UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Devine, S., Uanhoro, J. O., Otto, A. R., & Flake, J. K. (2024). Approaches for Quantifying the ICC in Multilevel Logistic Models: A Didactic Demonstration. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.94263

Methodology and Research Practice

Approaches for Quantifying the ICC in Multilevel Logistic Models: A Didactic Demonstration

Sean Devine[1][a], James O. Uanhoro[2], A. Ross Otto[1], Jessica K. Flake[1]

[1] Department of Psychology, McGill University, Montreal, Canada, [2] Education Psychology, University of North Texas, Denton, TX, USA

Keywords: Multilevel modeling, ICC, variance, tutorial

https://doi.org/10.1525/collabra.94263

Collabra: Psychology

Vol. 10, Issue 1, 2024

Multilevel modeling techniques have gained traction among experimental psychologists for their ability to account for dependencies in nested data structures. Increasingly, these techniques are extended to the analysis of binary data (e.g., correct or incorrect responses). Despite their popularity, the information in logistic multilevel models is often underutilized when researchers focus solely on fixed effects and ignore important heterogeneity that exists between participants. In this tutorial, we review four techniques for estimating and quantifying the relative degree of between-person variability in logistic multilevel models in an accessible manner using real data. First, we introduce logistic multilevel modeling, including the interpretation of fixed and random effects. Second, we review the challenges associated with the estimation and interpretation of within- and between-participant variation in logistic multilevel models, particularly computing the intraclass correlation coefficient (ICC), which is usually a first, simple step in a linear MLM. Third, we demonstrate four existing methods of quantifying the ICC in logistic multilevel models and discuss their relative advantages and disadvantages. Fourth, we present bootstrapping methods to make statistical inference about these ICC estimates. To facilitate reuse, we developed R code to implement the discussed techniques, which is provided throughout the text and as supplemental materials.

1. Introduction

From 2007 to 2017 there has been a threefold increase in published psychology articles that use multilevel modeling techniques (Huang, 2018), which is in part due to the *nested* structure of psychological data: individual observations clustered within participants or participants clustered within groups. Across multiple areas of psychology, behavioural economics, and other behavioural sciences, nested data often arise from experiments in which participants have completed multiple trials of a task. Analyzing these data with traditional statistical tests (e.g., fixed-effects ANOVA, linear regression) fails to account for this nested structure and violates the assumption of independence of observations. When responses come from the same participant, they are more similar than if they come from independently sampled participants. These response redundancies deflate the effective sample size, cause incorrect standard errors, and often cause higher Type I error rates (Snijders & Bosker, 2011).

In addition to these data being nested, the outcome variable of interest is many psychological datasets are often binary (e.g., accuracy or binary choices), rather than continuous. Such data can be modeled using *logistic multilevel models* (i.e., logistic mixed-effects regressions), which model the probability of success/correct choice while handling the dependencies that arise in nested datasets. However, these techniques are underexploited in practice. That is, while it is becoming common for researchers to fit multilevel models to nested experimental data, interpretation of the results is primarily, if not exclusively, focused on the 'fixed effects' of the model. As a result, estimates based on the variance of the random effects— e.g., the intraclass correlation (ICC), which capture the relative importance of variability *between* participants' responses (or between other types of clusters)—are seldom used to support or speak against conclusions.

Why should researchers care about the ICC in their multilevel logistic models? First, it quantifies theoretical phenomena, such as, how variable participants' responses are during a task or how stable an effect is across people. Relatedly, this provides a starting point for explanatory power: if most of the variability in the data comes from differences between persons, then model building is a useful exercise with person-level predictors, whereas if most of the variability in the data comes from within-person variations,

a Correspondence: seandamiandevine@gmail.com

then the researcher will want to explore the variability of responding to their task and investigate differences between conditions. At present, it is uncommon for researchers to report or even investigate the random effects variance or the ICC when outcomes are binary, which is inconsistent with best practices for the transparent reporting of any multilevel model (Luo et al., 2021).

To support researchers in expanding their understanding of their data and encourage best practice, this tutorial explains how to estimate and interpret the ICC in logistic multilevel models. A tutorial focusing on this topic is needed for at least three reasons: 1) the unique challenges that computing and interpreting the ICC in logistic multilevel models poses are seldom addressed in an accessible manner, 2) while other works have provided methods to compute the ICC in a logistic multilevel context (e.g., Goldstein et al., 2002; Merlo et al., 2006), there does not exist a didactic survey of methods aimed at increasing uptake by psychologists (see, for instance, examples in social epidemiology; Austin & Merlo, 2017; Merlo et al., 2006), and 3) texts that do address this issue tend to focus on only a single method (e.g., Snijders & Bosker, 2011) ignoring the differences between various methods.

Accordingly, we have four goals. First, for the uninitiated, we provide a brief introduction to linear multilevel regression (in the Supplemental Materials, sections **S6b**, explained below) and logistic multilevel regression, which includes the interpretation of fixed and random effects. Second, we review the challenges associated with the interpretation and estimation of the relative contribution of within- and between-participant variation in logistic multilevel models—namely in the computation of the intraclass correlation (also known as the variance partition coefficient, VPC). Third, we explain four existing methods of quantifying the ICC in logistic multilevel models, highlighting the differences between the methods. Finally, we describe bootstrapping methods which permit statistical inferences about these variance estimates and comment on the application of the reviewed techniques for models with additional predictors and random slopes. Throughout the tutorial we include *Code Boxes* which illustrate the concepts discussed in the main text using the R programming language, in a step-by-step fashion.[1]

To accomplish these goals, we use an example dataset from experimental cognitive psychology (see Section 2), where responses are nested within-participants. However, the topics discussed extend to cases where binary data may be nested within other types of clusters (e.g., groups, schools, countries, etc.). Namely, while the computations presented below do not change based on the nature of the grouping variable in the data, the interpretation of the ICC does. In contexts where the data are grouped by organizations or "clusters", such as data collected from different schools or hospitals, the ICC measures the proportion of the total variance that is attributed to the variability between clusters, sometimes referred to as "the intra-*cluster* correlation." This quantifies the impact of the clustering variable (e.g., schools) as well as how subjects within the same cluster are similar. Another interpretation of the ICC in this context then is the average correlation in the outcome data between measurements from the same cluster. When participants provide repeated measures (e.g., in a psychological experiment), the ICC measures the proportion of the total variance that is due to the differences between individuals. Here, the ICC can provide insights about how individual differences among subjects exert an influence on the outcome. It is this latter interpretation of the ICC, in the context of a binary outcome variable, that will be the focus of this tutorial.

We developed materials to be reusable and reproducible. We encourage readers to follow along with the supplementary material, which we refer to throughout (e.g., **S1**, referring to "supplement section 1") and includes all R code with explanatory narrative that can be used to reproduce the results highlighted in the main text. The main text of this tutorial is aimed at someone who has experience executing multilevel models in R and would like to expand their knowledge to logistic models. For those with less background knowledge, readers can review the relevant sections of the supplemental materials (see Section **S6**).

2. Illustrative Data Description

Data for this tutorial are taken from a study examining how acute stress engenders avoidance of cognitively effortful tasks (Bogdanov et al., 2021; data available at https://osf.io/26w4u/). Thirty-eight young adults (20 female; PID in the dataframe) completed 300 trials of the well-characterized demand selection task (DST) in blocks of 150 trials, in which participants repeatedly chose between two nondescript cues that represented demand levels of a cognitively demanding task-switching paradigm: a low-demand condition and high-demand condition (`effort_choice`, 0 = low demand, 1 = high demand; see Figure 1).

In Bogdanov et al., (2021), participants completed the DST before and after being exposed to acute social stress (condition, control or stress), in a repeated-measures design. Their research question was whether willingness to exert cognitive effort (i.e., to choose the high effort cue) would change after being exposed to acute social stress. For the purposes of this tutorial, we will focus on data from sessions in which participants were not exposed to social stress to predict effort aversion (**S1** demonstrates how to load this subset of the data).

1 For ease of use, function code for each method is available for download from the following GitHub repository, which compute the measure of interest on a fitted logistic multilevel regression model: https://github.com/seandamiandevine/logisticicc. A brief demonstration of how to call these functions is provided in the Supplemental Materials (**S6d)**.

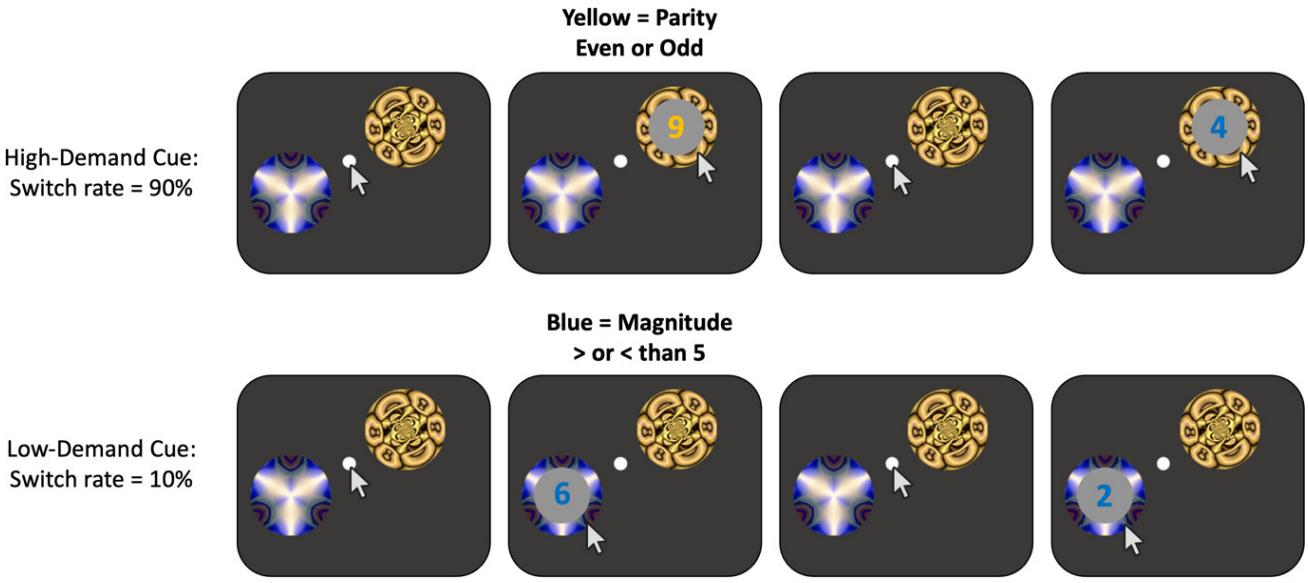

Figure 1. Schematic Representation of the DST.

In each trial of the task, participants chose between one of two pattern cues. After selecting a cue, a random number (range = 1–4 and 6–9) was presented in either yellow or blue at the center of the chosen cue. Participants then judged either the number's parity if the number was yellow (whether the number was odd or even) or its magnitude if the number was blue (whether the number was larger or smaller than 5). The choice of cue dictated the probability with which these tasks switched from trial-to-trial (the switch rate). Choosing the high-demand cue (upper row) resulted in a task-switching probability of .9, whereas choosing the low-demand cue (lower row) resulted in a task-switching probability of .1.

The traditional finding using this task is that participants will demonstrate a marked preference for the low-demand option over the high-demand option, reflecting a general bias against performing cognitively demanding activities, and in favour of less cognitively demanding activities (Kool et al., 2010). This recurring finding of demand-avoidant preferences has been interpreted as evidence that humans have a default tendency to avoid cognitively effortful tasks (Kool et al., 2010; Kool & Botvinick, 2018).

In this tutorial, the outcome variable of interest is participants' effort choices in the DST, which are binary (low demand or high demand) and nested (within participants; each participant makes many choices). We begin with how to model these data using logistic multilevel models, predicting the effect of effort level on cue preference in the DST, with the *lme4* package in R (Bates et al., 2015). For those less familiar with linear multilevel regression, review **S6a**, for those who need a refresher on logistic regression, review section 3.

3. Logistic Multilevel Regression

Linear multilevel regression models (see **S6b**) can be generalized to model binary outcomes. In this case, the goal is to model the probability that a given participant chose the high-effort option (`effort_choice`) on a given trial. If our binary data are coded as 0 (low effort option chosen) and 1 (high effort option chosen), this probability will be equal to the mean of the outcome: $p\left(Y_{ij}=1\right)=\overline{Y_j}$, where i refers to the i^{th} trial and j refers to j^{th} participant. However, predicting probabilities as a linear combination of predictors creates problems, because probabilities are bounded between 0 and 1 inclusive, whereas predicted values from a linear model (see **S6a** for more information) can, in theory, take on any value between negative and positive infinity. To circumvent this issue in logistic regression, a linear combination of the model's parameters and its predictors are computed on the logit scale—and then, afterwards, the regression predictions are transformed back to probabilities via a mean function. The regression on the logit scale is a linear model for the log of the odds of an event occurring (in this case, the odds of choosing the high effort option), where the odds represent the probability of the event occurring over its complement ($\frac{p(Y_{ij}=1)}{1-p(Y_{ij}=1)}$):

$$LO_{ij}=b_0+b_1X_{ij} \qquad \text{(Eq. 1)}$$

Here, the left-hand side of the equation represents the linear combination of b_0, b_1, and X on the logit scale (log-odds, LO), where X is some predictor variable of interest (presented here for completeness, though for most of this tutorial we will focus on intercept-only models). A feature of the logit scale is that it is simple to transform it into odds, and from there, into probabilities:

$$Odds_{ij}=\exp\left(LO_{ij}\right) \qquad \text{(Eq. 2)}$$

$$p\left(Y_{ij}=1\right)=\frac{\text{odds}_{ij}}{1+Odds_{ij}} \qquad \text{(Eq. 3)}$$

This gives us three ways to express the same linear combination of predictors in Eq. 1: log odds (i.e., logit scale), odds, and probabilities. In the case of effort choices in the DST, if $b_0 = -0.25$ and $b_1 = -.09$, assuming $X = 1$, then the log odds would be -0.34 (-0.25 + -0.09; Eq. 1), which is proportional to an odds of 0.71 (exp(-0.34); Eq. 2.), and a probability of 0.42 (0.71 / (1 + 0.71), Eq. 3), which all correspond to the same observed pattern: a participant is 0.71 times as likely to choose the high-demand option relative to low-demand option, or, equivalently, that a participant will select the high-demand option 42% of the time. This demonstrates an overall trend for participants to avoid the high demand option.

These techniques can be extended to a multilevel framework to account for the nested structure of the data (for more detail, see **S6b).**

$$\begin{aligned} \text{Level 1: } \log\left(\frac{p\left(Y_{ij}=1\right)}{1-p\left(Y_{ij}=1\right)}\right) &= b_{0j}+b_{1j}X_{ij} \\ \text{Level2: } b_{0j} &= \gamma_{00}+U_{0j} \\ \text{Level2: } b_{1j} &= \gamma_{10}+U_{1j} \end{aligned} \quad \text{(Eq. 4)}$$

Here, Y_{ij} represents the effort choice (0 or 1) on trial i for participant j. Similarly, X_{ij} corresponds to the value of X on trial i for participant j. Accordingly, b_{0j} is the person-specific log-odds for person j when X_{ij} is equal to 0. b_{1j} is the person-specific effect of X_{ij} on Y (in log-odds) for person j.

At the first level, participant j's effort choice on trial i are predicted from a linear combination of regression coefficients (b_{1j}) and trial-by-trial predictors (abstractly, X_{ij}), shifted by an intercept term (b_{0j}). The left-hand side of the level 1 equation (Eq. 4) corresponds to the log odds of choosing the high effort option.

As in the case of linear multilevel regression, Level 1 of equation 1 can be thought of as applying the standard (not multilevel) logistic regression to the data from each participant. Doing so would yield two coefficients on the log-odds scale unique to each participant—b_{0j} and b_{1j} described above. Conceptually, "averaging" both sets of coefficients (b_{0j} and b_{1j}) across participants would yield the fixed effects at level 2 (γ_{00}, γ_{10}) in Eq. 4. In other words, γ_{00} represents the average log-odds of $Y = 1$ assuming X_{ij} is zero for the average or typical person, and γ_{10} represents the average influence of X on the log odds of $Y = 1$ for the average person. These fixed effects are the *average of participant-specific effects* in log-odds. This is different from alternative techniques that return population-averaged fixed-effects when applied to correlated or nested data (e.g. Generalized Estimating Equations; Hardin & Hilbe, 2014).

γ_{00} and γ_{10} are estimated "averages" of participant-specific values in log-odds, but each participant will have coefficients that deviate from these average estimates. These deviations are captured by U_{0j} and U_{1j} in Eq. 4 and represent how much more or less a given participant's intercept or slope was from γ_{00} and γ_{10}, respectively. These values play a central role in a multilevel framework, as they capture variation between participants. They are assumed to be normally distributed with mean zero and variance τ^2: $U_{0j} \sim N\left(0, \tau_0^2\right), U_{1j} \sim N\left(0, \tau_1^2\right), \ldots U_{kj} \sim \left(N,\ 0,\ \tau_k^2\right)$. In this regard, the regression coefficients from Level-1 of equation 4 (the bs) are outcomes in Level-2. For example, the expected value of a given b_{0j} will be the sum of the average estimate, γ_{00}, plus the participant-level deviation, U_{0j}.

Such a model can be fit in R using the `glmer()` function from the *lme4* package (Bates et al., 2015; **S2**). First, a random intercept only model is fit, which captures participants' general preponderance for selecting the high effort tasks. In lme4 syntax, the model formula begins with `effort_choice ~ 1`, which asks R to fit a model where effort choice is predicted by its mean, the intercept, represented by a "1". Additionally, we add `(1|PID)` to this equation, which reflects the random intercept ("1") for each ("|") participant ("PID") in our dataset. We then specify the data frame that contains our variables, `data.Ctl`, and finally we indicate that the `family = 'binomial'`, which instructs R to fit a logistic regression model when it detects that the outcome variable is binary.

```
Library(lme4)

data.Ctl <- read.csv('Bogdanovetal2021/DST_data_osf2_ControlOnly.csv')

logistic_MLM0 <- glmer(effort_choice ~ 1 + (1|PID), data=data.Ctl,
family='binomial')

summary(logistic_MLM0)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: effort_choice ~ 1 + (1 | PID)
   Data: data.Ctl

     AIC      BIC   logLik deviance df.resid
 14234.5  14249.2  -7115.3  14230.5    11202

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7586 -1.0096  0.3631  0.9050  1.6890

Random effects:
 Groups Name        Variance Std.Dev.
 PID    (Intercept) 0.7554   0.8691
Number of obs: 11204, groups:  PID, 38

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.3338     0.1428   -2.338   0.0194 *
```

Code Box 1. Input and Output from an Intercept-Only Logistic Multilevel Model of Effort-Option Choices

Looking at the "Fixed effects" section of the output in Code Box 1, the intercept (`(Intercept)` in the output) is -0.33, which represents the average log-odds of choosing the high demand option for the typical person. This value can be converted to a probability using Eq. 2 and 3, which yields a 42% chance of choosing the high-demand option, which, consistent with past work, shows that, on average, people avoid high-demand options more often than chance (Kool et al., 2010; Patzelt et al., 2019).

Additionally, under the "Random effects" section of the output, the variance of participants' choices around this grand mean (τ_0^2, `Variance` in the output) is 0.76. This estimate is a quantification of the variability across individual participants. To gain a better intuition for this value, we can convert it from a variance to a standard deviation by taking the square root: $\tau_0 = \sqrt{\tau_0^2} = 0.87$. Since we assume normality roughly 95% of participants will have log-odds of choosing the high effort options within $\gamma_{00} \pm 2\tau_0 = -0.33 \pm 2 * 0.87 = [-2.07,\ 1.41]$. Converting these values into probabilities using Eq. 2 and 3, 95% of participants will choose the high effort option between 11% and 80% of the time. Thus, while the typical person avoids effort, some prefer it. Additionally, a researcher could visualize this variability by plotting the predicted participant-specific choice proportions (**S3**). To graphically depict this variance, we can extract the intercept for each participant from the model. These values are known as "empirical Bayes estimates" and are a weighted combination of person specific information and the average across all persons (γ_{00}). The EB estimates for persons with more informative data, e.g., from completing more trials, would be closer to their person-specific estimates. While EB estimates for persons with less informative data would be closer to the aver-

```
# extract random intercepts from model and add them to the fixed effect (gamma00)
eb  <- coef(logistic_MLM0)$PID
head(eb)


(Intercept)
1 -1.35826976
2 -0.28081092
3 -0.03887375
4 -0.03193523
5  0.77954561
6 -0.42367159
```

Code Box 2. Extraction of the Random Effect from an Intercept-Only Model in R

age across all persons (γ_{00}). The more trials a person completes, the more their EB estimates can be differentiated from the average across all persons. Alternatively stated, EB estimates reflect skepticism about differences between persons (or clusters more generally). Doing so we obtain a per participant measure of effort avoidance (b_{0j})—in other words, the fixed effect, -0.33 (γ_{00}), plus each participant's deviation from this fixed effect (U_{0j})—which yields each participant's estimated log-odds of choosing the high effort option. In R, this computation can be automated using the `coef(logsitic_MLM0)`.

The output of Code Box 2 shows the first 6 participants' estimated log-odds of choosing the high-effort option. To emphasize how these values correspond to participant behaviour, we have visualized the correspondence between each participant's estimated log-odds of choosing the high effort option and their actual proportion of high effort choices (Figure 2A). As can be seen, participants with higher estimated log-odds of choosing the high effort option also exhibited higher actual effort seeking behaviour. Notably, this relationship follows an "S-shaped" curve, which is the result of applying the logistic transformation discussed in Section 3 to the data (Eq. 1-3), allowing the log-odds to take both positive and negative values, whereas actual proportions are constrained to be between 0 and 1. It is important to note here that while above we have considered the fixed effects to be "averages" of person-level random effects, this is only true in the transformed log-odds scale. When computing log-odds back to probabilities, the non-linearity of the transformation makes it so large differences in log-odds correspond to only small differences in probabilities. For example, if we were to compare two (hypothetical) participants with estimated log-odds of choosing the high-effort option of 3 and 4 in terms of their raw probabilities, these values would correspond to probabilities 0.95 and 0.98—a difference of 3%. Conversely, if we compare two participants with the same difference in log-odds but lower absolute values, e.g., a participant with log-odds 1 and another with log-odds 2, then the corresponding probabilities are 0.73 and 0.88—a difference of 15%.

With this caveat in mind, we can also visualize between-participant variation in log-odds. As can be seen in Figure 2B, there is substantial variation in individual participants' effort-avoidance behaviour. As demonstrated in our numerical analysis above, a sizable proportion of participants demonstrate demand-*seeking* behaviour (red area in Figure 2B), preferring the high-demand option to the low-demand option on average (across trials). These individual differences expand theoretical understanding, because past work has often taken the fixed effects estimates of effort avoidance (Y_{00}) as evidence for a general and ubiquitous cognitive mechanism that aims to minimize effort exertion and maximize reward (e.g., Kool et al., 2010; Kool & Botvinick, 2018).

While visualizing these variations yields insight into the generality of the estimated fixed effect, it is useful for researchers to *quantify* the degree to which variability in the data stems from individual differences versus within-participant response noise (for example, Volpert-Esmond et al., 2018). In other words, in a multilevel context, it is important to distinguish between variability that stems from level-2 variance (i.e., differences *between* participants) and level-1 variance (variability *within* participants' response). The former provides insight into the generality of our conclusions—e.g., are *all* participants effort avoidant and, if not, which ones are and which ones are not?—while the latter speaks to the degree to which our experimental manipulation yields reliable responses from participants—e.g., how stable is one's aversion to effort over the course of the task? Generally-speaking, high within-person variability would require more explanatory within-person predictors, whereas high between-person variability would require a greater number of between-person predictors. With answers to these questions a researcher can move forward with model building.

In the context of linear multilevel models, variability attributable to between- versus within-participant differences is relatively easy to compute and outputted by default in statistical software (see **S6b**), but this is not the case with logistic models.

4. Challenges in Quantifying Variance in Logistic Multilevel Regression

In a linear random intercept model, the total variance in the outcome is the sum of between and within variance components: $var\left(Y_{ij}\right) = \tau_0^2 + \sigma^2$, where σ^2 is the residual, within-person, variance (see **S6b**). As such, one can compute the proportion of total variance accounted for by variations between individuals, τ_0^2,—a value called the intra-class correlation coefficient (ICC; a.k.a. variance partition coefficient), or simply ρ, as the ratio of variation about individuals to the total variance: $\rho = \frac{\tau_0^2}{var(Y_{ij})}$ (see **S6b** for more information). Conceptually, the ICC corresponds to the degree of nesting in a dataset, or the importance of the cluster variable (the variable in which observations are grouped; here, participants) in explaining variance in the outcome (Goldstein et al., 2002). This is why computing the ICC usually the best first step for a multilevel analysis (McCoach & Adelson, 2010). In cases where the data are nested within participants, the ICC summarizes the degree to which the outcome variable is sensitive to *individual differences*. Despite that many introductory texts on multilevel modeling discuss logistic MLMs as "simple" extensions of linear models, the ICC cannot be computed in the same way, and in

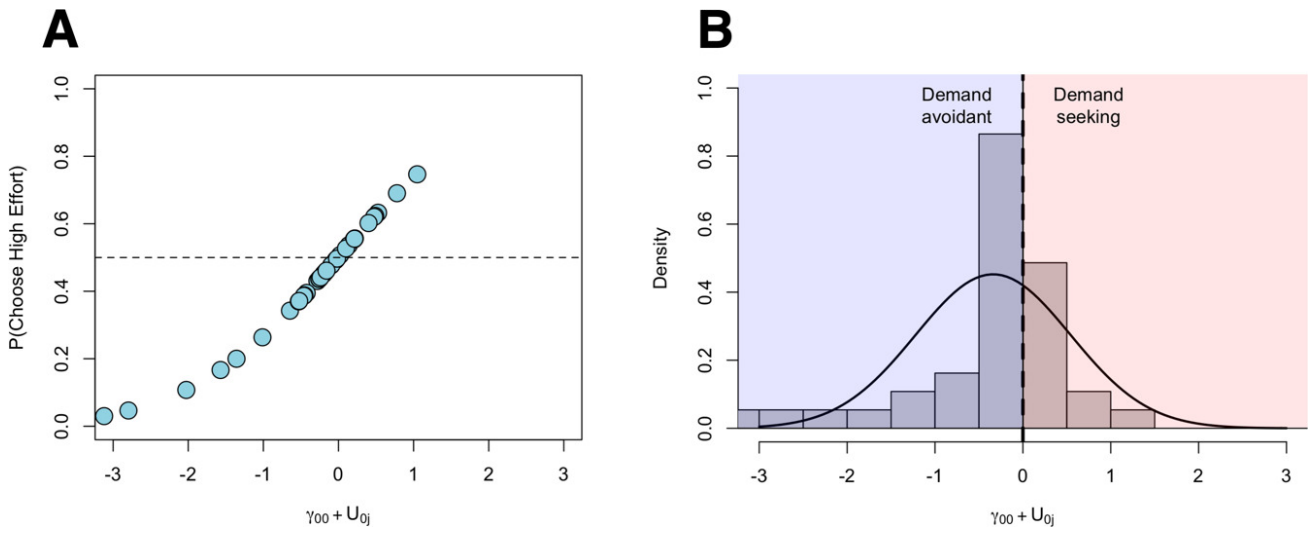

Figure 2. Correspondence between estimated demand preferences in log-odds during the DST from a logistic multilevel model and empirical behaviour.

The x-axis in both figures represents the fixed intercept of high effort choices plus each participant's deviation from this fixed intercept. Values on the x-axis are on log-odds scale. In A, the y-axis shows the proportion of high effort choices, and each point represents a single participant. In B, the y-axis represents the kernel density of participants within each bin of the histogram. As we can see, there is sizable variability in estimated effort aversion (i.e., the probability, in log odds, of choosing the high demand option). For some participants (in the red shaded area), the log odds are positive, indicating a probability above 0.50—that is, a *preference* for higher effort tasks.

fact not at all with the default output provided from *lme4* or other commonplace multilevel modeling software.

A key difference between linear and logistic multilevel models is that logistic models do not estimate a residual variance term, σ^2—that is, a term that represents within-participant variability in the outcome. In a linear model, the residual variance term is the variance of deviations, negative and positive, from a Gaussian variable's mean or prediction. In logistic models which use a Bernoulli distribution, the mean represents the probability of an event occurring. In turn, the variance of deviations about that mean/probability is entirely determined by the mean according to the following equation[2]: $var = P(Y = 1)\,(1 - P(Y = 1))$, .42(1-.42) = .24. As a result, a subtle, but important, difference between linear and logistic multilevel models is that logistic models, which are based on the Bernoulli distribution, lack a residual variance parameter, σ^2. The level 2 variance, τ^2, is on the logit scale. The variance of the Bernoulli outcome as implied by the mean, however, is on the probability scale. As such, we cannot combine the level-1 and level-2 variance to form the total variance denominator as is done in the standard computation of the ICC.

These caveats have consequences for the interpretation of variance parameters in a logistic multilevel model. First, there is no straightforward value for the residual variance (σ^2) that captures within-participant variation in the outcome—and accordingly, *lme4* does not provide one (the residual variance parameter when available would be printed under the Random effects section in Code Box 1). Thus, unlike a continuous outcome model, there is not a singular, "simple", way to quantify the relative contribution of individual differences to variability in the data, and thus the degree of unexplained variance (Goldstein et al., 2002). Consequently, variance in the outcome cannot be decomposed into within and between components in a straightforward manner and the standard ICC formula cannot be used to create a ratio of between-participant variance to total variance.

5. Within-Participant Variance and ICC in Logistic Multilevel Models

References texts usually discuss one method of quantifying the ICC in a logistic multilevel framework: the latent threshold method (e.g., Snijders & Bosker, 2011). However, other tested methods exist in the literature, and we review three additional approaches 1) the simulation approach, 2) linearization, and 3) the median odds ratio. While all four approaches attempt to quantify the degree of variability between people, they do not estimate the same exact statistical quantity and will often return different values. As we will see, the first three approaches estimate the ICC in the (binary) outcome variable but they do so differently. The

2 To demonstrate why this equation holds, let $\mathbf{X}$ be some binary variable coded as 0 and 1, θ be the mean of the distribution of this variable ($\mathbf{p}\,(\mathbf{Y} = \mathbf{1})$), and $\overline{\mathbf{X^2}} - \overline{(\mathbf{X})}^{\,\mathbf{2}}$ be the equation for the variance. We begin by solving for $\theta\,(\mathbf{X^2})$ as follows: $\theta\,(\mathbf{X^2}) = \mathbf{p}(\mathbf{X} = \mathbf{1})\mathbf{1^2} + \mathbf{p}\,(\mathbf{X} = \mathbf{0})\mathbf{0^2} = \mathbf{p}\,(\mathbf{X} = \mathbf{1}) = \theta$. Following the standard equation for computing variance, $\sigma^2 = \left(\frac{\sum \mathbf{X^2}}{\mathbf{N}}\right) - \mu^2$, keeping in mind that $\mathbf{N}$ is already captured by θ by virtue of it being a probability ($\frac{\sum \mathbf{X}}{\mathbf{N}}$) and $\mu = \mathbf{p}\,(\mathbf{X} = \mathbf{1}) = \theta$, we can compute the variance around θ as: $\mathbf{var}\,(\mathbf{X}) = \theta\mathbf{X^2} - \theta^2 = \theta - \theta^2 = \theta\,(\mathbf{1} - \theta)$. In this final form, we can see that the variance is just a restatement of θ.

median odds ratio, however, estimates a different value, the heterogeneity between people in terms of the odds of occurrence of a binary outcome variable—i.e., the median odds ratio that would be observed if multiple random samples were drawn from the population. After illustrating how to implement and interpret each method in R step-by-step, using the random intercept only model fit as an example above (`logistic_MLM0`, see Code Box 1), we summarize the differences between techniques (see Table 1).

Latent Threshold Approach

The simplest and most popular method of calculating the ICC in a logistic MLM is to assume that the within-cluster variation is equal to the variance of a logistic distribution with a scale parameter equal to one, yielding the value $\frac{\pi^2}{3}$ (Goldstein et al., 2002; Snijders & Bosker, 2011, p. 440). Using this approach then, the ICC is a quantification of the proportion of variance in (latent) logit units attributable to between-person differences.

This method arises from a common formulation or data generation mechanism for binary regression models. We begin by assuming a continuous outcome regression model, i.e. a weighted sum of predictors and an error term.[3] However, the continuous outcome is unavailable, and instead we have a dichotomized version of the variable. Under this formulation, binary regression models attempt to retrieve the parameters or regression coefficients of the underlying continuous variable regression. Since the continuous variable is unavailable, we make certain "identification" assumptions to permit model estimation. For logistic regression, the relevant assumption is that the error term in the continuous variable regression follows a standard logistic distribution i.e. it has mean 0 and variance $\frac{\pi^2}{3}$ (Goldstein et al., 2002).

Applying this formulation to the data at hand, we may assume a continuous distribution of values across participants, and participants with higher values have greater preference for high effort tasks. The exact preference value of a participant is a weighted combination of their predictors and the logistic error term. When this preference value exceeds some threshold, that varies by participant, the participant chooses the high effort task. A random-intercept multilevel logistic regression returns the variation across persons, τ_0^2, and we assume the variation within persons to be fixed at $\frac{\pi^2}{3}$. At which point one can calculate the ICC as a ratio of the variance component of interest and the total variance, $var = \tau_0^2 + \frac{\pi^2}{3}$ (Goldstein et al., 2002).

In R, we can compute the ICC this way as follows (**S4a**):

```
tau20         <- VarCorr(logistic_MLM0)$PID[1]
sigma2        <- pi^2/3
threshold_ICC <- tau20 / (tau20 + sigma2)

threshold_ICC
[1] 0.1867374
```

Code Box 3. Computation of the ICC using the Latent Threshold Method

In the Code Box above, The `VarCorr` command extracts the random variance components from a fitted model (here `logistic_MLM0`, i.e., the value .7554 from Code Box 1) . Alone, this yields a variance component per grouping variable and per number of random variance parameters ($\tau_0^2, \tau_1^2, \tau_2^2, \ldots$) specified in the model. In our case, we have one grouping variable, `PID`, and so we specify that we are interested in this the random intercept variance for this group using `$PID[1]`. We then use this value to compute the latent threshold ICC as described above. Executing the code in Code Box 3 yields an ICC = 0.19—that is, nearly 20% of variance in effort avoidant behaviour is attributable to differences between individuals or, conversely, 80% to within-participant variation.

In contexts where the modeler is primarily interested in a continuous outcome, but only has access to a dichotomized variable, this formulation for the ICC is ideal. For example, if a researcher were interested in a presumably continuous, but inaccessible, construct of effort aversion, then effort-related choices in the DST might be seen as a dichotomous representation of this latent construct. It is worth noting, however, that the assumption that unobservable continuous variables masquerade as dichotomous ones may be untenable in many experimental contexts. While effort choice could be thought to reflect a continuous underlying distribution of preference, some may think of it as a truly dichotomous outcome—a participant chooses either one option or the other—in the same way that a person is either a bachelor or not. In these cases, it may be desirable to assume responses as truly dichotomous when calculating the ICC.

Simulation Approach

As an alternative to the latent threshold approach, Goldstein et al. (2002) recommended a simulation-based approach as a more general means of estimating the ICC on a binary (observed) scale. Differently from the Latent Threshold approach which focuses on the logit scale, this is an ICC measure on the probability scale. The key idea is to compute the Bernoulli variance over a large number of simulated datasets as a proxy for the estimate of residual variance. This computes the ICC from two sets of values: first, the person-level estimated probabilities, which are obtained from the fixed and random effects of a fit model, and second, a measure of within-person variance based

3 This formulation is mathematically equivalent to the GLM formulation for binary regression models based on the Bernoulli distribution, with a linear model for probabilities on an unbounded scale.

on the Bernoulli variance seen above, $P(Y=1)\,(1-P(Y=1))$, which is averaged across people. These two components can then be used to compute the ICC. In practice, this approach is done in four steps.

1. From an already fitted model (`logistic_MLM0`), simulate a large number (M) of normally-distributed participant-level random effects using the fitted model's random intercept variance estimate, $U_{0jm} \sim N\left(0, \tau_0^2\right)$, where m refers to a single simulation, $m = 1 \ldots M$ and j refers to a particular (simulated) participant.
2. For each random effect, compute predicted probabilities for each level 1 observation, $\widehat{p}_{ijm}$, according to the model's fixed effects (γ_{00}) and the random effect obtained in step 1 (hence the m subscript). Using the example model, `logistic_MLM0`, predicted probabilities of selecting the high-effort option on each trial would be computed as follows: $\hat{p}_{ij_m} = \frac{\exp(\gamma_{00}+U_{0jm})}{1+exp(\gamma_{00}+U_{0jm})}$.
3. For each of these predicted probabilities, compute the level 1 variance according to a Bernoulli distribution: $\widehat{V}_{ij_m} = \hat{p}_{ij_m}\left(1-\hat{p}_{ij_m}\right)$.
4. The ICC is then estimated as the ratio of participant-level variance in predicted probabilities over total variance, which itself is composed of cluster-level variance and average level-1 variance: $ICC = \frac{Var_{m=1}^{M}(\widehat{p}_{ij_m})}{Var_{m=1}^{M}(\widehat{p}_{ij_m})+\frac{1}{M}\sum_{m=1}^{M}\widehat{V}_{ij_m}}$.

Implementing this approach in R (**S4b**):

```
# 0. Extract intercept-variance and fixed effects, specify number of simulations,
and set seed

set.seed(2022) # for reproducibility
tau20   <- VarCorr(logistic_MLM0)$PID[1]
gamma00 <- fixef(logistic_MLM0)
M       <- 100000

# 1. Simulate random effects (using the square root of the random variance; i.e.,
the SD)
U0j <- rnorm(M, 0, sqrt(tau20))

# 2. Predict probabilities
logit_p_hat <- gamma00 + U0j
p_hat       <- exp(logit_p_hat) / (1 + exp(logit_p_hat))

# 3. Compute level-1 variance (Bernoulli variance)
var_L1 <- p_hat * (1 - p_hat)

# 4. Compute ICC
sigma2 <- mean(var_L1)
simICC <- var(p_hat) / (var(p_hat) + sigma2)

# print result
simICC
[1] 0.1390484
```

Code Box 4. Computation of the ICC using the Simulation Method

At the start of Code Box 4, we use the `VarCorr` function to extract the between cluster variance, as before (Code Box 3). We then extract the fixed intercept (γ_{00} in Eq. 4) using the `fixef` command and specify the number of simulations, M. We then follow the steps described above to compute the ICC using the simulation method.

Executing the code in Code Box 4 yields an ICC of 0.14—that is, 14% of total variance in effort avoidance owes to differences between individuals. Notably, this estimate is smaller than that obtained by the latent threshold method. In the latent threshold method, the ICC is computed entirely on the logit scale, and the residual variance parameter is assumed known. In the simulation approach, both the between- and within-participant variances are computed on the probability scale—the calculations include the fixed effects. Hence, the fixed effects shift the mean of the probability in ways that affect the variance of probabilities—which in turn increase at a slower rate than the variance on the logit scale. These differences lead to discrepancies in ICC estimates between the methods.

It is worth bearing in mind that because simulated deviations from the fixed effect are based on the fixed effects themselves, estimates of the ICC in models with additional covariates (i.e., not a null model) will depend on the covariate patterns. That being said, differences in the ICC associated with different covariate values may themselves be of interest, though this process increases in complexity as more covariates are added.

Linearization

Instead of assuming a non-linear (logistic or Bernoulli) variance, it is possible to estimate a linear approximation of the logistic multilevel equation and use the variance from this approximation. Thus, linearization attempts to approximate the ICC otherwise obtained via simulation method (e.g., using the Simulation approach)—i.e., one in which variance is calculated on the probability scale. Goldstein et al. (2002) proposed the following approximation for the null model, where $\hat{p}_{ij}$ refers to the estimated probability that $Y_{ij} = 1$.

$$\hat{p}_{ij} = \frac{\exp\left(\gamma_{00}\right)}{1+\exp\left(\gamma_{00}\right)} \qquad \text{(Eq. 5)}$$

This approximation works for the case of the null model specifically (i.e., the model for which we typically compute the ICC), but Goldstein's et al. (2002) approach may also work in the presence of multiple predictors (a topic we return to later). Considering the null model, then the variance is equal to a combination of level-1 and level-2 variance, as:

$$\begin{aligned} \hat{p}_{ij} &= \frac{\exp\left(\gamma_{00}\right)}{1+\exp\left(\gamma_{00}\right)} \\ \hat{V}_1 &= \hat{p}_{ij}\left(1-\hat{p}_{ij}\right) \\ \hat{V}_2 &= \tau^2 \hat{p}_{ij}^2 (1+\exp\left(\gamma_{00}\right))^{-2} \\ ICC &= \frac{\widehat{V}_2}{\hat{V}_1+\hat{V}_2} \end{aligned} \qquad \text{(Eq. 6)}$$

where $\widehat{p}$ reflects the model-estimated probability of choosing the high effort option at the mean of the random intercepts (γ_{00}), and $\widehat{V}_1$ is the Bernoulli implied variance. $\widehat{V}_2$ is the variance of the level-2 outcomes as predicted by the model, including information about each person (τ^2; i.e., the variance in the (null) model-estimated value of choosing the high effort option for each participant). The total estimated variance is the sum of these values. This variance then appears in the denominator of the standard ICC calculation. Note that the equation above is a more general form of the linearization equation, as the present model (`logistic_MLM0`) is a null model and thus only contains one predictor (γ_{00}). In R (**S4c**):

```
# 0. Extract relevant parameters (tau20 and gamma00)
tau20   <- VarCorr(logistic_MLM0)$PID[1]
gamma00 <- fixef(logistic_MLM0)

# 1. Evaluate the probability of success at the mean of random effects (i.e., the
# fixed effect) -> Eq. 5 above
p <- exp(gamma00) / (1 + exp(gamma00))

# 2. Compute the Bernouilli variance of this fixed estimate first (last part of
# equation 6)
var1 <- p * (1 - p)

# 3. Compute the variance in the level-1 outcome (equation 6)
var2 <- tau20 * p^2 * (1 + exp(gamma00))^(-2)

# 4. Compute ICC
linICC <- var2 / (var1 + var2)

# print result
linICC
[1] 0.1551821
```

Code Box 5. Computation of the ICC using the Linearization Method

As before, at the start of Code Box 5, we extract the random effects and fixed effects using the `VarCorr` and `fixef` functions in R. Following this, we follow the steps described above to compute the ICC. Executing Code Box 5 yields an ICC of approximately 16%—closer to the simulation method and lower than the latent threshold approach.

While this linearization method does not require simulation, much like the simulation approach, its value is conditional on the covariate structure. That is, the ICC will be conditional on the values of the predictor variables. We discuss ways around this limitation later, but it is at the discretion of the modeler whether the unique ICCs for different patterns of covariates are of interest over a global ICC estimate.[4] Another consideration is that when τ_0^2 is large, the linearization transformation will be applied to values along a broad span of the logistic curve, which is non-linear and poorly approximated by a linear function. As such, the linearization method may only be appropriate when τ_0^2 is relatively small.

Median Odds Ratio

The methods reviewed so far have aimed to compute the ICC for logistic multilevel models using techniques analogous to linear models. An alternative approach is to abandon the framework of the ICC and instead compute a measure of participant-level variation that more readily captures the binary nature of the data, without the need for an estimate of within-participant variance. That is, a method in which the measure of variation is expressed in terms of odds, which is a more common scale for interpreting binary data models. Accordingly, the median odds ratio (MOR), proposed by Merlo et al. (2006), represents median participant-level variation on the odds ratio scale. Conceptually, the MOR represents the median odds of success across every pairing of participants in the dataset. Imagine that every participant is paired to each other participant in the sample and their odds of choosing the high demand option are calculated according to their participant-level deviation from the fixed intercept. The ratio of these pairs of odds are then computed, with the higher odds always placed in the numerator. Doing this for all pairs of participants would yield a distribution of odds ratios, the median of which would be the MOR. If the MOR is equal to 1, it would suggest no differences between participants (all participants are equally effort-avoidant; that is, they avoid high effort stimulus γ_{00} log-odds of the time). If the MOR is considerably larger than 1, it would suggest sizable individual differences. As such, the MOR is a useful technique for those interested solely in the between-participant effects, but offers no direct extension to quantify within-participant variability.

Practically, it is not necessary to carry out these computations in full, because the MOR can be readily computed using the following equation (see Merlo et al., 2006):

$$MOR = \exp\left(\sqrt{2\tau_0^2}\Phi(0.75)\right) \qquad \text{(Eq. 7)}$$

where Φ is the cumulative distribution function of a standard normal distribution, and $\phi(0.75)$ is roughly equal to 0.67 (see Larsen & Merlo, 2005 for a derivation of this equation). In R, the MOR can be computed as follows (**S4d**):

```
tau20 <- VarCorr(logistic_MLM0)$PID[1]
phi75 <- qnorm(.75) # 75th percentile of normal CDF
MOR   <- exp(sqrt(2 * tau20) * phi75)

# print result
MOR
[1] 2.291137
```

Code Box 6. Computation of the MOR.

Using the data at hand, the MOR is 2.29, which means that the median odds of avoiding the high-demand option increased by 2.29 times when randomly comparing two participants. In other words, the odds of effort avoidance in the present sample can vary, in median, by 2.29. Relative to the original estimate of between-participant variability from the output of fitting `logistic_MLM0` (τ_0^2), the MOR reflects a summary of how variation in effort aversion when comparing random subsets of participants, whereas τ_0^2 is an estimate of average variability in effort aversion across participants.

4 Nakagawa et al. (2017; 2013) have proposed a similar method for estimating the coefficient of determination (R^2) in generalized multilevel models. In the case of binary outcomes, this method estimates the proportion of variance explained in a model. This value is proportional to the ICC when considering null models, as we do here (Snijders & Bosker, 2011), and in fact the latent threshold ICC is exactly equal to the conditional R-squared value, assuming a null model. Accordingly, we do not review it in detail here, but point readers to Nakagawa's (2013; 2017) work on topic. Notably, one can estimate R^2 for models with other families of multilevel regression than binomial (e.g., Poisson), which may be of interest to some readers (see for instance the MuMIn package in R for an implementation of this method).

Table 1. Advantages and Disadvantages of Different Techniques for Computing ICC in Logistic MLM.

Method	Estimated ICC for `logistic_MLM0`	Pros	Cons
Latent Threshold	0.19	• Simple • Popular • Applicable to all models • Ideal when goal is to model the latent continuous variable	• Unintelligible estimate when latent continuous variable assumption is not tenable
Simulation Goldstein et al. (2002)	0.14	• Provides direct estimate of within-participant variance • Estimate should be meaningful for most applications	• Cannot be computed by hand • Estimates depend on covariate structure
Linearization (Goldstein et al., 2002)	0.16	• Closed-form solution—no simulation required • Provides estimate of (conditional) within-participant variance • Estimate should be meaningful for most applications	• Estimates depend on covariate structure • Is inappropriate when random intercept variance is high
Median Odds Ratio (Merlo et al., 2006)	2.29 (in median odds scale)	• Readily interpretable on the same scale as odds ratios • Closed-form equation	• Not expressed as a proportion, like other measures of between-participant variance • No within-participant variance term

While the MOR is unlike the other approaches discussed thus far, it is worth emphasizing its conceptual similarity to the ICC (and thus its inclusion in this tutorial). Much like the ICC, the MOR provides information about individual variability between participants impacts the odds of the outcome variable being equal to 1 (in this case, choosing the high effort option). Put simply, both the MOR and ICC quantify how *important* the participant is in determining the outcome. Moreover, like other ICC values (and holding total variance constant), larger MOR values imply reduced within-person variability, and vice-versa. In this respect, the MOR and the ICC convey similar information. Nevertheless, unlike the ICC, the MOR is not expressed as a proportion. As such, this method offers no quantification of residual, within-participant, variance, which could be of interest to experimental psychologists.

6. Estimating Uncertainty about the ICC With Bootstrapping

In the previous section, we outlined four methods for quantifying variation in logistic multilevel modeling, as well as some relative advantages and disadvantages of each method. However, quantifying uncertainty around any given estimate of the ICC, which in our in our example varies between 15-20% depending on the method, is not routinely discussed in the literature.

To quantify uncertainty in the ICC, Austin and Leckie (2020) proposed using a parametric bootstrapping process with percentile confidence intervals. In its simplest form, this process unfolds in three steps:

1. Simulate many datasets from the original logistic model, randomly sampling random effects from a normal distribution centered at the estimated mean and variance from a fit model.
2. Compute the quantity of the interest (the ICC or residual variance measure in this case).
3. Summarize these data (e.g., by taking the mean, computing a 95% interval, etc.) to make statistical inferences using the resulting sampling distribution.

```
bootstrap_MOR <- bootstrap_icc(logistic_MLM0, gr='PID', method='icc_thr', B = 100)

quantile(bootstrap_MOR, c(.025, .975))

2.5%    97.5%
0.1089664 0.2576925
```

Code Box 7. Example of bootstrapping procedure to obtain sampling distributions of the ICC, using the latent threshold method.

In principle it would be possible to hand-code a bootstrapping procedure to estimate these values using the information provided in Code Boxes throughout this tutorial (and, for the intrepid reader, we provide code to do so at https://anonymous.4open.science/r/logisticicc-F6C8/ and **S5b**). For ease however, we provide a function that draws bootstrapped samples of the ICC using the methods specified above. Below, we use this function to obtain 100 bootstrapped estimates of the ICC, using the latent threshold method (**S5a**).

`bootstrap_icc()` takes as arguments a model, here, `logisticMLM0`, a grouping variable, here `PID`, a string representing one of the methods for computing the ICC described above, here `icc_thre`, and the number of samples to draw, `B`. Applying this yields a bootstrapped sampling distribution that characterizes uncertainty around the point estimate. For instance, while the point estimate for the latent threshold ICC in the present data was 0.19, the bootstrapped confidence interval suggests the ICC estimates most compati-

Table 2. 95% Intervals for Bootstrapped Statistics (100 iterations)

Variable	Quantile	
	2.5%	97.5%
ICC (Latent Threshold)	0.11	0.26
ICC (Simulation)	0.11	0.23
ICC (Linearization)	0.10	0.22
MOR	1.83	2.77

ble with the current data range between 0.11 and 0.26 (see Table 2). In other words, anywhere between approximately a tenth to a quarter of the variance in effort avoidant behaviour is attributable to differences between individuals.

Notably, unlike confidence intervals calculated using standard errors, the lower bound of these bootstrapped confidence intervals will never go below 0 in the case of the ICC, or 1 in the case of an MOR, because these are the lowest possible values these metrics can take. Accordingly, 1 for the MOR and 0 for the ICC might be very conservative thresholds for minimal participant-level variation. The advantage of the bootstrap approach is that researchers can set their own lower bound for what constitutes minimal heterogeneity in the outcome measure and explore the 95% interval relative to this threshold.

The 95% interval of bootstrapped distributions for MOR, and ICC using all of the methods discussed thus far from 100 samples are summarized in Table 2. Regardless of the measure, there are substantial individual differences in effort avoidant behaviour, suggesting that not all people are equally averse to effort action—a finding echoed by recent work exploring individual differences in cognitive effort investment, which speaks to the theoretical insights examining the random effects can provide (Otto & Daw, 2019; Sandra & Otto, 2018).

7. Computing the ICC and MOR in Models with Predictors and Random Slopes

Thus far, the tutorial has focused on a null model with no predictors and random intercepts only. We recommend this as a first step to modeling building because it provides a starting point to understanding the variability in the outcome. Researchers are then likely to proceed to adding predictors and estimating random slopes. In this section we review how to use the methods we have discussed with more complex models.

Models with Within-Participant Predictors

First, we demonstrate how to estimate a model with a fixed within-participant predictor, trial number, and random intercepts. Code Box 8 shows the regression equation now includes `trial0`, which captures the number of trials

```
logistic_MLM1 = glmer(effort_choice ~ trial0 + (1|PID), data=data.Ctl,
family='binomial')

summary(logistic_MLM1)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: effort_choice ~ trial0 + (1 | PID)
   Data: data.Ctl

     AIC      BIC   logLik deviance df.resid
 13822.0  13843.9  -6908.0  13816.0    10906

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7442 -0.8961 -0.3605  0.9873  4.8429

Random effects:
 Groups Name        Variance Std.Dev.
 PID    (Intercept) 0.778    0.882
Number of obs: 10909, groups:  PID, 37

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.33826    0.14674  -2.305   0.0212 *
trial0      -0.06576    0.03624  -1.814   0.0696 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
       (Intr)
trial0 -0.002
```

Code Box 8. Multilevel logistic model with one predictor, random intercepts, and fixed slopes.

that have elapsed since the beginning of a block in the DST (where a block is a subset of trials). The predictor has been centered at the median and scaled to be between 0 and 1. In line with past work (Kool et al., 2010), we consider whether participants' desire to avoid higher effort tasked increased as their exposure to the experimental tasks increased. The fixed intercept, γ_{00}, reflects the average log-odds of selecting the high-effort tasks when `trial0` is equal to zero for the typical participant, and the fixed slope, γ_{10}, reflects the change in log-odds from the beginning to the end of the task for the typical participant (from when `trial0` equals 0 to when it equals 1). As before, we have specified that the intercept will vary per participant in a normally distributed fashion, with random variance τ_0^2.

The simulation and linearization approaches for computing the ICC described above rely on an estimate of intercept variance (τ_0^2), which depends on the value of the intercept in a model, which in turn depends on the covariate structure. For these approaches, if researchers are interested in various points, (e.g., when covariate 1 is high and covariate 2 is low, when both covariates are high, when both are low, and so on), they need to calculate multiple ICCs. In Figure 3A, we plot the ICC at each value of `trial0` using the linearization method, which demonstrates how the ICCs changes as we move along different values of the covariate. Conversely, both the ICC computed using the latent threshold approach and the MOR implicitly assume fixed residual variance and will be assumed to be appropriate for any covariate structure. In the presence of covariates, the latent threshold ICC and MOR return values that are adjusted for the covariates. For example, the latent threshold ICC functions similarly to the *residual ICC* in linear multilevel models in that it is the latent variance attributable

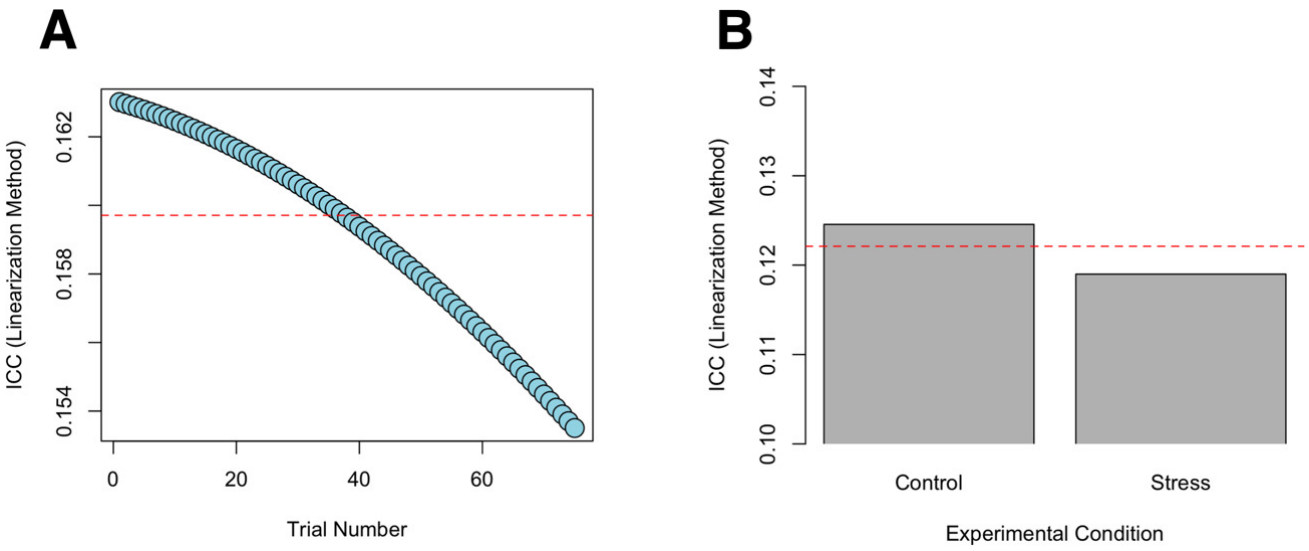

Figure 3. Linearization-based ICC in models where (A) within-person and (B) between-person predictors are included.

In A, the x-axis represents the trial number at which the ICC was computed. In B, the x-axis represents the experimental condition at which the ICC was computed. The y-axis shows the resultant estimated ICC. The red dotted line shows the estimated ICC using the averaging method discussed in the main text. Together, these figures illustrate the problems that accompany estimating the ICC when predictors are included in a logistic multilevel model.

to differences between individuals after accounting for the variance attributable to covariates.[5]

Another alternative to proliferation of simulation and linearization ICC values with increasingly complex covariate structures is to calculate the ICC at the *average* prediction of the observed data on the logit scale. To do so, the average predicted value from the model is used in place of the combination of fixed effects. The benefit of this method is that it is simpler while the cost is that it cannot capture the relative contribution of within- vs. between-participant variance at different levels of the predictor. Thus, faced with a model like our `logisticMLM1` we must decide if we want to compute the ICC at different levels of the predictor—at the beginning vs. end of a block, for instance—or compute a global ICC at the average—here, the middle of a block. We take the latter approach by centering the predictor at the median, but the former approach could be taken if `trial0` was recentered to the beginning or end of the task.

As an example, we can leverage the averaging approach described above for the linearization method described in Section 5 to compute a single ICC for `logisticMLM1` estimated above, which yields an estimate of the ICC that is close to the estimate for the intercept-only model in Section 5 ($ICC_{average}$ = 0.1554; ICC_{null} = 0.1552). As mentioned however, this estimate of the ICC will only be valid for the middle of the block (see Figure 3A). This is implemented in the `icc_lin()` function provided in the Supplemental Materials (**S6d;** see Code Box 9).

```
icc_lin(logistic_MLM1, 'PID', avg_preds = T)

[1] 0.1558737
```

Code Box 9. Simulation method for a model with within-person predictors

As an aside, in the case of the simulation method, this averaging approach can be extended via numerical integration to estimate the moments of the logit-normal distribution, for which, for example, the simulation method can be thought of as a method for approximating. This yields numerically very similar estimates of the ICC without the need for simulation. We describe this *Numerical Integration Approach* in the Supplemental Materials (**S6d**) and provide a separate function for it.

Models with Between-Person Predictors

The same issues apply when a level 2, in our example a between person, predictor is included. In Code Box 10, we fit a logistic multilevel model where effort choices are predicted by experimental condition: control (0) or stress (1). The original purpose of Bogdanov and colleagues' (2021)

5 Since the residual variance is fixed across logistic models, estimated model parameters are automatically scaled across models to reflect changes in residual variance across models (an issue referred to as non-collapsibility or unobserved heterogeneity depending on the literature; Greenland et al., 1999; Mood, 2010). For example, if one extends a random-intercept model by including within-participant predictors that do not explain any participant level variance (e.g. person-mean centered predictors), the intercept variance (τ_0^2) would increase from the random-intercept, producing a larger latent threshold ICC or MOR since the residual variance *has reduced* from the earlier model.

```
logistic_MLM2 <- glmer(effort_choice ~ condition + (1|PID), data=data,
family='binomial')

summary(logistic_MLM2)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: effort_choice ~ condition + (1 | PID)
   Data: data

     AIC      BIC   logLik deviance df.resid
 28492.0  28516.1 -14243.0  28486.0    22475

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.2636 -0.8786 -0.4627  0.9761  4.5435

Random effects:
 Groups Name        Variance Std.Dev.
 PID    (Intercept) 0.5838   0.7641
Number of obs: 22478, groups:  PID, 38

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.44017    0.12483  -3.526 0.000422 ***
condition1   0.12024    0.01417   8.484  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr)
condition1 -0.002
```

Code Box 10. Multilevel logistic model with one between-person predictor and random intercepts

```
icc_lin(logistic_MLM2, 'PID', avg_preds=T)

[1] 0.1221064
```

Code Box 11. Linearization method for a model with between-person predictors

study was to examine the role that acute stress played on effort avoidance. The model below tests this prediction, examining average effort avoidance rates depending on if participants were stressed or not before completing the DST.

Just as with the continuous predictor example above, the intercept and the ICC using the simulation and linearization approaches will depend on the value of `Condition`. We illustrate this in Figure 3B. In Code Box 11, we compute the averaged ICC using the linearization method. As before, we obtain a value not too far from, though slightly lower than, the linearized ICC computed for `logisticMLM_0` (ICC_{null} = 0.16).

As before, these issues do not apply to the latent threshold ICC nor the MOR, which will be identical regardless of the covariate pattern.

Notably, while continuous predictors may yield an unwieldy number of possible ICCs, categorical variables may have very restricted covariate patterns. As a result, researchers may in fact be interested to examine how person-level variability in the intercept (indexed by the ICC) varies per categorical conditions. Moreover, when a model includes categorical predictors, the average prediction may not reflect any actual case in the data, such it may be wiser to consider the ICC at specific values of the categorical predictors.

```
logistic_MLM2 = glmer(effort_choice ~ trial0 + (trial0|PID), data=data.Ctl,
family='binomial')
summary(logistic_MLM2)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: effort_choice ~ trial0 + (trial0 | PID)
   Data: data.Ctl

     AIC      BIC   logLik deviance df.resid
 13803.0  13839.5  -6896.5  13793.0    10904

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7942 -0.8994 -0.3641  0.9937  4.9218

Random effects:
 Groups Name        Variance Std.Dev. Corr
 PID    (Intercept) 0.78982  0.8887
        trial0      0.07572  0.2752   -0.39
Number of obs: 10909, groups:  PID, 37

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.33915    0.14781  -2.295   0.0218 *
trial0      -0.04598    0.05958  -0.772   0.4402
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
       (Intr)
trial0 -0.297
```

Code Box 12. Multilevel logistic model with one predictor, random intercepts, and random slopes.

Models with Random Slopes

Finally, a modeler may be interested in estimating how much the influence of a level 1 predictor varies per participant. For example, we might expect that over the course of a block, participants become increasingly effort averse, but that some participants remain steadfast in their initial rate of effort aversion while others rapidly reject all effortful tasks after only a few trials. This type of conditional between-participant variability can be modeled using random slopes. In Code Box 12 we estimate `logistic_MLM2` by adding (`trial0 | PID`) to estimate the slope variance across people. As can be seen in Code Box 12, a new random effects variance is included, which encodes how variable different participants' effort preferences are to the trial number of the task.

Though adding random slopes to the model and computing the ICCs in the manner we have demonstrated is relatively straightforward in *lme4*, there is debate in the literature about whether the ICC should be computed for models that contain random slopes. For instance, Kreft & De Leeuw (1998) state that "The concept of intra-class correlation is based on a model with a random intercept only. No unique intra-class correlation can be calculated when a random slope is present in the model." (p. 63). Conversely, Goldstein et al. (2002) highlight that it is, in theory, possible to estimate the ICC for a model with random slopes, but that doing so will be conditional on the pattern of covariates and that doing so changes the interpretation of the ICC—and its components: between- and within-participant variation—to a point that it no longer serves the same purpose of assessing the relative contribution of between- versus within-participant variation to total variance. We will not settle this debate in this tutorial and to our knowledge, no work exists that tackles this issue in logistic multilevel

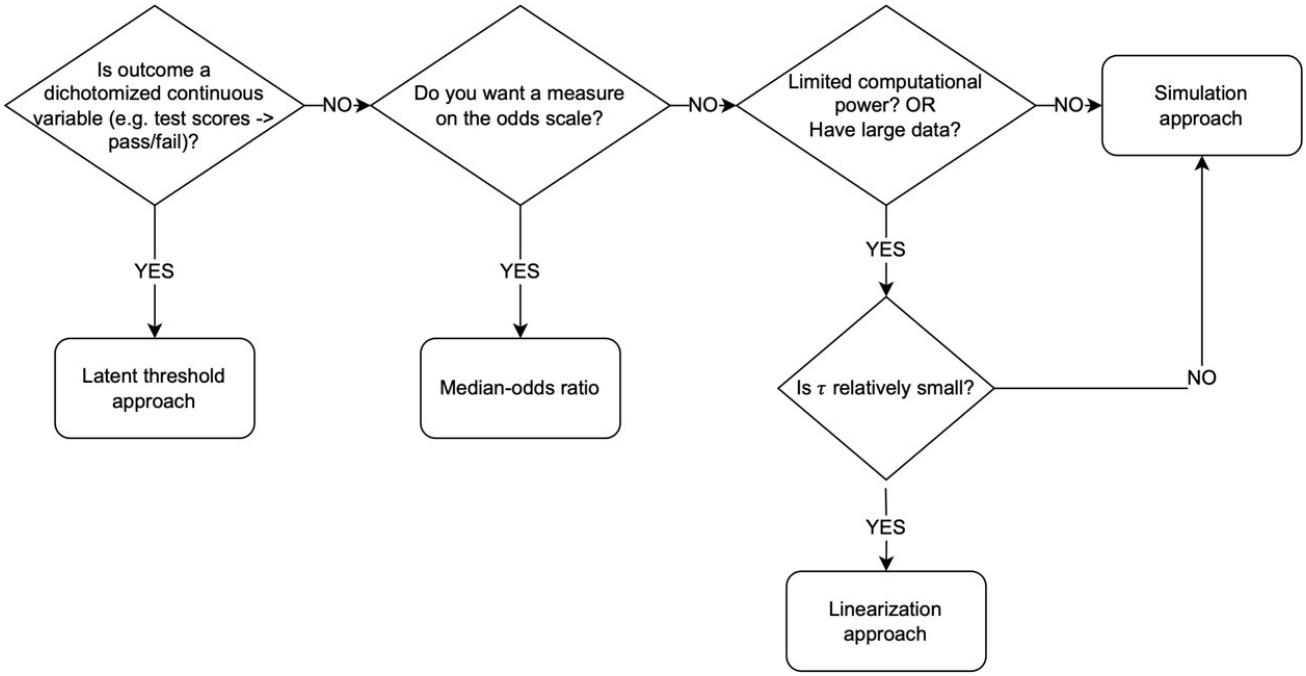

Figure 4. Article Flowchart.

Readers are encouraged to use this flowchart to direct and their selection of an estimate of ICC (and MOR) for logistic multilevel modeling.

models. For now, we recommend implementing the ICC using the techniques described above for models with random intercepts, and we caution that the interpretations we've provided do not hold for models with random slopes.

8. Conclusion

Calculating the ICC to describe the effect of clustering in the dependent variable is a best practice for the early stages of multilevel model building. However, for logistic models, this is not straightforward to do, with popular programs not outputting the necessary estimates. In this tutorial, we have reviewed four methods for computing estimates of between- and within-participant variability from logistic multilevel models and demonstrated how to calculate the ICC. We have highlighted relative advantages and disadvantages of each technique and provided R code to compute these estimates, both manually (using didactic code from the Code Boxes or the supplement to reproduce all analyses reported herein) or functionally (using the function code provided in the repository associated to this paper and described in **S6d**). Finally, we reviewed bootstrapping techniques to quantify uncertainty about these variance estimates.

Here, we focused on modeling binary data using a model with a logit link function, alternative link functions, like the probit function, are commonly employed. The techniques to quantify the ICC (excluding the MOR, because it is in odds ratios) described could be used for a probit model. That is, in a probit model, the value of the residual variance in the ICC using the threshold approach would be 1, rather than $\frac{\pi^2}{3}$, in the simulation approach, the mean function from inverse-logit function would change to a normal cumulative density function, and for the linearization approach, the formula presented in equations 5 and 6 would also change.

At this point, readers may ask which method is the best overall, or which method they should use for a specific scenario. To help arbitrate between the different methods presented here, we point readers to Table 1 and Figure 4 for practical considerations. Additionally, the decision between methods might also be made on theoretical grounds. If researchers view the observed binary outcome as the result of a measurement process that dichotomized a continuous variable/trait, then the latent threshold approach is appropriate. If they are instead concerned only in binary preference, the linearization or simulation methods should be considered. Finally, if researchers wish to express between-person heterogeneity in terms of odds, the MOR is a good option.

Finally, another way to identify which measure to report is to consider the type of effect size the researcher plans to report. If the researchers' primarily focus on log-odds and odds-ratios, then the latent approach and MOR are adequate, because they are already expressed in that scale. If the researchers are instead interested in transforming effects into probabilities or attempt to communicate effects on the probability scale (i.e., on the observed scale), then the simulation and linearization approaches may be more appropriate.

Overall, making these approaches easier to implement will enable researchers to investigate variability in their binary data and expand their theoretical considerations beyond effects in the aggregate.

Funding

This work was supported by a fellowship (awarded to S.D.) from the Natural Sciences and Engineering Research Council of Canada.

Author Contributions

S.D. conceived of the initial research question, reviewed literature, and wrote the initial draft of the tutorial. J.O.U. provided expert commentary throughout the writing and coding process. A.R.O. provided critical feedback throughout the writing process. J.K.F. helped write and edit the manuscript and supervised the project.

Competing Interests

The authors declare no conflict of interest.

Data and Materials Availability

All code and data used for this tutorial can be found at https://github.com/seandamiandevine/logisticicc.

Submitted: April 03, 2023 PDT, Accepted: December 19, 2023 PDT

References

Austin, P. C., & Leckie, G. (2020). Bootstrapped inference for variance parameters, measures of heterogeneity and random effects in multilevel logistic regression models. *Journal of Statistical Computation and Simulation*, *90*(17), 3175–3199. https://doi.org/10.1080/00949655.2020.1797738

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Statistics in Medicine*, *36*(20), 3257–3277. https://doi.org/10.1002/sim.7336

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *Fitting linear mixed-effects models using lme4*. https://doi.org/10.48550/arXiv.1406.5823

Bogdanov, M., Nitschke, J. P., LoParco, S., Bartz, J. A., & Otto, A. R. (2021). Acute Psychosocial Stress Increases Cognitive-Effort Avoidance. *Psychological Science*, *32*(9), 1463–1475. https://doi.org/10.1177/0956797621100546 5

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(4), 223–231. https://doi.org/10.1207/s15328031us0104_02

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*(1), 29–46. https://doi.org/10.1214/ss/1009211805

Hardin, J. W., & Hilbe, J. M. (2014). Generalized estimating equations: Introduction. *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat06899

Huang, F. L. (2018). Multilevel modeling myths. *School Psychology Quarterly*, *33*(3), 492. https://doi.org/10.1037/spq0000272

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, *2*(12), 899–908. https://doi.org/10.1038/s41562-018-0401-9

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665. https://doi.org/10.1037/a0020198

Kreft, I. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Sage. https://doi.org/10.4135/9781849209366

Larsen, K., & Merlo, J. (2005). Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *American Journal of Epidemiology*, *161*(1), 81–88. https://doi.org/10.1093/aje/kwi017

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 years. *Review of Educational Research*, *91*(3), 311–355. https://doi.org/10.3102/0034654321991229

McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly*, *54*(2), 152–155. https://doi.org/10.1177/0016986210363076

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., Råstam, L., & Larsen, K. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology & Community Health*, *60*(4), 290–297. https://doi.org/10.1136/jech.2004.029454

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, *123*, 92–105. https://doi.org/10.1016/j.neuropsychologia.2018.05.006

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). The transdiagnostic structure of mental effort avoidance. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-018-37802-1

Sandra, D. A., & Otto, A. R. (2018). Cognitive capacity limitations and Need for Cognition differentially predict reward-induced cognitive effort expenditure. *Cognition*, *172*, 101–106. https://doi.org/10.1016/j.cognition.2017.12.004

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publishing.

Volpert-Esmond, H. I., Merkle, E. C., Levsen, M. P., Ito, T. A., & Bartholow, B. D. (2018). Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. *Psychophysiology*, *55*(5), e13044. https://doi.org/10.1111/psyp.13044

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/94263-approaches-for-quantifying-the-icc-in-multilevel-logistic-models-a-didactic-demonstration/attachment/197237.docx?auth_token=ki6On9Z8Rb3B3QWp6p1W

Supplemental Material

Download: https://collabra.scholasticahq.com/article/94263-approaches-for-quantifying-the-icc-in-multilevel-logistic-models-a-didactic-demonstration/attachment/197238.pdf?auth_token=ki6On9Z8Rb3B3QWp6p1W

Cover Letter

Download: https://collabra.scholasticahq.com/article/94263-approaches-for-quantifying-the-icc-in-multilevel-logistic-models-a-didactic-demonstration/attachment/197239.pdf?auth_token=ki6On9Z8Rb3B3QWp6p1W