $1 \to 1 \to 1$ , or lower level mediation because each 120 variable is measured at the lowest level (Kenny et al., 2003; level 1, trials, in contrast to level 121 2, the individuals; Krull & MacKinnon, 2001; Preacher, 2015). In the current work, we focus 122 on this model. Other multilevel mediation models include  $2 \to 1 \to 1$  mediation, where the 123 X variable does not vary at the lower level (Preacher, 2015; Raudenbush & Sampson, 1999). 124 and  $2 \to 2 \to 1$  mediation, where only the outcome variable (Y) is repeatedly measured 125 (Krull & MacKinnon, 2001). 126

After the topic was introduced (Kenny et al., 1998), multilevel mediation has been attracted interest both in methodology development and application (Preacher, 2015). It has been successfully applied in  $1 \to 1 \to 1$  models (Kenny et al., 2003), multilevel models with moderation (Bauer, Preacher, & Gil, 2006), but requires care in application, such as considerations of within-cluster centering variables to isolate between- and within-subject effects from each other (Zhang, Zyphur, & Preacher, 2009). Multilevel mediation has also been extended to multilevel structural equation modeling (Preacher, Zyphur, & Zhang, 2010), but SEM approaches are outside of the scope of the current work.

Multilevel mediation equations. In this article, we consider a mediation model 135 applied to data where the independent variable is manipulated within individuals, and the 136 outcome and hypothesized mediating variables are measured on each trial. These data then 137 afford two levels of analysis: At the lower level are trial-level observations, which are clustered within individual persons at the upper level. The following equations refer to (and the software presented below requires) data sets structured in long format. That is, each 140 observation is on a separate row, and each variable has its own column. When an experiment 141 consists of multiple trials, then each row would represent a trial, and the different variables 142 measured (or manipulated) during the trials would be in their own columns. The example 143

162

data are presented in this format in Table 1.

The multilevel mediation equations include both population- and subject-level 145 parameters. The population-level parameters describe the distribution of the parameters in 146 the population, and are usually of key interest in the analysis. The means of these 147 distributions (sometimes known as "fixed effects") describe the effects "for the average 148 person" (Bolger & Laurenceau, 2013), whereas their standard deviations (sometimes known 149 as "random effects") describe the extent to which people differ from one another in the 150 population. The subject-level parameters are specific to the individuals in the current sample 151 of subjects, and are considered random draws from the population-level distribution. We 152 avoid using the terms "fixed" and "random" because they can be confusing, and are less 153 meaningful in the Bayesian context where all parameters are random in some sense (Gelman & Hill, 2007, p. 245; Gelman et al., 2013, p. 383). 155

We denote the subject-level effects with the same letters as the population-level effects but prepend them with u and append with the index variable j to specify that they vary across units of j (the individual subjects; Figure 2)<sup>1</sup>. An important addition to the equations is the covariance of the subject-level  $a_j$  and  $b_j$  parameters, illustrated on top of Figure 2 (and explained in further detail below).

In writing the model, we deviate from the common "error term" representation, where the stochastic component of the model is added separately to indicate that the errors are normally distributed, and instead represent the models as probability distributions

<sup>&</sup>lt;sup>1</sup>We refer to the subject-level parameters as effects, although they more accurately represent deviations of the subject-level parameters from the population-level average effects. We use this nomenclature for two reasons: First, we believe it is more straightforward and avoids an unimportant technicality. Second, although we could have directly written the model so that the subject-level parameters are effects, not deviations from the population-level effects, we found that the MCMC algorithms were more efficient when using the parameterization presented here. Importantly, the subject-specific effects returned by the software (such as those Figure 7) are not deviations from the average effect, but instead have the average effect added to them and can therefore be considered as the subject-specific effects.

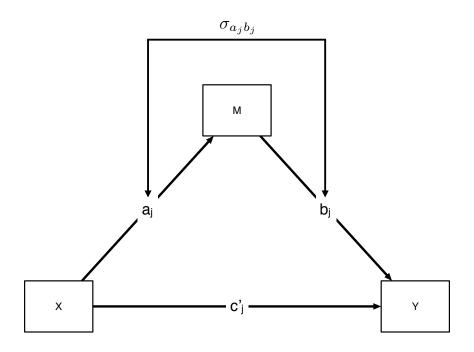


Figure 2. Diagram of the within-subjects mediation model.

166

167

168

169

170

themselves, an approach we think is more natural in the Bayesian context.

Analyzing the causal paths in Figure 2 consists of estimating two multilevel regression equations. For path a (X's effect on M), we model each observation of  $M_{ij}$  (observation in row i for individual j; see Table 1) as a random draw from a Gaussian distribution with mean  $\mu_{M_{ij}}$  and standard deviation  $\sigma_M$ . Note that  $\sigma_M$  is the standard deviation of the lower-level residual. The linear model is then represented as a regression equation for  $\mu_{M_{ij}}$ ,

$$M_{ij} \sim N(\mu_{M_{ij}}, \sigma_M^2) \tag{2}$$

$$\mu_{M_{ij}} = (d_M + u_{dM_j}) + (a + u_{a_j})X_{ij}. \tag{3}$$

The first term in equation 3,  $d_M$ , is the population-level intercept for M, and  $u_{dM_j}$  is
the subject-level intercept for subject j (see footnote 1). The M  $\rightarrow$  Y and X  $\rightarrow$  Y slopes
(paths b and c') are captured by modeling  $Y_{ij}$  (observation of Y in row i for individual j; see
Table 1) as random draws from a Gaussian distribution with mean  $\mu_{Y_{ij}}$  and standard

deviation  $\sigma_Y$  (which, again, is the lower-level residual of Y).

$$Y_{ij} \sim N(\mu_{Y_{ij}}, \sigma_Y^2) \tag{4}$$

$$\mu_{Y_{ij}} = (d_Y + u_{dY_j}) + (c' + u_{c'_i})X_{ij} + (b + u_{b_j})M_{ij}.$$
(5)

This regression predicts Y from the combination of population-level and subject-level intercepts  $d_Y$  and  $u_{dY_j}$ , respectively; population and subject-level direct effects of X on Y (c' and  $u_{c'_j}$ ); and population- and subject-level effects of M on Y (b and  $u_{b_j}$ ).

The multilevel nature of this model is captured by specifying the subject-level

parameters as draws from a multivariate normal distribution with a 5 x 1 vector of means of zero and a 5 x 5 covariance matrix  $\Sigma$ ,

$$\begin{bmatrix} u_{dM_j} \\ u_{dY_j} \\ u_{a_j} \\ u_{b_j} \\ u_{c'_i} \end{bmatrix} \sim N(0, \Sigma). \tag{6}$$

Together, equations 2-6 constitute the multilevel mediation model. From the model's 183 estimated parameters, we can calculate for each individual, and the population average, all 184 the additional parameters that are used to assess mediation, such as me (population-level 185 mediation effect; also known as indirect effect) or  $u_{me_3}$  (mediation effect for person 3). 186 However, calculating the mediation effect, and therefore the total effect of X on Y (c) for the 187 multilevel model differs in important ways from the single-level (between-subject) 188 calculations (eqn. 1). To obtain the population-level mediation effect, we must add the 189 covariance of  $a_j$  and  $b_j$  to the product of the population-level a and b (as shown by Kenny et 190 al., 2003, eqn. 9): 191

$$me = ab + \sigma_{a_ib_i}. (7)$$

 $\sigma_{a_jb_j}$ , the covariance of  $a_j$  and  $b_j$ , is an element of the covariance matrix  $\Sigma$ , and 192 indicates the degree to which subjects with (say) greater values of  $a_i$  are likely to have 193 greater values of  $b_j$  (in the case where the covariance is positive). To fight et al. noted that 194 this covariance term can indicate an omitted variable that interacts with the a and b slopes, 195 and therefore including it in the model leads to a more general model that allows 196 misspecification of the model at the between subject level (Tofighi, West, & MacKinnon, 197 2013). These authors suggested that researchers estimate the model both with and without 198 this covariance term (Tofighi et al., 2013, p. 301), but in our view it is more straightforward 199 to allow this parameter to be estimated from the data. Note, however, that researchers can 200 effectively force this parameter to be zero by specifying a prior on it that spikes at zero (see 201 below), but we don't recommend this approach unless there is abundant prior information to 202 suggest such a model. This covariance is illustrated in Figure 2, with a double-headed arrow connecting  $a_i$  and  $b_i$ .

Finally, the population-level total effect of X on Y is given by

$$c = me + c'. (8)$$

The software also allows estimating the same multilevel mediation model, but for a binary Y variable (coded as 0s and 1s.) In this case, the model for Y (equations 4 & 5) is a multilevel logistic regression:

$$Y_{ij} \sim Bernoulli(\mu_{Y_{ij}})$$
 (9)

209

205

$$\mu_{Y_{ij}} = \frac{1}{1 + exp(-\eta_{Y_{ij}})} \tag{10}$$

210

211

$$\eta_{Y_{ij}} = (d_Y + u_{dY_j}) + (c' + u_{c'_i})X_{ij} + (b + u_{b_j})M_{ij}. \tag{11}$$

The Bernoulli distribution in equation 9 is the Binomial distribution for a single trial.

### 12 Alternatives to multilevel models

Although multilevel modeling is not the only approach to within-subject mediation, we believe that in the types of experiments most commonly employed in cognitive psychology and neuroscience, it is the most parsimonious and applicable mediation model.

A common alternative to multilevel modeling in within-subject mediation is 216 longitudinal modeling, where change processes within individuals are modelled to occur over 217 time (Cheong, MacKinnon, & Khoo, 2003; Cole & Maxwell, 2003; MacKinnon, 2008; Selig & 218 Preacher, 2009). While a tremendously valuable approach in many areas of psychology, 219 longitudinal models are less relevant in cognitive psychology and neuroscience, because 220 experiments in these fields rarely track participants over time. Instead, experimental 221 conditions are usually randomised trial-wise (or block-wise), and the experiments usually are 222 less than an hour or two in duration, meaning that any possible change within individuals 223 occurs repeatedly over the course of the experiment (when conditions change across trials) 224 with no meaningful purely temporal pattern with respect to the causal effect. 225

Another method of addressing mediation in within-subject designs focuses on using
change scores between experimental conditions (Judd, Kenny, & McClelland, 2001; Montoya
& Hayes, 2017), a valuable approach for experiments where participants provide only two
measures. This method was recently advanced to include designs with control groups
(Valente & MacKinnon, 2017). While useful for e.g. pre-post-test designs, this method does
not easily address designs where the manipulated variable (X) continuous, or where
participants are measured more than twice.

While useful in many situations, these other approaches to within-subject mediation
are less practical and applicable than multilevel models in the context of cognitive
psychology and neuroscience experiments. Therefore, we have decided to focus and
implement a multilevel modeling approach to within-subject mediation.

## 237 Bayesian estimation

249

250

251

253

254

Traditional procedures of estimating (within-subject) mediation models have focused 238 on various frequentist methods, such as Ordinary Least Squares (OLS) and Maximum 239 Likelihood Estimation (MLE). We instead advocate—and implement in the software package discussed below—Bayesian estimation, because it offers several advantages over these conventional methods. The benefits include the natural incorporation of uncertainty in 242 estimated parameters in the form of a posterior distribution, and the probability 243 interpretation afforded by it; the ability to incorporate prior information in the statistical 244 model; and a more natural interpretation of multilevel models. We discuss these benefits 245 below, then highlight some similarities of the Bayesian method to classical procedures before 246 briefly introducing the precise method by which the Bayesian estimation is conducted in our 247 programming package. 248

Traditional MLE methods for assessing (multilevel) regression models, such as those described above, provide point estimates and standard errors of the estimated parameters, which are in turn based on assumptions about the parameters' sampling distributions, and from which Confidence Intervals can be calculated. In contrast, Bayesian analyses provide, for each parameter, full posterior probability distributions of plausible parameter values, and therefore directly interpretable representations of uncertainty (Kruschke, 2014).

This fact is important when the investigation focuses on transformations of the
estimated parameters at multiple levels, such as equations 7 and 8, because the uncertainty
in the estimated parameters is conveniently carried forward to uncertainty in the
transformations of the estimated parameters, such as c, me and pme (at both levels) in the
multilevel mediation model. The posterior distributions can then be summarized by X%
Credible Intervals or displayed visually to effectively communicate the relative plausibilities
of various (transformed) parameter values. We believe the visual inspection of histograms
and violin plots (Figures 5 and 6) can benefit inference by helping users focus on
distributions of plausible values, instead of point estimates. Visual inspection of the

distribution of plausible values is especially important when the underlying distribution may
be non-Gaussian, such as for indirect effects (or *pme*) in mediation, or when sample sizes are
small.

Further, unlike the standard error given by MLE methods, the Bayesian posterior 267 distribution is a probability distribution and allows statements of relative probabilities of 268 parameter values. For instance, we often wish to discuss the plausibility of various parameter 269 values, and our subjective confidence in these values. The posterior distribution obtained by 270 a Bayesian analysis allows just that: Throughout this article we refer to "X\%" most plausible 271 values", and "credible" parameter values. These statements can be made based on 272 summaries of the posterior distribution, such as the Credible Interval, which contains some 273 percentage of the central values of the distribution. Confidence Intervals based on more 274 common frequentist estimation methods (such as MLE) do not allow such statements of 275 plausibility or subjective confidence, although they are sometimes (wrongly) so interpreted 276 (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015).

The Bayesian framework also allows incorporating prior information in the statistical model. When prior information on the magnitudes of parameters is available, it can be naturally incorporated in the analysis, thus improving the estimates. This information can come from expertise in the field of study, earlier studies, or knowledge about the natural constraints in the data. For example, researchers are sometimes aware of limits beyond which parameters are unlikely to be found; this information can be incorporated in the form of a prior distribution which will decrease the variance of the estimate. Information used in this manner is sometimes called a "regularizing" prior, and can be very useful especially in contexts when the data are uninformative about the parameter (McElreath, 2016).

A related benefit of incorporating prior information into a statistical model relates to
the stability of the estimated parameters. A well-known problem with MLE methods in the
context of multilevel models is that point estimates of the between-person heterogeneity
parameters (often denoted  $\tau$ , see equation 6) cannot be distinguished from zero, even though

the parameter's likelihood function contains a considerable range of non-zero values. A 291 consequence of this is that with MLE the person-level parameters conditional on the zero 292 heterogeneity would be erroneously estimated as identical. This situation occurs especially in 293 applying generalized multilevel models, such as logistic regression. However, the Bayesian 294 analysis provides a distribution of values for the heterogeneity parameter, which 295 consequently is not "stuck" at zero and thus allows the subject-level parameters to vary. In 296 these situations, the data can be relatively uninformative about the actual value of the 297 parameter, and the posterior distribution may be unnecessarily wide. 298

In addition, when this happens, the Bayesian analysis allows including prior
information in the form of a heavy tailed distribution, such as Cauchy with appropriate
hyperparameters (Gelman, 2006). This information can then effectively regularize the
inference toward more realistic values, and thereby allow estimating the between-person
heterogeneity parameter even in situations where MLE methods fail and the data are
relatively uninformative about the underlying parameter values.

Although the topic is beyond the scope of this article, prior information allows

Bayesian hypothesis testing in a straightforward manner using Bayes Factors. Bayes Factors

can be thought of as quantifying the "extent to which data cause revision in belief" (Kass & Raftery, 1995; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016, p. 533) where

"belief" is the prior probability distribution. Furthermore, Bayes Factors can be fairly easily

estimated using MCMC samples (see below) using the Savage-Dickey density ratio method;

details are given by Wagenmakers et al. (2010).

Multilevel mediation is both conceptually and computationally significantly more
complicated than single-level mediation. One conceptual difficulty is in how the parameters
are interpreted at various levels of analysis: Classical methods consider the upper-level
parameters as "fixed", and the lower-level parameters as "random". This may not be a
fundamental problem, but at least presents difficulties in teaching and communication (see
footnote 2 in Gelman & Hill, 2007, p. 245 for an exposition of the problem; see also Yuan &

MacKinnon, 2009, p. 312). In the Bayesian context, in contrast, all estimated
quantities—irrespective of their level in the model's hierarchy—are considered random. In
the context of the multilevel mediation model, the upper- or population-level parameters can
then be considered as empirically informed priors for the lower- or subject-level parameters
(equation 6; Gelman & Hill, 2007).

On the computational side, conventional methods are often more difficult to apply to complex models, such as multilevel mediation (Kenny et al., 2003; Yuan & MacKinnon, 2009). On the other hand, Bayesian methods allow for a relatively straightforward estimation, especially when it is implemented with efficient MCMC methods (as discussed below). Although out of the scope of this manuscript, the Bayesian method will easily fit more complicated models with large numbers of covariates, levels of analysis, and parameter transformations—even in situations where more traditional MLE methods may fail or be too difficult to implement in practice.

Similarities of Frequentist and Bayesian Methods. Having detailed some
benefits of a Bayesian approach to estimating the multilevel model, is is also important to be
aware of some important similarities between results obtained with Bayesian and more
traditional MLE based methods. For one, when the sample size is very large, and the
sampling error correspondingly small, the point estimate of a parameter can be considered a
sufficient description of the posterior distribution (Gelman et al., 2013). However, in practice
sample sizes are rarely that large.

A more important similarity between classical and Bayesian methods is that if no prior information is included ( $\theta \sim U(-\infty, \infty)$ ), and the model estimation presents no problems, the obtained intervals often have identical bounds. This fact has led some authors to suggest that a classical confidence interval can sometimes be given a Bayesian interpretation (Gelman et al., 2013, sec. 4.5). Notice, however, that we would rarely want to give a Bayesian interval a frequentist interpretation.

Furthermore, under similar assumptions as given above, the frequentist one-sided

p-value corresponds to Bayesian posterior probabilities (that a parameter is greater or
 smaller than a comparison value, such as zero.) (Marsman & Wagenmakers, 2016). However,
 the probability interpretation naturally afforded to the Bayesian quantity seems to us to
 suggest its conceptual superiority, at least insofar as its interpretation does not immediately
 invite a binary significant-or-not attitude (Gelman et al., 2013, p. 95). As a side note, the
 two-sided p-value does not have a straightforward Bayesian counterpart.

We are not the first to suggest the use of Bayesian methods in mediation analyses: 351 Yuan and MacKinnon (Yuan & MacKinnon, 2009) discussed using it for single- and 352 multi-level mediation analyses, and provided copy-and-paste WinBUGS code for conducting 353 the analyses. Another paper discussed the use and benefits of Bayesian methods in the 354 specific context of moderated mediation (Wang & Preacher, 2015). To further these efforts, 355 we provide a fully functional software package for conducting Bayesian multi-level mediation 356 analyses in a common and free programming environment, using state of the art Bayesian 357 estimation procedures, which we turn to next. 358

MCMC and Stan. The computer program we provide and discuss below uses

Markov Chain Monte Carlo (MCMC) procedures as implemented in the Stan<sup>2</sup> modeling

software (Stan Development Team, 2016b) to fit the multilevel mediation model. MCMC is a

class of computational procedures that allow approximating a probability distribution by

drawing random samples from it (for an excellent introduction to MCMC, see van

Ravenzwaaij, Cassey, & Brown, 2016). This technique is extremely valuable for Bayesian

inference, because complex multidimensional posterior distributions are often difficult or

impossible to obtain with analytic calculations.

Stan's effective MCMC algorithms (No-U-Turn Sampler, Hamiltonian Monte Carlo;
Hoffman & Gelman, 2014) are well suited for the problems presented in multilevel mediation
models, such as large numbers (potentially hundreds or thousands) of parameters at multiple
levels. Unlike early popular MCMC algorithms that relied on Gibbs sampling (BUGS,

<sup>&</sup>lt;sup>2</sup>http://mc-stan.org/.

JAGS), Stan's Hamiltonian Monte Carlo is very effective even when the posterior
distributions are highly correlated, making it especially useful for path analyses—such as
mediation—and other more complex structural models.

Furthermore, the Stan language allows placing priors on correlation matrices and standard deviations, instead of (co)variances, making (arguably) the prior specification easier, as discussed next.

Prior distributions on parameters. For Bayesian analysis, all population-level
parameters must also be assigned prior distributions that represent the analyst's state of
knowledge and uncertainty, before seeing the data. The priors depend on the specifics of the
data and context, and should be chosen by the researcher—although the priors often have
little influence on the posterior distribution as the amount of data increases. The software
package we introduce below has default values for the priors that we believe are reasonable,
minimally informative priors in most contexts.

For the regression parameters, the prior distributions are zero-centered Gaussians, with 384 user-defined standard deviations (defaults to 1000). Effectively, a zero-centered Gaussian 385 with a small standard deviation "regularizes" (makes large positive or negative parameter 386 values less plausible, a priori) the estimated parameters, thereby improving inference on 387 average by preventing over-fitting (McElreath, 2016). A more technical definition of what 388 constitutes a "small" standard deviation depends on the theoretical context and 389 measurement scale, but for large amounts of data, the values must be very small indeed to 390 make a meaningful difference in the posterior. The default value of 1000 will have practically 391 no impact on inference for data sets where effects are on the range of z-scores. If greater 392 effects are plausible (say, a manipulation has an effect on the order of thousands of milliseconds), users may increase the value. The  $\sigma = 1000$  we have placed on the population-level regression parameters considers values further away from zero as increasingly unlikely, such that 95% of the a priori most plausible values are between -1960 and 1960.

The second class of prior distributions relates to the variances and covariances of the

subject-level effects. To allow placing priors directly on standard deviations and correlations, 398 we construct the covariance matrix  $\Sigma$  from a vector of standard deviations  $\tau$  and a 399 correlation matrix  $\Omega$  (Stan Development Team, 2016b). We use folded Cauchy distributions 400 with user-defined scale parameters (defaults to 50) for the subject-level effects' standard 401 deviations (Gelman, 2006; Gelman & Hill, 2007). The Cauchy distribution is recommended 402 for these parameters over alternatives, such as inverse-gamma or uniform distributions, 403 especially when the number of clusters (subjects) is very small (Gelman, 2006). Specifically, 404 the hyperparameters of inverse-gamma prior distributions may be more difficult to specify 405 when minimally or non-informative priors are desired, and uniform prior distributions may 406 lead to overestimation of  $\tau$ . The folded (positive-only) Cauchy distributions concern the 407 variability of the effects between subjects: The default scale = 50 implies that increasingly 408 large values of variation (standard deviation of their respective Gaussian distributions) between subjects are increasingly unlikely, such that a priori 50% of the most plausible 410 values are under 50, and 95% are under 1272.

For the correlation matrix  $\Omega$ , we use an LKJ prior with a user-defined shape parameter  $\nu$  (defaults to 1) (Lewandowski, Kurowicka, & Joe, 2009; Stan Development Team, 2016b). With older MCMC sampling programs relying on Gibbs sampling, such as BUGS and JAGS, it was more convenient to use conjugate inverse Wishart distributions as priors on the covariance matrices. However, Stan doesn't require conjugacy for multivariate priors, and it is often easier to think of plausible correlations rather than covariances, and we therefore use the  $\nu$  the  $\nu$  prior distribution on  $\nu$ .

The default hyperparameter value for the LKJ prior ( $\nu = 1$ ) assigns equal plausibility across the range of possible values (-1 to 1), and values of  $\nu$  greater than 1 increase the a priori skepticism of large correlations (McElreath, 2016, p. 393). Because this distribution is relatively unknown and difficult to conceptualize (it is a distribution of matrices), Figure 3 shows four sets of random draws from LKJ distributions with different values of  $\nu$  (McElreath, 2016).

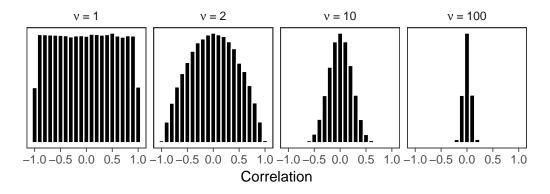


Figure 3. Histograms of 100000 random draws from the LKJ prior distribution for four values of  $\nu$ .

For general information on the choice of prior distributions, we refer the readers to 425 excellent textbooks on Bayesian statistics (Gelman, 2006; Gelman et al., 2013; Kruschke, 426 2014; McElreath, 2016). However, if users wish to estimate models without prior information, 427 they can specify very large standard deviations to the Gaussian prior distributions and large 428 scale parameters to the subject-level effects' standard deviations (Kruschke, 2014). Overall, we chose default values for the prior distributions which would have minimal impact on the resulting posterior distributions, given common ranges of data values and effect sizes. The 431 default priors are easy to change by simply passing named arguments to the estimation 432 function (as detailed below). 433

### Software package for Bayesian multilevel mediation: bmlm

We developed a free open-source software package ("bmlm" for Bayesian Multi-Level
Mediation) for the R programming language (R Core Team, 2016) for easy estimation,
summarizing, and plotting the results of the multilevel mediation model presented above
(Vuorre, 2016). The software can be installed from within the R environment (source code,
detailed installation and use instructions are provided online at the package's website
https://github.com/mvuorre/bmlm). To install the software package, please ensure that you
have the latest versions of R and Xcode with command line tools (OS X users) or Rtools
(Windows users). Then run the following command in the R console (the installation process

may take a few minutes, because the models are compiled to C++ during installation):

# install.packages("bmlm")

After the package has been installed, it must be loaded to the current R workspace to make the functions contained in it available to the user.

## library(bmlm)

Next, we illustrate the functionality and use of **bmlm** with an empirical example.

# Example: Judgments of performance in a video game task

In a series of experiments, Metcalfe and colleagues have examined the informational 448 bases of people's judgments of control; that is, to what extent various experimental 449 manipulations in a computer game task influence people's experiences of control (Metcalfe & 450 Greene, 2007). In these experiments, participants play an arcade style computer game, in 451 which they use the computer mouse to move a game cursor (a light square) horizontally on 452 the bottom of the screen, while Xs and Os fall from top to bottom of the screen. The 453 objective of the game is to catch as many falling Xs as possible, while avoiding all the falling 454 Os. After each game trial (about 20 seconds), the participants provide a judgment of their 455 experienced control using an analog slider scale. These ratings have been found to be highly 456 sensitive to various manipulations of the game, such as artificially introduced spatial and 457 temporal discrepancies between the mouse and square movements, and the speed of the 458 falling Xs and Os (Metcalfe, Eich, & Castel, 2010; Vuorre & Metcalfe, 2016). Additionally, 459 the participants have provided Judgments of Performance (JoP), their subjective evaluations of how well they did in the game on each trial, by using an analog slider scale. Here, we focus on how these JoPs are influenced by a specific experimental manipulation in this computer game task. 463

One experiment introduced, on some trials, a small temporal lag (250ms) between the participant's mouse movements, and the movements of the game cursor on the screen

(Metcalfe et al., 2010, Experiment 1). This manipulation led to a reliable decrement in the 466 players' ratings of performance. In the analysis below we ask: "How does the temporal lag 467 between one's mouse movements, and the movements of the game cursor, decrease ratings of 468 performance?" To answer this question, we propose a straightforward mediational 469 explanation: The temporal lag decreases performance (as measured by hit rate, the 470 percentage of Xs caught in a trial), and people's Judgments of Performance depend on their 471 hit rates. In other words, we expect that hit rate (performance) completely mediates 472 temporal lag's effect on judgments of performance. This hypothesis implies that people are 473 making metacognitively accurate judgments of performance, by basing their performance 474 judgments on an actual performance signal (hit rate), rather than, say, a general feeling of 475 abnormally delayed mouse control. 476

#### 477 Data set

Multilevel models assume that the observed variables have at least two potential levels 478 of variation. Because temporal lag was experimentally manipulated within subjects, it does 479 not vary between subjects. On the other hand, hit rates (HR) vary both between and within 480 participants: At the lower (within-person) level, HR varies from trial to trial. At the upper 481 level, we may also expect that HR varies, on average, between participants. We are most 482 interested in the within-person process, and therefore it is useful to transform the variables 483 such that these two levels are explicitly separated from each other (Bolger & Laurenceau, 484 2013). Notice that this transformation is not strictly required, but this reasoning suggests 485 that it is often useful and meaningful in data sets where the predictor values vary both between and within subjects. We first averaged the grand-mean-centered trial-level HR for 487 each person to create a between-person component of HR. We subtracted these means from the raw HR to create within-subject trial-by-trial deviations from the subject-means that 489 represent an entirely within-person version of HR. Isolating the within-person process from 490 variables can be done by using **bmlm**'s **isolate()** function: 491

Table 1

First six rows of the example data set.

subj	lag	hr	jop	hr_cw	
1	1	36.7	30	-23.3	
1	1	50.0	80	-10.0	
1	0	56.7	90	-3.3	
1	1	56.7	32	-3.3	
1	1	56.7	50	-3.3	
1	0	70.0	76	10.0	

Note. Variables with "\_cw" are centered within-person.

The isolate() function takes three arguments. On the first row, we specify the data set to be MEC2010, which contains data described in (Metcalfe et al., 2010) and is included with the **bmlm** package. The next line specifies the column containing values to isolate the within- and between person processes by (the subject numbers). Finally, the third line identifies the variable to be isolated. After this transformation, the example data frame is ready, and can be seen in Table 1.

Table 1 illustrates the structure and variables of the example data set. Each
participant (43 individuals) is assigned a unique id number (subj); the two experimental
conditions are represented by a dichotomous indicator variable, where 1 indicates a lag trial
(lag); hr is the raw percent of Xs caught in a trial; and jop is the judgment of performance

from low [1] to high [100]). Finally, hr\_cw is the isolated within-subject component of hit
rate. All the variables must be numeric; if the experiment contained two conditions, as in
the example here, the conditions would need to be dummy coded with integers. The data set
contains 8 observations per individual, 4 in the lag condition, and 4 in a control condition.
Eight observations per person may seem a prohibitively small sample, but because the
multilevel model pools uncertainty across subjects, we are able to estimate the
within-subject causal process with these data, as shown below.

## Estimating the multilevel mediation model with bmlm

To estimate the multilevel mediation model with bmlm, users need to specify the data (an R data frame) in the current R environment, and variables within the data frame identifying individuals, and the X, M, and Y variables. Here, MEC2010 is our data frame, subj the column identifying individuals, and lag, hr\_cw and jop the X, M, and Y variables, respectively. These variables are entered into a call to the mlm() function, which estimates the model using Stan's MCMC algorithms (Stan Development Team, 2016b).

This function has two other important features, the control of prior distributions and various controls of the underlying Stan MCMC procedures. As is usual in R, more information can be found by entering ?mlm in the R console. Although the software package sets default priors, users may also input arguments identifying the prior parameters they would like to change (see the documentation included with the software package, or the

package's website, for details). Users may also control the behavior of the underlying MCMC sampler; here, to ensure stable results we increased the number of iterations from the default of 2,000 to 10,000, and ran the program simultaneously on 4 CPU cores.

Depending on the size of the data set and your computer, the MCMC sampling may take from a few seconds to several minutes. By default, mlm() runs 4 MCMC chains, and 525 uses the first half of each chain for warmup (Stan Development Team, 2016b). The iter 526 argument specifies the total number of iterations per chain, so this example results in 20,000 527 samples (4 x 10,000 / 2) from the model's posterior distribution. During and after sampling 528 from the posterior distribution, Stan will print progress information in the R console. 529 Occasionally, these prints may include warnings about abnormal parameter values, but these 530 are usually not a cause for worry but simply a part of the random MCMC sampling 531 procedure. After the procedure ends, and the model has been estimated (the desired number 532 of posterior samples have been obtained), the estimated parameters can be summarized 533 using **bmlm**'s functions. 534

# Summarizing the multilevel mediation model

Population-level estimates. We first focus on the population-level parameters of
the multilevel model. These estimates describe the results of the mediation analysis for the
average person, and are often the parameters of greatest interest. Users may print the
model's focal estimated parameters directly to the R console by running mlm\_summary(fit),
where fit is the R object containing the estimated model:

The output in Table 2 consists of the main population-level parameters of the mediation model. The names correspond to the parameters introduced in Figures 1 and 2, and equations 2-8 (cp is c', pme is proportion of effect that is mediated [see below]). For each parameter, the output shows the posterior mean, standard deviation (abbreviated SE for standard error), and median (which may be a more representative point estimate for skewed posterior distributions.) "2.5%" and "97.5%" are the lower and upper limits of a 95%

Table 2
Summary of results.

Parameter	Mean	SE	Median	2.5%	97.5%	n_eff	Rhat
a	-35.53	1.27	-35.53	-38.00	-33.01	20,000	1.00
b	0.95	0.07	0.95	0.80	1.09	20,000	1.00
cp	-0.49	2.85	-0.49	-6.09	5.16	20,000	1.00
me	-33.70	2.88	-33.67	-39.50	-28.11	20,000	1.00
c	-34.19	2.30	-34.20	-38.70	-29.71	20,000	1.00
pme	0.99	0.08	0.99	0.83	1.16	20,000	1.00

Note. SE (for Standard Error) is the posterior standard deviation.

Credible Interval, which is the central 95% of the corresponding distribution. The limits of 547 the CI can be controlled by specifying the level argument of this function, e.g. an 80% CI 548 would be obtained with level = .8.  $n\_{\it eff}$  indicates the number of effective posterior 549 samples, taking into account the MCMC chains' autocorrelation; this value should be large 550 to allow confident estimates of the quantities. Finally, Rhat is the potential scale reduction 551 factor, and should be 1.00 for accurate estimates of the posterior distribution (Gelman et al., 552 2013, pp. 285–288). If  $n_{eff}$  is too small, or Rhat not 1.00, simply increase the number of MCMC iterations and re-estimate the model using mlm(). We recommend to increase the number of iterations until Rhat is within .05 of 1, and although Gelman et al. (2013) suggest that  $n\_eff$  greater than 10 or 100 is acceptable, in practice when extreme quantiles (such as 95% CIs) are of interest, we recommend increasing iterations until  $n_eff > 100$  (at least). 557 First, we interpret the total effect of temporal lag on judgments of performance (c, 558 equation 8). 95% of the most plausible values of this parameter lie in the interval between 559 -39 and -30, and the mean value of the posterior distribution is -34. Therefore, people gave 560 about 34 points lower ratings of performance (on a scale from 1 to 100) in the lag condition 561

Table 3
Standard deviations of the regression parameters, and their covariance (and correlation).

Parameter	Mean	SE	Median	2.5%	97.5%	n_eff	Rhat
tau_a	1.61	1.26	1.34	0.06	4.74	6,877	1.00
tau_b	0.18	0.08	0.19	0.02	0.33	3,197	1.00
tau_cp	4.37	2.90	4.01	0.21	10.62	5,441	1.00
covab	-0.07	0.17	-0.02	-0.52	0.18	10,288	1.00
corrab	-0.15	0.40	-0.18	-0.81	0.66	20,000	1.00

Note.

576

versus the control condition, with 95% of the most plausible values ranging from -39 to -30. Our mediation hypothesis was that this effect would be mediated by hit rate. Therefore, we next focus on the magnitude of the mediation effect, the me parameter (equation 7). In support of our conjecture, me appears very strong, and of approximately equal magnitude to the total effect c. 95% of most plausible values of me lie between -39 and -28, and the mean value is -34. Further, after taking the hit rates into account, the direct effect of lag is approximately zero and has a narrow credibility interval (cp = -0.49, 95% CI [-6.09, 5.16]), indicating that the lag  $\rightarrow$  JoP relationship is completely mediated by hit rate.

Multilevel models also naturally estimate the between-subject variability around the population-level estimates, and the covariance of the subject-level parameters (i.e. the so-called random effects). The variability is captured in the standard deviations of the subject-level effects (see eqn. 6 and **Prior distributions on parameters** above). These estimates are useful summaries of the heterogeneity and covariance of effects, and can be obtained from the model by calling mlm\_summary(fit, pars = "random")):

For instance,  $tau\_a$  in Table 3 is the estimated standard deviation of the lag  $\rightarrow$  hit

rate relationship in the population. Because posterior distributions of standard deviations tend to be non-normal, instead of a point estimate (the posterior mean or median) we focus on the 95% CI: 95% of the most plausible values of  $tau_a$  are between 0.06 and 4.74. In context of an a effect of -35, the between-subject variation of this effect is very small. To obtain summaries of other parameters in the model, such as subject-specific effects, users need to enter the names of the parameters to the pars argument of mlm summary().

# Visualizing the estimated parameters

bmlm offers quick access to summary plots from estimated models. To draw a path diagram with the variable names, and estimated path parameters as means and X% CIs (default 95%), use mlm\_path\_plot().

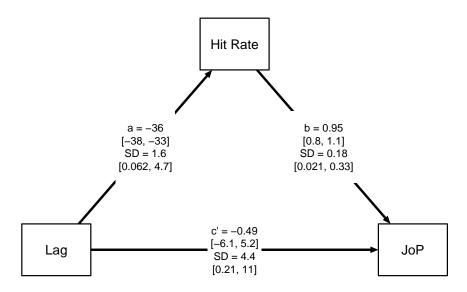


Figure 4. Path diagram with point estimates (posterior means) of the parameters and associated 95 percent Credible Intervals (in square brackets below the point estimates). Under each estimated average effect, "SD" shows the associated effect's standard deviation, which indicates the degree to which that effect varies between people (in standard deviation units).

Second, although the path plot affords a rapid visual display of the main conclusions of
the model, it is often more informative to plot the parameters themselves. **bmlm** offers
three default plots of the parameters, illustrated below. To access these figures, use the
mlm\_pars\_plot() function. This function draws histograms, violin plots, or point estimates
with CIs, of the estimated parameters. The type of the plot can be specified by setting the
type = X argument to this function call, where X is either "hist" (Figure 5), "violin"
(Figure 6), or "coef" (Figure 7).

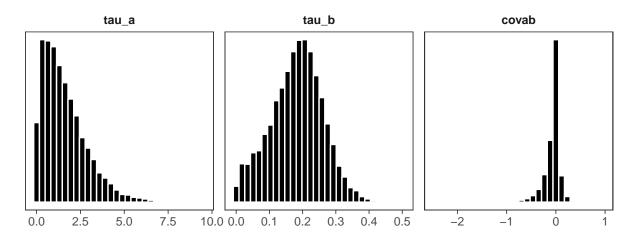


Figure 5. Histograms of the marginal posterior distributions of the standard deviations of a and b (left, middle), and the a-b covariance (right). Quick visual inspection tells us that these distributions are unlikely to be Gaussian, and therefore using the distributions' means or standard deviations as numerical summaries may be misleading. It is better to interpret the entire distribution.

We prefer displaying the main parameters of interest as "violin" plots (also known as "cat's eye" plots, Figure 6). These offer a view of the distributions such that the width of the

shape is proportional to the frequency of those values. In other words, the "violins" are filled 596 density curved turned sideways and mirrored. The violin shapes in Figure 6 illustrate that 597 the most plausible values of cp (population-level direct effect of lag on JoPs), for example, 598 are found near zero, and offer a visual depiction of the relative credibility (width of violin) of 599 the parameter values (y axis). The extremely thin tails of cp beyond about  $\pm 5$  indicate that 600 these values are implausible. The code snippet below also illustrates that the object returned 601 by mlm pars plot() is a ggplot2 object, and can be further customized by functions in the 602 ggplot2 R package (Wickham, 2016). Here, we specify the y axis breaks to run from -50 to 603 10 in increments of five.

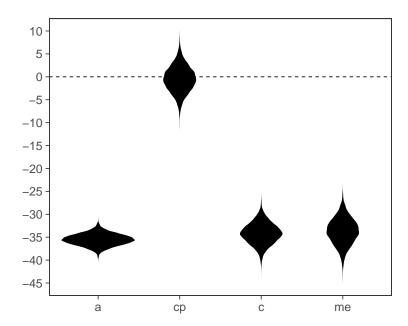


Figure 6. Violin plots of the estimated parameters. Each dark violin shape is a mirrored top-down view of a density plot.

Unlike more familiar point-and-error-bar representations of uncertainty, violin plots put

visual emphasis on the relative plausibility of values within the distribution itself, and may
therefore allow a more efficient display of information.

Subject-level estimates. The multilevel model provides, for each person, their own mediation model with empirical Bayes estimates of the parameters. Numerical representations of these values would quickly overwhelm us, but a graphical representation of the subject-level parameter values offers valuable insight about the between-subject variability in the estimated effects. For example, the subject-specific values of me show relatively little variation, and indicate that the mediation effect is present for each individual person. To obtain subject-specific parameters, simply call the function with the parameter name prepended with "u\_". (We again specify the y axis breaks as in the Figure 6.)

Figure 7 displays the subject-level (black) and population-level (red) estimates of *me*, the mediation effect, and shows that while there is some variation in the subject-specific effects, there appears to be a strong mediation effect (indirect effect) for every person.

## 619 Mediation with binary outcomes

Binary outcomes are common in cognitive psychology, such as in learning and memory
experiments where the Y variable may be a binary indicator for a correct / incorrect or
remembered / not remembered response. **bmlm** allows estimating the multilevel mediation
model with binary outcomes, and assumes that the outcome variable is coded as 0s and 1s.

To illustrate how to estimate the model with a binary outcome variable, we created a
binary (0/1) outcome variable in the current example data by within-person-median splitting

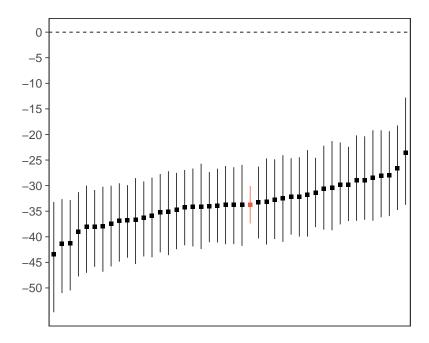


Figure 7. Coefficient plot of the person-level estimates of the mediation effect. Each square represents the mean of an individual's estimated ab parameter, and the lines cover the 80% CI of the parameters. By adjusting the "pars" argument, we also include the average level estimate, which is automatically displayed in red.

the original outcome variable (Table 4). Some authors also recommend standardizing the M variable (MacKinnon & Dwyer, 1993; Winship & Mare, 1983), but we omit standardizing M here for simplicity.

Because Y is now binary, we must specify binary\_y = TRUE when using mlm() to estimate the model. We also take the opportunity here to illustrate how to adjust the prior scale parameters when estimating the model. We assign the b parameter's Gaussian prior distribution's SD to 1. This (somewhat arbitrary) prior assumption means that our prior knowledge about b is described by a Gaussian distribution centered on zero with a standard deviation of 1. For example, if the estimated parameter was exactly 1, then the effect of one unit of hr\_cw on the log-odds of jop\_bin was 1; a rather implausibly large effect. In effect, this prior then a priori constrains plausible parameter values to be closer to zero as described by the N(0,1) distribution.

Table 4						
Example	data	with	a	binary	Y	variable.

subj	lag	hr	jop	hr_cw	jop_bin
1	1	36.67	30	-23.33	0
1	1	50.00	80	-10.00	1
1	0	56.67	90	-3.33	1
1	1	56.67	32	-3.33	0
1	1	56.67	50	-3.33	0
1	0	70.00	76	10.00	1

*Note.* jop\_bin is a within-person median split version of the original jop variable.

The estimated model is now in an R object called fit\_bin. All the summarizing and plotting functions illustrated above can be used with the binary Y model as well. However, all the model's coefficients that refer to Y (and their transformations, such as the mediation effect) are products of a linear regression coefficient (path a), and a logistic regression coefficient (path b). These values are usually difficult to interpret, and we therefore recommend users to visualize the fitted values of the model.

## 44 Visualizing the model's fitted values

A helpful function for visualizing the fitted values is mlm\_spaghetti\_plot(), which is
used to draw "spaghetti" plots that show fitted values at the population- and subject-levels.
Spaghetti plots make the relationships between the variables particularly salient, by plotting
the model's fitted values in the data space (i.e. the *b* path is plotted in probability space).
We illustrate how to use this function below:

```
mlm_spaghetti_plot(
    mod = fit_bin,
    d = MEC2010,
    x = "lag", m = "hr_cw", y = "jop_bin", id = "subj",
    fixed = TRUE, random = FALSE, binary_y = TRUE, n = 20)
```

The input arguments to mlm spaghetti plot() are the model (mod), the data frame

used to fit the model (d), and the X, M, Y, and id variable names (as they are in the data). 651 Further arguments allow the user to decide to visualize the population-level effects (fixed), 652 subject-specific effects (random), or both. Finally, we also specified that the model has a 653 binary outcome variable binary\_y = TRUE, and ensured that the lines look smooth by 654 specifying that the fitted lines should be evaluated along 20 points on the x-axis (n = 20). 655 The resulting figure (Figure 8) is especially useful for understanding the model when Y 656 is binary, because the estimated parameters referring to Y are in log-odds, and the mediation 657 effect is a product of a linear coefficient and a log-odds coefficient. This visualization is 658 directly interpretable because the slope of the b path is shown in probability space, and visualized such that the x-axis values are the fitted values of M from the a path regression, and thus carry the effect size of the a path to plausible effects of the b path. Finally, the 661 gray shades surrounding the regression lines in both panels of the figure are by default 95% 662 Credibility Intervals, but the percentage can be adjusted with the level argument of 663 mlm\_spaghetti\_plot(). 664

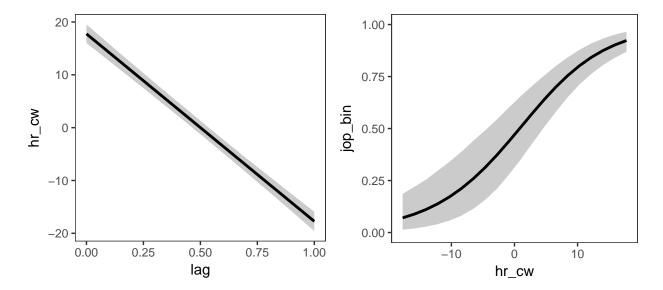


Figure 8. Fitted values of the multilevel mediation model with a binary Y. Left panel: Population-level regression line for path a and its 95%CI as a grey shade. Right panel: Population level regression line for path b in probability space, and its 95%CI.

The same function can also be used to display regression lines of the a and b paths for every individual in the study by setting the function's argument random = TRUE. Figure 9 shows the resulting "spaghetti" plot. These figures are especially helpful in illustrating the heterogeneity of effects among participants, which in this study was very small.

## 669 Estimating the magnitude of mediation

665

666

667

668

670

population-level mediation and direct effects, respectively, another approach to assessing the 671 magnitude of mediation is to calculate the proportion of the total effect that is mediated,  $\frac{me}{a}$ 672 (MacKinnon et al., 2007; MacKinnon, Warsi, & Dwyer, 1995; Shrout & Bolger, 2002). 673 Because the Bayesian framework provides a full multivariate posterior distribution, 674 obtaining the posterior distribution of  $\frac{me}{c}$  is straightforward. This quantity is saved in the 675 estimated model as pme, for proportion mediation effect (or "proportion of effect that is 676 mediated"). The resulting marginal posterior distribution from the current example is 677 illustrated in Figure 10. It is important to note that interpreting pme is straightforward only 678

While me and c' together provide information on the magnitude of the

690

691

692

693

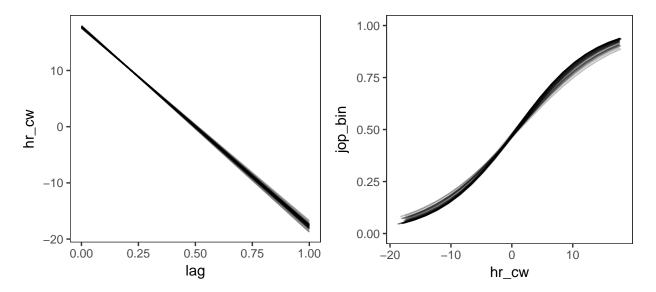


Figure 9. Fitted values of the multilevel mediation model with a binary Y, for every individual in the study. Left panel: Subject-specific regression lines for path a. Right panel: Subject-specific fitted values for path b in probability space.

if the mediated and direct effects are of the same sign (Shrout & Bolger, 2002). For this, and 679 other reasons, estimated values of pme may exceed 1 or be negative, and therefore do not 680 represent a true proportion. We recommend interpreting values greater than 1 as 1. 681 Although the usefulness of this metric can be disputed because it does not represent a true 682 proportion, it can sometimes be useful—especially as a quick and rough estimate of the 683 importance of the mediated effect—and we include it in the model's output, but remind 684 researchers to be cautious when interpreting it. Keeping this in mind, the posterior 685 distribution of pme in Figure 10 suggests that most plausible values are very close to 1, again 686 reinforcing our conclusion of total mediation of lag's effect on judgments of performance 687 through hit rates. 688

Other effect size measures for mediated effects have also been proposed, such as  $\kappa^2$ , which is a standardized effect size denoting the "proportion of the maximum possible indirect effect that could have occurred, had the constituent effects been as large as the design and data permitted" (Preacher & Kelley, 2011, p. 106). However, the usefulness and definition of the  $\kappa^2$  metric has been contested for a number of reasons, including that it can

698

699

708

709

710

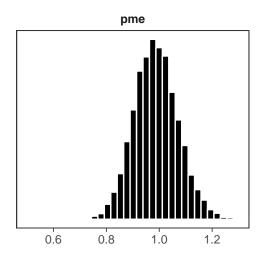


Figure 10. MCMC samples from the posterior distribution of the population-level proportion of effect that is mediated.

decrease even though the underlying mediation effect increases (Wen & Fan, 2015). 694

Furthermore, non-standardized effect sizes are often easier to interpret, and as such can 695 be more informative about the measures used in the experiment (Baguley, 2009). Therefore, we recommend describing the results of multilevel mediation analyses using unstandardized effect sizes such as the coefficients a, b, c', and their transformations such as me. This recommendation echoes Tukey's note that "being so disinterested in our variables that we do not care about their units can hardly be desirable." (Tukey, 1969, p. 89). 700

In the case of a binary Y variable, the regression coefficients b and c' are more difficult 701 to interpret directly, because they report unstandardized effect sizes in the log-odds scale. 702 Consequently, the mediated effect me is a product of a linear effect on the data scale (a), 703 and a linear effect on the log-odds scale (b). However, this complication reflects the fact that 704 mediation effects are inherently combinations of many variables, and there may not be a 705 single metric that adequately captures the mediation effect size in all situations (Wen & Fan, 706 2015). 707

Consequently, we recommend reporting not only the mediated effect me, but its constituent parts a and b as well (and covariance  $\sigma_{a_ib_i}$ , if it is important in the current analysis). If the a and b paths are of the same sign, the proportion of the total effect that is

Table 5

Main functions of bmlm.

Function	Purpose Inputs		
isolate()	Create within-person variables	Data, variable names	
mlm()	Estimate a multilevel model	Data, variable names, MCMC options	
$mlm\_summary()$	Print parameters to R console	ole Model, parameters, credibility level	
$mlm\_pars\_plot()$	Plot mediation model's parameters	rameters Model, plot type, parameters	
$mlm\_path\_plot()$	Plot the model as a path diagram Model, variable names		
$mlm\_spaghetti\_plot()$	Plot fitted values (regression line)	Model, data, variable names	
tab2doc()	Create a Word summary document	Results of mlm_summary()	

*Note.* To learn more about each function, type the function's name prepended with a question mark in the R console. This will bring out the function's help page.

mediated,  $\frac{me}{c}$ , should also be reported. Because the Bayesian analysis automatically provides posterior distributions of these parameters, inference (and communication) should not focus on point estimates, but their associated uncertainty intervals should also be reported.

Additionally, the communication and interpretation of mediation analyses is greatly facilitated by graphical descriptions of the data and estimated model, such as Figures 8 and 9. We have specifically designed **bmlm**'s plotting functions to facilitate the interpretation and communication of results.

#### 718 Summary of bmlm's functions

Above, we have illustrated the functionality of **bmlm** with an empirical example.

Table 5 provides a quick reference table to its main functions.

721 Discussion

## • Comparison to other software

Software options for implementing multilevel mediation are limited, and to date have 723 been limited to commercial software. Bolger & Laurenceau (2013) used Mplus (L. K. 724 Muthén & Muthén, 2017) to estimate a  $1 \to 1 \to 1$  multilevel mediation with continuous Y, 725 and here we provide a summary table comparing estimated parameters from bmlm and Mplus for the example discussed in Bolger & Laurenceau (2013). (This data set is included in bmlm as BLch9.) 728 Table 6 shows that the point estimates, and their standard errors (for **bmlm** these are 729 posterior means and standard deviations), obtained from bmlm and Mplus are in agreement, 730 and numerical differences are small. However, it is important to emphasize that only 731 focusing on the point estimate and SE of some of the estimated parameters can be less 732 informative than viewing the full posterior distribution, because the shape of the distribution 733 can be very non-Gaussian, in which case these two numbers may be misleading. For example, 734 Figure 5 shows that the posterior distributions of standard deviation and covariance 735 parameters characterizing the random effects may be very skewed; summarizing these with 736 just a mean and SD may lead to inaccurate inferences. In summary, numerical results from 737 bmlm are the same as would be obtained using commercial software (Mplus). 738 Why, then, should researchers choose to use **bmlm** over the more general Mplus 739 software? First, on our reading, the Mplus software—and its modeling language—is not well 740 known or commonly used within cognitive psychology and neuroscience. The R software, is 741 well known within these fields, and because at its core **bmlm** is only a library of R functions, users who are familiar with R can estimate the model within minutes of installation. A second benefit of using the comparably more limited **bmlm** has to do with the fact that it is emphatically not a general purpose (structural equation) modeling tool; it does few things, 745 but it does them well. For example, the figures illustrated above are very useful in 746 interpreting and communicating results from the analysis, and are available to users by

Table 6

Comparison of parameters estimated with bmlm and Mplus, from a model using example data presented in Bolger and Laurenceau (2013), chapter 9. The bmlm estimates are posterior means, and SEs are posterior standard deviations (based on 100,000 MCMC samples).

Parameter	Estimate (bmlm)	Estimate (Mplus)	SE (bmlm)	SE (Mplus)
a	0.19	0.19	0.04	0.04
b	0.15	0.15	0.03	0.03
ср	0.10	0.10	0.02	0.02
c	0.16	0.16	0.03	0.03
me	0.06	0.06	0.01	0.01
pme	0.36	0.36	0.08	0.08
covab	0.03	0.03	0.01	0.01
tau_a	0.26	0.26	0.04	NA
tau_b	0.22	0.21	0.03	NA
tau_cp	0.08	0.09	0.03	NA
sigma_y	0.93	0.92	0.02	NA
sigma_m	1.09	1.09	0.02	NA

Note. The variance components' SEs are missing from the Mplus results because they were reported in the variance scale in Bolger and Laurenceau (2013), and therefore not directly comparable to the current results.

simply using their associated functions. Because Mplus is a general purpose modeling tool, it does not easily provide these types of figures for specific purposes, such as detailed here.

The two most important reasons for using **bmlm** over its commercial alternatives, 750 such as Mplus, however, are openness and price. Regarding the former, there has recently 751 been an enormous push toward increasing scientific reproducibility, openness, and 752 transparency (e.g. Eglen et al., 2017; Munafò et al., 2017; Vuorre & Curley, 2017). Because 753 the source code of our program is freely available, it is easily accessible to public scrutiny, 754 improvement and communication, thereby potentially increasing the aforementioned goals. 755 Second, providing the package within the R ecosystem makes literate programming (Knuth, 756 1984) more accessible than standalone programs, thereby possibly improving reproducibility 757 (literate programming is the combining of computer code and language to enhance technical 758 communication). Finally, perhaps the most import difference between **bmlm** and its 759 commercial alternatives, such as Mplus, is that our program and code is free to use, modify 760 and extend. For many researchers, software licenses can be too expensive, but free programs 761 don't require re-allocation of research funds toward programs, and thereby make these useful 762 methods available to a broader audience of researchers. 763

#### 764 Limitations

Currently, **bmlm's** implementation of multilevel mediation requires that the data set be submitted to the analysis with complete rows. That is, missing cells within rows are not allowed, and users are required to either drop all rows of data that are not complete, or fill the data before entering it to the mlm() function. However, the software does not require that the data set is balanced either across individuals or across conditions within individuals. We believe that not allowing missing values is not a great limitation, because in cognitive experiments data is usually collected with a computerized experiment, making missing rows (e.g. outliers for M or Y leading to the entire trial being rejected from analysis; allowed) much more common than missing values (e.g. value for a single variable not logged for one

trial; not allowed.)

Another limitation of **bmlm** is that it currently implements only the  $1 \to 1 \to 1$ 775 mediation model (with continuous / binary Y), and more complex mediation models are not allowed. Furthermore, issues such as covariates and latent variables (e.g. Cheong et al., 2003) are often discussed in the literature on mediation, possibly making the model presented here seem as limited in scope. However, latent variables, longitudinal models, and covariates are 779 not widely applicable in experimental studies in cognitive psychology and neuroscience: 780 While future work might address these issues, we feel that this relatively simple model is 781 widely (and easily) applicable to a wide range of data within these fields. Furthermore, the 782 model's Stan source code is extensively commented and modular, thus making it easier for 783 experienced users to expand it to more complex models. We plan to implement some 784 common but more complicated models in the software in our future work, but believe that 785 the models currently provided cover a large number of common use cases in cognitive 786 psychology and neuroscience. 787

# Considerations for Analyzing Causal Models

We also remind readers of the general limitations and pitfalls of analyzing causal models with statistical mediation, and the additional complications related to allowing the hypothesized causal effects to vary randomly between individuals. One of the primary benefits of controlled experiments is that hypothesized causal variables (X) are manipulated, and the causal assumptions are therefore easier to accept. With mediation models, the situation is more complicated because the mediating variable (M) is thought to exert a causal influence on Y, yet it is not experimentally manipulated.

Ultimately, to adjudicate causation from correlation in the M-Y relationship, strong theoretical, logical, and experimental considerations need to be taken into account. For instance, it is important to ensure that the temporal sequence of X, M, and Y within an experimental trial supports causation from M to Y and not vice versa. In the example presented above, we can fairly certainly assume that the actual game performance during a trial (hit rate) occurred and was determined before the subject's judgment of performance (JoP) at the end of the trial. Another important consideration is that there should be no other mediator, correlated with the proposed M, that would instead explain the mediated effect.

A second issue with statistical mediation, and any regression method, is that measured 805 variables are assumed to be measured without error. If M is a very noisy measure of the 806 underlying construct, its association to M may be very difficult to find, or the relationship 807 may be otherwise unrepresentative of the relationship between the actual construct that M 808 represents and Y. Although outside the scope of this article, Bayesian methods allow 800 relatively straightforward relaxation of this assumption: If researchers have information 810 about the measurement error associated with a variable, they can include it in the model. 811 This approach is used, for example, in the Bayesian regression R package brms (Buerkner, 812 2016). Because M is measured, the assumption of no measurement error is often more 813 difficult to satisfy in mediation models than in models where all predictors are 814 experimentally manipulated. 815

Finally, the multilevel model allows all parameters to vary between subjects. It is 816 therefore possible that while the population level estimated parameter might indicate a 817 mediated effect for the average person, the subject-specific effects might indicate that the 818 effect is very weak—or might be in the other direction—for a subset of individuals. In these 819 cases, some researchers might object to the assertion that the mediation effect holds in the 820 population. However, we think that significant between-subject variance is a source of 821 inspiration for future research. We also note that the important issue of between-person heterogeneity is not specific to multilevel mediation, but applies to all research and analyses where effects can vary between individuals. Multilevel models are useful-among other reasons—because they bring this heterogeneity to researchers' attention, and possibly deepen 825 their understanding of the research question. 826

### 27 Software dependencies and development

At its core, **bmlm** uses the Stan programming language through the rstan R interface 828 for estimating the mediation model (Stan Development Team, 2016a, 2016b). After 829 estimating the model, users may export the underlying Stan code (run 830 cat(rstan::get stancode(fit)) in R) and use it to extend the mediation model to 831 answer more complex questions. bmlm's core functions also depend on the R packages 832 dplvr (Wickham & Francois, 2016) and Rcpp (Eddelbuettel & Francois, 2011). bmlm's 833 plotting functions depend on R packages ggplot2 (Wickham, 2016) and ggraph (Epskamp, 834 Cramer, Waldorp, Schmittmann, & Borsboom, 2012). Situating **bmlm** in the R ecosystem 835 also makes it easy for users to write reproducible reports and manuscripts using the R 836 packages knitr (Xie et al., 2016), R Markdown (Allaire et al., 2016), and papaja (Aust & 837 Barth, 2016). Users may also use the tab2doc() function to directly export bmlm's results 838 to a Word table (Gohel, 2016). 839 For more extensive user instructions, we direct users to the package's website<sup>3</sup>. For 840 comments and feedback, such as suggestions of new features, users may visit the package's 841 GitHub website and leave a request for a new feature. This website also allows more 842 advanced users to copy the package's source code for extending its functionality. 843

#### 844 Conclusion

Statistical mediation allows researchers to address questions about causal mechanisms in which the effect of one variable on another is mediated by a third variable. Such research questions about causal relations are commonplace and important in psychological science, but to date have most commonly concerned between-person causal relations. The analysis of mediation at the within-person level is relatively uncommon and presents additional complexities, but comes with great benefits: When individual participants provide multiple measures of the IV, DV, and mediating variable, mediation can be assessed for each

<sup>&</sup>lt;sup>3</sup>https://mvuorre.github.io/bmlm/.

individual and the population average, and the inference to within-person psychological 852 processes is more straightforward. Additionally, multilevel mediation analysis provides 853 estimates of the between-person variability (heterogeneity) in the effects, which are important 854 when considering the generalizability of the observed effects (Bolger & Laurenceau, 2013). 855 Here, we discussed the multilevel modeling approach to investigating within-person 856 mediation (Kenny et al., 1998, 2003), and introduced a free, open-source software package 857 for the R programming environment for conducting Bayesian multilevel mediation analyses 858 (bmlm; Vuorre, 2016). This software package allows users to easily estimate multilevel 859 mediation models, and summarize and visualize its results. The software package is freely 860 available at https://cran.r-project.org/package=bmlm. 861

## Acknowledgements

We would like to thank Janet Metcalfe, Teal Eich, and Alan Castel for making their
data available. We also acknowledge Ben Goodrich and Jonah Gabry's help with the
development of bmlm on GitHub. This research was supported, in part, by Institute of
Education Science grant (R305A150467). The authors are solely responsible for the content
of this article.

References 868

```
Allaire, J. J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., ... (Ionicons), D.
869
          (2016). Rmarkdown: Dynamic Documents for R (Version 1.3). Retrieved from
870
          https://cran.r-project.org/web/packages/rmarkdown/index.html
871
   Atlas, L. Y., Bolger, N., Lindquist, M. A., & Wager, T. D. (2010). Brain mediators of
872
           predictive cue effects on perceived pain. The Journal of Neuroscience, 30(39),
873
          12964–12977. https://doi.org/10.1523/JNEUROSCI.0057-10.2010
874
   Aust, F., & Barth, M. (2016). Papaja: Create APA manuscripts with RMarkdown. Retrieved
          from https://github.com/crsh/papaja
876
   Baguley, T. (2009). Standardized or simple effect size: What should be reported? British
877
          Journal of Psychology, 100(3), 603-617. https://doi.org/10.1348/000712608X377117
878
   Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social
879
          psychological research: Conceptual, strategic, and statistical considerations. Journal
880
          of Personality and Social Psychology, 51(6), 1173–1182.
881
          https://doi.org/10.1037/0022-3514.51.6.1173
882
   Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random
883
          indirect effects and moderated mediation in multilevel models: New procedures and
884
          recommendations. Psychological Methods, 11(2), 142–163.
          https://doi.org/10.1037/1082-989X.11.2.142
886
   Bolger, N., & Laurenceau, J.-P. (2013). Intensive longitudinal methods: An introduction to
887
          diary and experience sampling research. Guilford Press. Retrieved from
888
          http://www.intensivelongitudinal.com/
880
   Buerkner, P.-C. (2016). Brms: Bayesian Regression Models using Stan. Retrieved from
890
          http://CRAN.R-project.org/package=brms
891
   Cheong, J., MacKinnon, D. P., & Khoo, S. T. (2003). Investigation of Mediational Processes
892
           Using Parallel Process Latent Growth Curve Modeling. Structural Equation Modeling:
893
          A Multidisciplinary Journal, 10(2), 238–262.
```

```
https://doi.org/10.1207/S15328007SEM1002_5
895
    Cole, D. A., & Maxwell, S. E. (2003). Testing Mediational Models With Longitudinal Data:
896
           Questions and Tips in the Use of Structural Equation Modeling. Journal of Abnormal
897
          Psychology, 112(4), 558–577. https://doi.org/10.1037/0021-843X.112.4.558
   Eddelbuettel, D., & Francois, R. (2011). Rcpp: Seamless R and C++ Integration. Journal
899
          of Statistical Software, 40(1), 1–18. https://doi.org/10.18637/jss.v040.i08
900
    Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P., ... Poline,
901
          J.-B. (2017). Toward standard practices for sharing computer code and programs in
902
          neuroscience. Nature Neuroscience, 20(6), 770–773. https://doi.org/10.1038/nn.4550
903
   Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012).
904
           Qgraph: Network Visualizations of Relationships in Psychometric Data. Journal of
905
           Statistical Software, 48(4), 1–18. Retrieved from http://www.jstatsoft.org/v48/i04/
    Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models
907
          (comment on article by browne and draper). Bayesian Analysis, 1(3), 515–534.
908
          Retrieved from http://projecteuclid.org/euclid.ba/1340371048
909
    Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical
910
          models. Cambridge University Press.
911
    Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013).
912
          Bayesian data analysis, third edition. Boca Raton: Chapman; Hall/CRC.
913
    Gohel, D. (2016). ReporteRs: Microsoft Word and PowerPoint Documents Generation.
914
          Retrieved from https://CRAN.R-project.org/package=ReporteRs
915
   Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path
          lengths in hamiltonian monte carlo. Journal of Machine Learning Research, 15(1),
917
          1593-1623.
918
   Ishigami, Y., & Klein, R. M. (2009). Is a hands-free phone safer than a handheld phone?
919
          Journal of Safety Research, 40(2), 157–164. https://doi.org/10.1016/j.jsr.2009.02.006
920
    Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not)
921
```

```
and towards logit mixed models. Journal of Memory and Language, 59(4), 434–446.
922
          https://doi.org/10.1016/j.jml.2007.11.007
923
   Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation
          and moderation in within-subject designs. Psychological Methods, 6(2), 115–134.
925
          https://doi.org/10.1037/1082-989X.6.2.115
   Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. Journal of the American Statistical
927
          Association, 90(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572
928
   Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D.
          T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The handbook of social psychology, vols.
930
          1 and 2 (4th ed.) (pp. 233–265). New York, NY, US: McGraw-Hill.
931
   Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel
932
          models. Psychological Methods, 8(2), 115–128.
933
          https://doi.org/10.1037/1082-989X.8.2.115
934
   Knuth, D. E. (1984). Literate Programming. The Computer Journal, 27(2), 97–111.
935
          https://doi.org/10.1093/comjnl/27.2.97
936
   Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level
937
          mediated effects. Multivariate Behavioral Research, 36(2), 249–277.
938
          https://doi.org/10.1207/S15327906MBR3602_06
   Kruschke, J. K. (2014). Doing bayesian data analysis: A tutorial introduction with r (2nd
940
          Edition). Burlington, MA: Academic Press.
941
   Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices
          based on vines and extended onion method. Journal of Multivariate Analysis, 100(9),
943
          1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008
   MacKinnon, D. P. (2008). Introduction to Statistical Mediation Analysis. Routledge.
945
   MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating Mediated Effects in Prevention
          Studies. Evaluation Review, 17(2), 144–158.
```

```
https://doi.org/10.1177/0193841X9301700202
948
   MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. Annual
949
          Review of Psychology, 58(1), 593-614.
950
          https://doi.org/10.1146/annurev.psych.58.110405.085542
   MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect
952
          measures. Multivariate Behavioral Research, 30(1), 41–62.
953
          https://doi.org/10.1207/s15327906mbr3001_3
954
   Marsman, M., & Wagenmakers, E.-J. (2016). Three Insights from a Bayesian Interpretation
955
          of the One-Sided P Value. Educational and Psychological Measurement.
956
          https://doi.org/10.1177/0013164416669201
957
   McElreath, R. (2016). Statistical rethinking: A bayesian course with examples in r and stan.
958
           CRC Press.
959
   Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. Journal of Experimental
960
          Psychology: General, 136(2), 184–199. https://doi.org/10.1037/0096-3445.136.2.184
961
   Metcalfe, J., Eich, T. S., & Castel, A. D. (2010). Metacognition of agency across the lifespan.
          Cognition, 116(2), 267–282. https://doi.org/10.1016/j.cognition.2010.05.009
963
   Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical
964
          mediation analysis: A path-analytic framework. Psychological Methods, 22(1), 6.
965
          https://doi.org/10.1037/met0000086
966
   Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The
967
          fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review.
968
   Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N.
          P., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature
970
          Human Behaviour, 1, 0021. https://doi.org/10.1038/s41562-016-0021
   Muthén, L. K., & Muthén, B. O. (2017). Mplus User's Guide (8th ed.). Los Angeles, CA:
972
           Muthén & Muthén.
973
   Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new
```

```
developments. Annual Review of Psychology, 66(1), 825–852.
975
           https://doi.org/10.1146/annurev-psych-010814-015258
976
    Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models:
           Quantitative strategies for communicating indirect effects. Psychological Methods,
978
           16(2), 93–115. https://doi.org/10.1037/a0022658
970
    Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for
980
           assessing multilevel mediation. Psychological Methods, 15(3), 209–233.
981
           https://doi.org/10.1037/a0020141
982
    R Core Team. (2016). R: A language and environment for statistical computing. Vienna,
983
           Austria: R Foundation for Statistical Computing. Retrieved from
984
           https://www.R-project.org/
985
    Raudenbush, S. W., & Sampson, R. (1999). Assessing Direct and Indirect Effects in
986
           Multilevel Designs with Latent Variables. Sociological Methods & Research, 28(2),
987
           123–153. https://doi.org/10.1177/0049124199028002001
988
    Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is
989
           There a Free Lunch in Inference? Topics in Cognitive Science, 8(3), 520–547.
990
           https://doi.org/10.1111/tops.12214
991
    Selig, J. P., & Preacher, K. J. (2009). Mediation Models for Longitudinal Data in
992
           Developmental Research. Research in Human Development, 6(2-3), 144–164.
           https://doi.org/10.1080/15427600902911247
    Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies:
995
           New procedures and recommendations. Psychological Methods, 7(4), 422–445.
996
           https://doi.org/10.1037/1082-989X.7.4.422
    Stan Development Team. (2016a). RStan: The R interface to Stan. Retrieved from
998
           http://mc-stan.org/
990
    Stan Development Team. (2016b). Stan: A C++ Library for Probability and Sampling,
1000
```

```
Version 2.14.1. Retrieved from http://mc-stan.org/
1001
    Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of
1002
           Experimental Psychology, 18(6), 643-662. https://doi.org/10.1037/h0054651
1003
    Tofighi, D., West, S. G., & MacKinnon, D. P. (2013). Multilevel mediation analysis: The
1004
           effects of omitted variables in the 1-1-1 model. British Journal of Mathematical and
1005
           Statistical Psychology, 66(2), 290–307.
1006
           https://doi.org/10.1111/j.2044-8317.2012.02051.x
1007
    Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? American
1008
           Psychologist, 24(2), 83–91. https://doi.org/10.1037/h0027108
1000
    Valente, M. J., & MacKinnon, D. P. (2017). Comparing Models of Change to Estimate the
1010
           Mediated Effect in the Pretest-Posttest Control Group Design. Structural Equation
1011
           Modeling: A Multidisciplinary Journal, 24(3), 428–450.
1012
           https://doi.org/10.1080/10705511.2016.1274657
1013
    van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2016). A simple introduction to Markov
1014
           Chain Monte-Carlo sampling. Psychonomic Bulletin & Review, 1–12.
1015
           https://doi.org/10.3758/s13423-016-1015-8
1016
    Vuorre, M. (2016). Bmlm: Bayesian Multilevel Mediation. Retrieved from
1017
           https://cran.r-project.org/package=bmlm
1018
    Vuorre, M., & Curley, J. P. (2017). Curating Research Assets in Behavioral Sciences: A
1019
           Tutorial on the Git Version Control System. PsyArXiv Preprints.
1020
           https://doi.org/10.17605/OSF.IO/TXGN8
1021
    Vuorre, M., & Metcalfe, J. (2016). The relation between the sense of agency and the
1022
           experience of flow. Consciousness and Cognition, 43, 133–142.
1023
           https://doi.org/10.1016/j.concog.2016.06.001
1024
    Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian
1025
           hypothesis testing for psychologists: A tutorial on the Savage–Dickey method.
1026
```

```
Cognitive Psychology, 60(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001
1027
    Wang, L. (Peggy), & Preacher, K. J. (2015). Moderated Mediation Analysis Using Bayesian
1028
           Methods. Structural Equation Modeling: A Multidisciplinary Journal, 22(2), 249–263.
1029
           https://doi.org/10.1080/10705511.2014.935256
1030
    Wen, Z., & Fan, X. (2015). Monotonicity of effect sizes: Questioning kappa-squared as
1031
           mediation effect size measure. Psychological Methods, 20(2), 193–203.
1032
           https://doi.org/10.1037/met0000029
1033
    Wickham, H. (2016). Gaplot2: Elegant Graphics for Data Analysis (2nd ed.). Springer
1034
           Science & Business Media.
1035
    Wickham, H., & Francois, R. (2016). Dplyr: A Grammar of Data Manipulation. Retrieved
1036
           from http://CRAN.R-project.org/package=dplyr
1037
    Winship, C., & Mare, R. D. (1983). Structural Equations and Path Analysis for Discrete
1038
           Data. American Journal of Sociology, 89(1), 54–110. https://doi.org/10.1086/227834
1039
    Xie, Y., Vogt, A., Andrew, A., Zvoleff, A., Simon, A., Atkins, A., ... Foster, Z. (2016).
1040
           Knitr: A General-Purpose Package for Dynamic Report Generation in R (Version
1041
           1.15.1). Retrieved from https://cran.r-project.org/web/packages/knitr/index.html
1042
    Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. Psychological Methods,
1043
           14(4), 301–322. https://doi.org/10.1037/a0016972
1044
    Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using
1045
           hierarchical linear models problems and solutions. Organizational Research Methods,
1046
```

12(4), 695-719. https://doi.org/10.1177/1094428108327450

1047