A practical guide and power analysis for GLMMs: Detecting among treatment variation in random effects

Morgan P Kain, Ben M Bolker, Michael W McCoy

In ecology and evolution GLMMs are becoming increasingly used to test for differences in variation by treatment at multiple hierarchical levels. Yet, the specific sampling schemes that optimize the power of an experiment to detect differences in random effects by treatment/group remain unknown. In this paper we develop a blueprint for conducting power analyses for GLMMs focusing on detecting differences in variance by treatment. We present parameterization and power analyses for random-intercepts and random-slopes GLMMs because of their generality as focal parameters for most applications and because of their immediate applicability to emerging questions in the field of behavioral ecology. We focus on the extreme case of hierarchically structured binomial data, though the framework presented here generalizes easily to any error distribution model. First, we determine the optimal ratio of individuals to repeated measures within individuals that maximizes power to detect differences by treatment in among-individual variation in intercept, among-individual variation in slope, and within-individual variation in intercept. Second, we explore how power to detect differences in target variance parameters is affected by total variation. Our results indicate heterogeneity in power across ratios of individuals to repeated measures with an optimal ratio determined by both the target variance parameter and total sample size. Additionally, power to detect each variance parameter was low overall (in most cases >1,000 total observations per treatment needed to achieve 80% power) and decreased with increasing variance in non-target random effects. With growing interest in variance as the parameter of inquiry, these power analyses provide a crucial component for designing experiments focused on detecting differences in variance. We hope to inspire novel experimental designs in ecology and evolution investigating the causes and implications of individual-level phenotypic variance, such as the adaptive significance of within-individual variation.

Title: A practical guide and power analysis for GLMMs: Detecting among treatment variation in random effects

Authors: Morgan P. Kain[1*], Ben M. Bolker[2], Michael W. McCoy[1]

[1]Department of Biology East Carolina University, N108 Howell Science Complex, Greenville, NC 27858
[2] Department of Mathematics & Statistics, McMaster University, 314 Hamilton Hall, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada

[*] Correspondence author.
N108 Howell Science Complex
East Carolina University
Mail Stop 551
Greenville, NC 27858
E-mail: morganpkain@gmail.com

Abstract

In ecology and evolution GLMMs are becoming increasingly used to test for differences in variation by treatment at multiple hierarchical levels. Yet, the specific sampling schemes that optimize the power of an experiment to detect differences in random effects by treatment/group remain unknown. In this paper we develop a blueprint for conducting power analyses for GLMMs focusing on detecting differences in variance by treatment. We present parameterization and power analyses for random-intercepts and random-slopes GLMMs because of their generality as focal parameters for most applications and because of their immediate applicability to emerging questions in the field of behavioral ecology. We focus on the extreme case of hierarchically structured binomial data, though the framework presented here generalizes easily to any error distribution model. First, we determine the optimal ratio of individuals to repeated measures within individuals that maximizes power to detect differences by treatment in among-individual variation in intercept, among-individual variation in slope, and within-individual variation in intercept. Second, we explore how power to detect differences in target variance parameters is affected by total variation. Our results indicate heterogeneity in power across ratios of individuals to repeated measures with an optimal ratio determined by both the target variance parameter and total sample size. Additionally, power to detect each variance parameter was low overall (in most cases >1,000 total observations per treatment needed to achieve 80% power) and decreased with increasing variance in non-target random effects. With growing interest in variance as the parameter of inquiry, these power analyses provide a crucial component for designing experiments focused on detecting differences in variance. We hope to inspire novel experimental designs in ecology and evolution investigating the causes and implications of individual-level phenotypic variance, such as the adaptive significance of within-individual variation.

Key-words: individual variation, behavioral ecology, reaction norm, plasticity, binomial distribution, hierarchical, sampling scheme

Introduction

Recent advances in computing power and access to increasingly sophisticated statistical tools such as generalized linear mixed effects models are changing research in ecology, evolution and behavior. Research questions and data analyses are no longer confined to the assumptions of clean experimental designs based on agricultural plots and Normal error distributions. Researchers now commonly incorporate multiple levels of hierarchical nesting (e.g. repeated measures) and can analyze data using a wide array of non-Gaussian error distribution models. This change is epitomized by the recent increase in use of linear and generalized linear mixed models ([G]LMMs: Touchon, J. & McCoy, W.M. unpublished data). These powerful tools permit appropriate modeling of variation among groups and across space and time, allowing for more accurate extrapolation of statistical results to unobserved data, as well as statistical tests of variance components (Gelman & Hill, 2006; Bolker et al., 2009; Zuur et al., 2009; Zuur, Hilbe & Leno, 2013).

The upsurge in the use of LMM and GLMM has been facilitated by several recent methods papers (Bolker et al., 2009; Martin et al., 2011; Dingemanse & Dochtermann, 2013; Schielzeth & Nakagawa, 2013) and textbooks (Gelman & Hill, 2006; Zuur et al., 2009; Zuur, Hilbe & Leno, 2013; Bolker, 2015) specifically aimed at non-statisticians. While these resources have accelerated the adoption of these tools, there are still too few resources guiding researchers through the choices that must be made *prior to* the initiation of a new experiment, such as the sampling scheme that will optimize the power of an experiment requiring analysis by linear (Moineddin, Matheson & Glazier, 2007; Scherbaum & Ferreter, 2009; Martin et al., 2011) and generalized linear (Johnson et al., 2014) mixed models. In this paper, we develop a blueprint for conducting power analyses

for GLMMs using the `lme4` package (Bates et al., 2014) in the R statistical programming environment (R Development Core Team, 2015). We focus on a specific application aimed at detecting differences in variance by treatment at multiple hierarchical levels.

Power analysis is fundamental to good experimental design, but is often overlooked (Jennions & Moller, 2003), or in the case of GLMMs, simply too difficult to implement for many practitioners. Power analyses can be especially daunting for GLMMs because they require large simulations with complex, non-Normal and non-independent data structures (Johnson et al., 2014). In this paper we take advantage of recent developments in the `lme4` package in R that simplify the process of simulating appropriate data. Despite the increasing use of GLMMs in ecology and evolution and growing interest in variance, we are aware of no papers that present power analyses for statistical tests on variance using GLMMs, and only one paper presenting power analyses for fixed effects in GLMMs (Johnson et al., 2014). Indeed, Johnson et al.'s (2014) analysis illustrates that power analyses conducted for hierarchically structured experiments that do not incorporate random effects can generate biased estimates of fixed effects, highlighting the need for a better understanding of these approaches.

While most applications of GLMMs to date have focused on detecting differences in fixed effects while appropriately accounting for random effects (e.g. Johnson et al., 2014), GLMMs are under rapid development and many new applications are now possible (e.g. modeling heterogeneous error variance: Kizilkaya & Tempelman 2005, Cernicchiaro et al., 2013). With growing interest in variance as the parameter of inquiry (Moore, Brodie & Wolf, 1997; Lynch & Walsh, 1998; Benedetti-Cecchi, 2003; Hill & Zhang, 2004; Nussey, Wilson & Brommer, 2007; Dingemanse et al., 2010; Tonsor,

Elnaccash & Scheiner, 2013; Westneat, Wright & Dingemanse, 2014), there is an increased need for accessible, flexible simulation-based power analyses that assess power to detect differences in random effects by treatment—the magnitude of variation present among repeated measures at a specific hierarchical level (Gelman & Hill, 2006; Zuur et al., 2009).

Here we present parameterization and power analyses for random-intercepts and random-slopes GLMMs that test for differences in variation by treatment in three key parameters: 1) Among-group variation in intercept; 2) Within-group variation in intercept; 3) Among-group variation in slope. We examine each of these comparisons in two contexts. First, we describe the optimal ratio of groups to observations within groups that maximizes power to detect differences in each variance parameter. In experiments with binomially distributed response variables, observations within groups are organized into j sampling occasions, each containing n Bernoulli observations. Here we discuss the ratio of groups to total observations within groups ($n*j$), and consider different partitions of n and j. Second, we explore how power to detect differences in specific variance parameters is affected by increasing variation in non-target parameters (e.g., how power to detect differences in among-group variation decreases as within-group variance increases). We consider both random-intercepts and random-slopes models because of their generality as focal parameters for most applications, and choose to focus on the extreme case of hierarchically structured binomial data because binary response data (e.g. the presence or absence of a behavior) contains the least possible amount of information per observation and yet is a common data format for a variety of endpoints measured in ecology.

We use vocabulary and examples from behavioral ecology to illustrate our models because of their immediate applicability to emerging questions in this field. Specifically, we evaluate power to detect significant differences in among-individual variation in reaction norm intercept and slope, and within-individual variation in intercept between groups of individuals (Nussey, Wilson & Brommer, 2007; Dingemanse et al., 2010). Our methods extend current approaches used in behavioral ecology for quantifying among-individual variation away from simply testing whether there is significant deviation from a null model of no variation (Martin et al., 2011; Van de Pol et al., 2012; Dingemanse & Dochtermann, 2013) toward quantifying and contrasting the magnitude of among- and within-individual variation among multiple groups of individuals.

In an effort to present a framework that is customizable for a diversity of research problems, we focus on a general sampling scheme in which several Bernoulli observations ($n > 1$) within multiple sampling occasions ($j > 1$) are available for each individual. Under this sampling scheme multiple probabilities of "success" (e.g. the probability of displaying a behavior) are available for each individual, which is necessary for quantifying within-individual variation (variation among sampling occasions in the probability an individual displays a behavior). However, we note that often in behavioral ecology only a single Bernoulli observation ($n = 1$) is available for each sampling occasion j. We include a description on how to modify this general case to accommodate single observations per sampling occasion in Supplement 1. Finally, while we focus on the binomial GLMM, the framework presented here generalizes easily to other error distribution models such as Normal, log-Normal, or Gamma (for continuous responses) or Poisson or negative binomial (for count responses).

Methods

Linear Mixed Model

We begin by introducing a general linear mixed model (LMM) to illustrate the variance components we are interested in (Figure 1) and their applications in behavioral ecology. We provide only a brief introduction to LMMs here because they have been extensively discussed in several recent reviews and textbooks (Gelman & Hill, 2006; Zuur et al., 2009; Stroup, 2012; Zuur, Hilbe & Leno, 2013; Dingemanse & Dochtermann, 2013; Bates et al., 2014; Bolker, 2015). We use the notation of Stroup (2012) to facilitate a transition to the binomial GLMM model, which is the focus of our power analyses. A two treatment linear mixed model can be written as:

[1] $y_{ijk} \mid b_{0ik}, b_{1ik} \sim \text{Normal}(\mu_{ijk}, \sigma^2_{\varepsilon k})$

[2] $\eta_{ijk} = \beta_{0k} + b_{0ik} + (\beta_{1k} + b_{ik})X_{ij}$

[3] Identity link: $\eta_{ijk} = \mu_{ijk}$

[4] $\begin{bmatrix} b_{0ik} \\ b_{1ik} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{0k} & \sigma_{01k} \\ \sigma_{01k} & \sigma^2_{1k} \end{bmatrix}\right)$

Here, a single phenotypic measurement y_{ijk} of individual *i,* in environment *j* and treatment *k* is composed of three components: the treatment mean in environment *j* ($\beta_{0k} + \beta_{1k} X_{ij}$), the unique average response of individual *i* across the environmental gradient ($b_{0k} + b_{1k} X_{ij}$), and a residual error due to the variation around the mean of individual *i* ($\sigma^2_{\varepsilon k}$), which is assumed to be homogenous across *X* and among all individuals in treatment *k,* but is allowed to vary by treatment. Individuals vary from the treatment mean reaction norm in both their intercept (b_{0ik}) and slope (b_{1ik}), which together compose the total phenotypic variance attributable to among-individual variation. This individual

116 contribution is quantified using a random intercepts and slopes model with a multivariate
117 Normal (MVN) distribution [4]. Variation among individuals in intercept and slope are
118 σ^2_{0k} and σ^2_{1k} respectively; covariance between intercept and slope is given by σ_{01k}. In a
119 LMM, the linear predictor directly predicts the mean, as shown by the identity link
120 function in equation [3]. In a GLMM, the linear predictor predicts a function of the mean
121 $g(x)$, which must be linearized through the use of non-identity link functions; for
122 example, we use the standard logit (log-odds) link for Binomial GLMM.
123
124 *Among-individual variation in intercept*
125 In behavioral ecology among-individual variation in intercept σ^2_{0k} describes the
126 amount of variation around average behavior that occurs among individuals (Figure 1). In
127 field studies, σ^2_{0k} describes variation in individuals' average behavior in the mean-
128 centered environment (Nussey, Wilson & Brommer, 2007; Westneat et al., 2011).
129 Previous work has demonstrated that individuals from a diversity of taxa vary in their
130 average behavior across different environments (Bell, Hankison & Laskowski, 2009).
131 Yet, comparisons of among- and within-individual variation in average behavior (or other
132 forms of plasticity) among groups, populations, or treatments remain underrepresented
133 (e.g. Westneat et al., 2011; Dingemanse et al., 2012). For example, Westneat et al.,
134 (2011) found that female house sparrows vary less from one another in their average
135 provisioning behavior than male sparrows. In the model presented here, the random
136 intercept (b_{0ik}) for each individual (e.g. male and female nest provisioning rates are drawn
137 from Normal distributions with different variances) is drawn from a treatment-specific
138 Normal distribution.

139

140 *Within-Individual Variation in Intercept*

141 Within-individual variation in intercept ($\sigma^2_{\varepsilon k}$) is defined as the amount individuals
142 vary around their own average behavior. Within-individual variation is routinely used for
143 the calculation of repeatability in studies of animal personality (Bell, Hankison &
144 Laskowski, 2009; Dingemanse et al., 2010) or more often is simply regarded as noise,
145 despite the well established ecological and evolutionary implications of within-individual
146 variation (Stamps, Briffa & Biro, 2012; Biro & Adriasenssens, 2013; Westneat, Wright &
147 Dingemanse, 2014; Cleasby & Nakagawa, 2015). For example, a variable predator
148 environment may select for individual prey that vary greatly around their mean behavior
149 to remain unpredictable (Stamps, Briffa & Biro, 2012). LMMs can directly quantify
150 patterns of within-individual variation when repeated measures within multiple
151 individuals are available, facilitating comparisons of individual consistency between
152 groups of individuals (Dingemanse et al., 2013). Here we are interested in determining if
153 $\sigma^2_{\varepsilon k}$ differs by treatment. In other words, do individuals in one population or treatment
154 exhibit more intra-individual behavioral variation than individuals from a second
155 population or treatment?

156

157 *Among-Individual variation in slope*

158 Substantial empirical work has shown that individual animals in a variety of taxa
159 display variation in phenotypic plasticity (Martin & Réale, 2008; Mathot et al., 2011;
160 Dingemanse et al., 2012); using mixed models to quantify this variation has been the
161 primary focus of several recent papers (Martin et al., 2011; Van de Pol, 2012;

Dingemanse and Dochtermann, 2013). Among-individual variation in phenotypic plasticity has implications for the rate of evolutionary change, population stability and population persistence (Wolf & Weissing, 2012; Dingemanse & Wolf, 2013); thus defining those populations exhibiting greater individual variation in plasticity could help distinguish stable populations and populations with a high probability of micro-evolutionary change (Pigliucci, 2001; Ghalambor, Angeloni & Carroll, 2010). To quantify group differences in plasticity variation, multiple measurements within each individual across an environmental gradient are required. Here we are interested in determining if σ^2_{1k} differs by treatment.

Binomial GLMM

We assess power of a binomial GLMM for detecting differences in variation by treatment. This model can be written as:

[5] $y_{ijk} \mid b_{0ik}, b_{1ik,} v_{ijk} \sim \text{Binomial}(N_{ijk}, \pi_{ijk})$

[6] $\eta_{ijk} = \beta_0 + b_{0ik} + (\beta_1 + b_{1ik})X_{ij} + v_{ijk}$

[7] Inverse-logit: $\pi_{ijk} = 1/(1 + e^{-\eta_{ijk}})$

[8] $\begin{bmatrix} b_{0ik} \\ b_{1ik} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{0k} & \sigma_{01k} \\ \sigma_{01k} & \sigma^2_{1k} \end{bmatrix}\right)$

[9] $v_{ijk} \sim \text{Normal}(0, \sigma^2_{vk})$

Here, y_{ijk} is the number of "successes" in N_{ijk} observations of the *i*th individual in treatment *k* at the *j*th sampling occasion. When an environmental covariate (*X)* is present, one sampling occasion occurs at each level of the covariate *j*. In the absence of an environmental covariate, the linear predictor reduces to $\eta_{ijk} = \beta_0 + b_{0ik} + v_{ijk}$ and the *j*th occasion is simply a repeated sampling occasion in the same conditions. Note, when N_{ijk}

185 = 1 there is only 1 observation per sampling occasion *j*, making y_{ijk} a Bernoulli response
186 variable (see Supplement 1). When y_{ijk} is Bernoulli, overdispersion (v_{ijk}) and thus within-
187 individual variation is not identifiable.

188 In this model π_{ijk} describes the underlying probability of individual *i* in treatment
189 *k* at occasion *j* exhibiting a behavior. Variation in π is determined by the linear
190 combination of predictors on the logit (log-odds) scale: group intercept (β_0), group slope
191 (β_1), individual unique intercept (b_{0ik}), slope (b_{1ik}), and observation level overdispersion
192 that decreases predictive power at each observation (v_{ijk}). This linear predictor is
193 transformed with the inverse logit link to produce π_{ijk}, which follows a logit-Normal-
194 binomial mixed distribution.

195 We use an observation-level random effect to model additive overdispersion
196 (Browne et al., 2005), which models increased variance (following a Normal distribution
197 with variance σ^2_{vk}) in the linear predictor on the link scale (Nakagawa & Schielzeth,
198 2010). Overdispersion is used to quantify within-individual variation because it models
199 variation in π between each sampling occasion *j* for each individual. Here the magnitude
200 of overdispersion is allowed to vary by treatment (for an example of multiple data sets
201 where this occurs see Hinde & Demetrio, 2007), which is a focus of our power analysis.

202 The transformation through the inverse-logit function makes each of the three
203 target variance components difficult to visualize with a concise figure. However, because
204 the binomial GLMM model follows similar patterns as the LMM, we present power
205 analyses for the binomial GLMM using the visual aid presented for the LMM (Figure 1).
206 Finally, we simulate data for a fully balanced design without losing generality. See
207 Martin et al., 2011 and Van de Pol, 2012 for a discussion on experimental designs where

208 individuals are assayed in partially overlapping environments and when only single
209 measurements are obtained for some individuals.
210
211 ***Simulations***
212 All data were simulated in the R statistical programming environment using
213 newly developed simulation capabilities of the `lme4` package (Bates et al., 2014).
214 Guidelines for parameterizing the GLMMs and running data simulations and power
215 analyses are provided in Supplement 1. For a given total sample size, we present
216 simulations for determining the optimal ratio of total number of individuals versus the
217 number of repeated measures within individuals needed to provide power to detect a
218 difference among treatments 80% of the time. We conducted simulations for multiple
219 ratios of individuals to total observations within individuals, varying both sampling
220 occasions (j) and Bernoulli observations within sampling occasions (n). Next, we
221 describe simulations that evaluate how increasing "noise" (variation in non-target random
222 effects) affects power to detect differences in targeted variance comparisons.
223 For both scenarios we simulate data with biologically relevant parameter values
224 that illustrate common trends in power. At extreme parameter values the trends presented
225 here may not hold due to interactions between the variance components that arise at the
226 boundaries of binomial space. We do not dwell on these exceptions since they are
227 unrealistic for most empirical data sets, but suggest exploration of these exceptions with
228 code provided in Supplement 1.
229 We ran 2800 simulations for each combination of parameter values. The
230 significance of a given random effect was assessed using likelihood ratio tests (LRTs)

231 between models with and without the focal random effect. To correct for the known
232 conservatism of the LRT when testing for $\sigma^2 = 0$ (due to a null value on the boundary of
233 parameter space), we adopted the standard correction of dividing all p-values by 2
234 (Pinheiro & Bates, 2000; Verbeke & Molenberghs, 2000; Fitzmaurice, Laird & Ware,
235 2004; Zuur et al., 2009). This correction was appropriate for all p-values because each
236 LRT compared models that differed in only a single degree of freedom. Power is
237 estimated as the percentage of simulations that provide a corrected p-value smaller than
238 0.05. We insured the validity of a nominal p-value of 0.05 by confirming that 2800
239 simulations of a scenario where standard deviations did not differ at all did not result in
240 rejecting the null hypothesis more than 5% of the time. Under extremely low numbers of
241 individuals (~2-4) power to detect differences in the null case exceeded 5% (~10-15%),
242 possibly inflating power in these cases. Regardless, random effects cannot be reliably
243 estimated with such low sample sizes and therefore in most cases such experimental
244 designs should be avoided.
245
246 ***Scenario 1: Determining the optimal sampling scheme***
247 Most researchers face limitations imposed by time, money and access to samples,
248 and are therefore confronted with the question of how resources should be divided
249 between individuals and measures within individuals. To investigate the optimal
250 allocation of sampling effort between the number of individuals and number of
251 observations per individual, we simulated two data sets for each variance comparison
252 (See Table 1 for a summary of all simulations).

First, using three hypothetical total numbers of Bernoulli observations *per treatment* (total sample size per treatment, TSS_T), we manipulated either the ratio of individuals to sampling occasions (σ^2_{0k} and σ^2_{1k}), or the ratio of individuals to Bernoulli observations within sampling occasions ($\sigma^2_{\nu k}$). For comparisons of σ^2_{0k} and σ^2_{1k} we manipulated the ratio of individuals to sampling occasions, holding the number of Bernoulli observations constant at 5, because power follows a non-monotonic pattern across these ratios for σ^2_{0k} and σ^2_{1k} (Figures 2, 3). Conversely, for comparisons of $\sigma^2_{\nu k}$ we manipulated the ratio of individuals to Bernoulli observations and held the number of sampling occasions constant at 5 because power follows a non-monotonic pattern across ratios of individuals to Bernoulli observations for $\sigma^2_{\nu k}$ (Figure 4). For comparisons of σ^2_{0k}, and $\sigma^2_{\nu k}$ we simulated TSS_T of 600, 1200 and 2400, and for comparisons of σ^2_{1k} TSS_T were 300, 600, and 1200. For example, for b_{1ik} with a TSS_T of 300, the most extreme ratios were 30 individuals with 2 sampling occasions and 2 individuals with 30 sampling occasions. While using only 2 samples for a grouping variable (individuals) is never suggested for a random effect, we include this combination as an illustration of the low power that results from an ill-conceived sampling scheme. For each variance comparison we simulated three different effect sizes (2, 2.5, and 3 fold difference in standard deviation by treatment).

Next, we simulated data sets with increasing numbers of Bernoulli observations for comparisons of σ^2_{0k} and σ^2_{1k} (Figure 5A, B) and with increasing numbers of sampling occasions for comparisons of $\sigma^2_{\nu k}$ (Figure 5C). For these simulations we used 1, 3, 5, 10 and 15 Bernoulli observations or sampling occasions. Ratios of individuals to sampling occasions (σ^2_{0k} and σ^2_{1k}) or individuals to Bernoulli observations ($\sigma^2_{\nu k}$) followed the

intermediate TSS_T from the simulations described above. For example, for comparisons of σ^2_{0k} we simulated 1, 3, 5, 10 and 15 Bernoulli observations for ratios of individuals to sampling occasions ranging from 120:2 to 2:120. For all comparisons we simulated data using an effect size of a 2.5 fold difference in standard deviation by treatment.

In all Scenario 1 simulations, both β_0 and β_1 were constrained to a single value for all treatments. For comparisons of among-individual variation in intercept no environmental covariate was used, and $\sigma^2_{\nu k}$ was held constant among treatments. For comparisons of among-individual variation in slope we held $\sigma^2_{\nu k}$ constant. Finally, for comparisons of within-individual variation in intercept, no environmental covariate was included and σ^2_{0k} was held constant among treatments. All parameter values used in simulations for both Scenarios can be found in Table S1.

Our goal in Scenario 1 was to isolate changes in a single variance parameter, but exploration of the dependence among multiple variance components and the mean may be warranted if it is relevant for a specific problem. Incorporating concurrent changes in intercept, slope and overdispersion parameters can be easily implemented with slight modifications to the code presented in the online supplement. We show initial results of relaxing some of these assumptions in Scenario 2, but full exploration of these possibilities are beyond the scope of this paper.

Scenario 2: Measuring the ratio of overdispersion to effect size

Decreasing the ratio of the variance in the target random effect to total variance influences power to detect differences in the target variance among treatments. Therefore, we simulated four levels of "noise" (magnitude of non-target random effect variance)

299 assuming a Normal distribution with increasing standard deviations (0.1, 0.5, 1.0, 2.0)
300 (Figure 6). These correspond to ratios of target variance parameter effect size to non-
301 target variance of 25:1, 5:1, 5:2, and 5:4. For comparisons of σ^2_{0k} and $\sigma^2_{1k,}$"noise" was
302 simulated with increasing variation in within-individual variation (σ^2_{vk}), while for
303 comparisons of σ^2_{vk} noise was simulated with among-individual variation in intercept
304 (σ^2_{0k}). For each variance parameter ratios of individuals to repeated measures followed
305 the largest TSS_T sampling scheme used in Scenario 1 and an ES of a 2.5x difference in
306 standard deviation by treatment.
307
308 **Results**
309 ***Scenario 1: Determining the optimal sampling scheme***
310 Power to detect differences between treatments for each variance component
311 increases with total sample size (TSS_T) and effect size (ES) (Figures 2-5). For a given
312 TSS_T power depends on the ratio of the number of individuals to the number of repeated
313 measures per individual. However, the optimal ratio of individuals to repeated measures
314 varies depending on TSS_T and target variance parameter. For example, power to detect
315 both σ^2_{0k} and σ^2_{1k} is non-monotonic across ratios of individuals to sampling occasions
316 (Figures 2, 3), but is an increasing function of the number of Bernoulli observations
317 within sampling occasions (Figure 5A, B). Power to detect σ^2_{0k} is maximized at a ratio of
318 individuals to repeated measures of approximately 6:5 under low sample sizes (TSS_T =
319 600) (Figure 2A) but a ratio of approximately 2:1 is optimal under larger sample sizes
320 (TSS_T = 2400) (Figure 2C).

321 At low sample sizes ($TSS_T = 300$), power to detect σ^2_{1k} is maximized at a ratio of
322 approximately 12:5 (Figure 3A), while larger sample sizes ($TSS_T = 600, 1200$) favor a
323 ratio heavily weighted towards having more individuals (approximately 5:1) versus more
324 repeated measures (Figure 3B, C). Power to detect σ^2_{1k} is higher overall and less sensitive
325 to deviations from the optimum ratio than power to detect σ^2_{0k} (Figure 3).
326 Power to detect σ^2_{vk} follows a strikingly different pattern than σ^2_{0k} and σ^2_{1k}. Power
327 to detect σ^2_{vk} is non-monotonic across ratios of individuals to the number of Bernoulli
328 observations within sampling occasions (Figure 4), and is an increasing function of the
329 number of sampling occasions (Figure 5C). At low sample sizes (e.g. $TSS_T = 600$) power
330 to detect σ^2_{vk} is maximized by devoting nearly all of the available resources to repeated
331 measures within individuals (Figure 4A); however, at larger sample sizes (e.g. $TSS_T =$
332 2,400) power is maximized at a ratio of individuals to Bernoulli observations of
333 approximately 1:2 (Figure 4).
334
335 ***Scenario 2: Power under increasing non-target random effect variance***
336 Power to detect differences in variance components is strongly affected by the
337 proportion of total variance that can be attributed to the target variance component
338 (Figure 6). Increasing variance in non-target random effects decreases power to detect
339 differences in the target variance parameter by treatment. However, the ratio of target to
340 non-target variance does not alter the optimal ratio of individuals to repeated measures
341 for the target variance comparison (Figure 6). Panel A demonstrates that power to detect
342 σ^2_{0k} decreases substantially as the magnitude of within-individual variation increases.
343 Detecting differences in σ^2_{1k} depends only on total random effect variation at extreme

ratios of individuals to sampling occasions (e.g. 80:3) (Figure 6B). Finally, detection of σ^2_{vk} is largely independent of the magnitude of among-individual variation at large ratios of ES to non-target variance, as indicated by overlapping curves in Figure 6C. However, when among-individual variation in intercept is very large (Figure 6C: Red curve), power to detect σ^2_{vk} decreases because individual mean responses approach 0 or 1, reducing the amount of detectable within-individual variation.

Discussion

The power analyses presented here establish a framework for designing experiments focused on detecting differences in variance components by treatment using GLMMs. These results should serve as a baseline upon which researchers can expand to address their own specific problems. Nevertheless, our findings reveal some important general trends that should be considered when designing experiments. Our results demonstrate heterogeneity in power across sampling schemes (ratio of individuals to repeated measures and partitioning of repeated measures into sampling occasions and Bernoulli observations), and differences in which sampling scheme maximizes power for different components of variance (Figures 2-5). As expected, power declines rapidly for low sample sizes and small effect sizes (Figures 2-4). However, for large TSS_T and relatively large effect sizes (3 SD difference between treatments), > 80% power is retained across many different combinations of individuals to repeated measures for each component of variance (Figures 2-5). Not surprisingly, power to detect differences in the target random effect declines with increasing variance in the non-target random effects (Figure 6).

Power to detect σ^2_{0k} is non-monotonic across ratios of individuals to sampling occasions, and is an increasing function of the number of Bernoulli observations per sampling occasion. Power is maximized with ratios weighted towards having more individuals (Figure 2), and quickly declines with alternative sampling ratios when total sample sizes and effect sizes are small. The analyses are however more robust to deviations from this ratio when TSS_T and ES are large (Figure 2C). Finally, of all the random effect parameters we analyzed, power to detect σ^2_{0k} is the most sensitive to the amount of "noise" present in the model, decreasing rapidly with increasing within-individual variation (Figure 6).

Power to detect σ^2_{1k} is also non-monotonic across ratios of individuals to sampling occasions, and is maximized with a ratio of individuals to sampling occasions ranging from 2:1 to 5:1 as TSS_T increases (Figure 3). On average, testing for differences in σ^2_{1k} are more powerful than for σ^2_{0k} across all sampling schemes and ES (Figures 2, 3), and requires fewer samples to obtain 80% power.

Finally, power to detect $\sigma^2_{\nu k}$ is non-monotonic across ratios of individuals to Bernoulli observations and is an increasing function of the number of sampling occasions. Depending on sample size, sampling schemes ranging from maximizing Bernoulli observations to ratios of individuals to Bernoulli observations of 1:2 maximizes power (Figure 4). Unlike σ^2_{0k}, power to detect $\sigma^2_{\nu k}$ is largely independent of additional variance in the model (Figure 6C), such that power to detect $\sigma^2_{\nu k}$ is nearly equivalent at all levels of σ^2_{0k} except under the case of extreme values of $\sigma^2_{0k.}$

Collectively these results indicate the importance of clearly defining a biological question, designating the focal random effect, and knowing the expected magnitude of

total variation when determining the appropriate experimental sampling design and TSS_T. Even at larger effect sizes, failure to account for system noise can lead to insufficient power and a failed experiment. Our findings should serve as a strong warning to empiricists interested in variance components that power analyses should be performed when designing experiments in order to overcome the problems of overall low power, large heterogeneity in power to detect different variance components, and heterogeneity in sampling scheme required to optimize power.

By introducing new strategies for analyzing variance among treatments we hope to inspire novel experimental designs in ecology and evolution. For example, the power analyses presented here can inform the design of experiments aimed at quantifying heterogeneous within-individual variation by environment, which may lead to novel insights on the adaptive significance of within-individual variation (Westneat, Wright & Dingemanse, 2014).

In addition, these analyses answer the calls of researchers over the last decade for methods to investigate effects of treatment level variance on the variance of dependent variables (Benedetti-Cecchi, 2003). Transitions from one discrete environment to another (e.g. presence or absence of predators) are often classified as a form environmental variation, but switching between two distinct but relatively constant environments does not reflect environmental variation *per se*, such as temporal changes in the magnitude, pattern, and/or frequency of the environmental over time (Benedetti-Cecchi, 2003; Benedetti-Cecchi et al., 2006; Miner & Vonesh, 2004; Lawson et al., 2015). When this form of environmental variation is manipulated or natural variation exploited in an experimental context, within-individual variation can be described as the variable

response of individuals to this variation in the environment. In this context, within-individual variation may itself be a form of phenotypic plasticity, and may have profound implications for understanding the evolution of environmentally induced plasticity, and the evolution of labile traits generally (Stamps, Briffa & Biro, 2012; Biro & Adriasenssens, 2013; Westneat, Wright & Dingemanse, 2014).

Further Considerations

Heterogeneous within-individual variation

In our power analyses we have made a few important simplifying assumptions. First, we assume that within-individual variation in both intercept and slope is homogenous among individuals within the same treatment. Additionally, we assume homogeneity of within-individual variance across an environmental gradient. However, these assumptions may not be true for some natural or experimental populations. In fact, it has recently been proposed that assessing the magnitude of variation in within-individual error variance within a single individual across an environmental gradient or among individuals exposed to the same environment/treatment is an important metric that may help to explain the evolution of plasticity (Cleasby & Nakagawa, 2015; Westneat, Wright & Dingemanse, 2014). Power to detect differences in the magnitude of among-individual variation in within-individual variation by treatment (Cleasby & Nakagawa, 2015) and heterogeneity of variance across an environmental gradient are interesting research questions that deserve attention, but are beyond the scope of this article. We also note that practicality limits exploration of increasingly complicated scenarios, despite their conceivable statistical feasibility and intrinsic charm due to complex novelty.

436

437 *Covariance among intercept, slope, and variance components*

438 All of our simulations assessed power to detect differences in a single target
439 variance comparison between treatments, holding all other variance parameters constant
440 (Table S1). However, manipulating non-target variation generates additional variation
441 that is expected to decrease power to detect differences in the target variance parameter.
442 Because we assumed no slope variation in models where intercepts were allowed to vary
443 and no intercept variation in the models focused on variation in slopes, we did not discuss
444 power to detect covariance terms. However, these parameters can co-vary and the
445 covariation among these parameters may contain a wealth of biologically relevant
446 information. For example, covariation between phenotypic plasticity and within-
447 individual variation may be tightly linked via developmental tradeoffs, which can lead to
448 greater developmental instability in highly plastic individuals (Tonsor, Elnaccash &
449 Scheiner, 2013). Indeed, it is not known whether an individual's reaction norm slope and
450 within-individual variation around that reaction norm are always linked or if these
451 relationships can be context-dependent. Similarly, we do not know if stronger behavioral
452 responses lead to greater canalization of behavior. Understanding how to parameterize
453 GLMM and how to optimize experiments to detect these covariances will be a useful step
454 toward advancing evolutionary theory on adaptive, maladaptive and random patterns of
455 variation.

456 Covariance between intercept and slope has been described extensively in
457 theoretical papers and has been explored in earlier power analyses for LMM
458 (Dingemanse & Dochtermann, 2013); however, empirical studies documenting

459 significant covariance between these parameters remain rare (Mathot et al., 2011;
460 Dingemanse et al., 2012). While covariance among these parameters may be uncommon,
461 it is also likely that most experiments have insufficient power to detect such covariance.
462 Additional analyses that determine power to detect significant differences in intercept and
463 slope covariation for GLMMs is another important step considering the lack of current
464 evidence for covariation reported in the literature.
465
466 *Within-individual variation in slope*
467 Research, including ours, on among-individual variation in plasticity assumes
468 fully repeatable plasticity within each individual, causing among-individual differences in
469 phenotypic plasticity to be calculated using a single reaction norm for each individual
470 (Dingemanse & Wolf, 2013). However, quantifying only a single reaction norm for each
471 individual fails to capture any potential variation in plastic responses within an individual
472 around its mean reaction norm, which may inflate estimates of among-individual
473 variation and mask important variation that is subject to selection (Dingemanse & Wolf,
474 2013). Despite the reasonable assumption that each experimental individual would
475 exhibit variation in their reaction norm if it were repeatedly measured, we are aware of no
476 studies that demonstrate repeatable behavioral plasticity for a single individual when
477 assessed multiple times.
478
479 *Heterogeneity in sampling scheme and environment*
480 In our simulations all individuals were measured an equal number of times and all
481 treatments contained the same number of individuals, a luxury often not available to

empiricists that often deal with missing data and unbalanced designs. Intuitively, unbalanced sampling schemes will lower the power to detect among-individual variation (Van de Pol, 2012); however we do not know the rate at which statistical power is lost with the magnitude of imbalance for a particular sampling design. Future research should follow the lead of Van de Pol, 2012 to determine how power to assess differences in variance for GLMM is affected by incomplete sampling, specifically when only a single measure is available for some individuals.

Experiments with more than two treatments

Finally, these power analyses were created for a two-treatment scenario--"homogenous" environmental variation treatment and a "variable" environmental variation treatment. However, it is commonplace to have more than two treatments. Fortunately, our framework for conducting power analyses can be easily generalized for exploring power for experiments with more than two treatments (see supplemental material). In addition, syntax for the `lme4` package in R for specifying GLMM is highly flexible and can be written to restrict variance components to be the same in any number of treatments, while unique variance estimates can be obtained for any other given treatment. For example in a four treatment experiment composed of four levels of predator cue, two variance estimates could be obtained for among-individual variation (e.g. a single estimate for the three treatments with the lowest levels of predator cue and one estimate for the highest level of predator cue).

Conclusions

505 Random intercepts and slopes GLMMs are well established both ecology and
506 evolution and behavioral ecology. Despite their ubiquity, the use of GLMMs to compare
507 variance components among populations or among experimental treatments is rare. We
508 call for future work analyzing the accuracy and precision of estimates comparing random
509 effects by treatment for GLMMs (which our code facilitates) similar to the work of
510 Moineddin, Matheson & Glazier, 2007 and Van de Pol, 2012 on the accuracy and
511 precision of random effects estimates. As Van de Pol points out, just because power is
512 high does not ensure the accuracy and precision of estimates. Finally, with expanding
513 interest in a variety of variance parameters (e.g. heterogeneity in within-individual
514 variation), we hope the power analyses presented here will spur novel empirical research
515 and assist readers in constructing appropriate experimental designs and statistical models
516 to test how variance components are shaped by ecological and evolutionary processes.

517 **Acknowledgements**

518 This research was supported through a new faculty start up grant from East Carolina
519 University. We thank the lab members of Dr. Michael McCoy's lab and Dr. Krista
520 McCoy's lab for helpful comments during the preparation of this manuscript.

521 **Data Accessibility**

522 R scripts are available in Supplement 1

References

Bates, D., Maechler, M., Bolker, B., & Walker S. 2014. Fitting linear mixed-effects models using lme4. arXiv reprint. arXiv: 1406.5823.

Bell, A.M., Hankison, S.J. & Laskowski, K.L. 2009. The repeatability of behaviour: a meta-analysis. *Animal Behaviour*, *77*, 771–783.

Benedetti-Cecchi, L. 2003. The importance of the variance around the mean effect size of ecological processes. *Ecology*, 84, 2335–2346.

Benedetti-Cecchi, L., Bertocci, I., Vaselli, S. & Maggi, E. 2006. Temporal variance reverses the impact of high mean intensity of stress in climate change experiments. *Ecology*, 87, 2489-2499.

Biro, P.A. & Adriaenssens, B. 2013. Predictability as a personality trait: consistent differences in intra-individual behavioral variation. *The American Naturalist,* 182, 621–629.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution,* 24, 127–135.

Bolker, B.M. 2015. Linear and generalized linear mixed models. *Ecological Statistics: Contemporary theory and application* (eds G.A. Fox, S. Negrete-Yankelevich & V.J. Sosa), pp. 309–334. Oxford University Press, USA.

Browne, W.J., Subramanian, S.V., Jones, K. & Goldstein, H. 2005. Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society Series A-Statistics in Society*, 168, 599–613.

Cernicchiaro, N., Renter, D.G., Xiang, S., White, B.J. & Bello, N.M. 2013. Hierarchical Bayesian modeling of heterogeneous variances in average daily weight gain of commercial feedlot cattle. *Journal of animal science*, 91, 2910–2919.

Cleasby, I.R., Nakagawa, S. & Schielzeth, H. 2015. Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-individual variance. *Methods in Ecology and Evolution*, 6, 27–37.

Dingemanse, N.J., Kazem, A.J.N., Réale, D. & Wright, J. 2010. Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology and Evolution*, 25, 81–89.

Dingemanse, N.J., Bouwman, K.M., van de Pol, M., van Overveld, T., Patrick, S.C., Matthysen, E. & Quinn, J.L. 2012. Variation in personality and behavioural plasticity across four populations of the great tit Parus major. *Journal of Animal Ecology*, 81, 116–126.

566 Dingemanse N.J. & Dochtermann N.A. 2013. Quantifying variation in behaviour:
567 mixed-effect modeling approaches. *Journal of Animal Ecology*, 82, 39–54.

568 Dingemanse, N.J. & Wolf, M. 2013. Between-individual differences in behavioural
569 plasticity within populations: causes and consequences. *Animal Behaviour*, 85,
570 1031–1039.

571 Fitzmaurice, G.M., Laird, N.M. & Ware, J.H. 2004. *Applied Longitudinal Analysis*, 2nd
572 edn. Wiley, New Jersey, USA.

573 Gelman, A. & Hill, J. 2006. *Data Analysis Using Regression and Multi-*
574 *level/Hierarchical Models*. Cambridge University Press, England.

575 Ghalambor, C.K., Angeloni, L.M. & Carroll, S.P. 2010. Behavior as phenotypic
576 plasticity. *Evolutionary behavioral ecology* (eds D.F. Westneat & C.W. Fox), pp.
577 90–197. Oxford University Press, USA.
578
579 Hill, W.G. & Zhang, X.S. 2004. Effects on phenotypic variability of directional selection
580 arising through genetic differences in residual variability. *Genetical Research*, 83,
581 121–132.
582
583 Hinde, J. 1982. Compound Poisson regression models. *GLIM 82: Proceedings of the*
584 *International Conference on Generalised Linear Models* (ed R. Gilchrist), pp. 109–
585 121. Springer, New York.
586
587 Jennions, M. & Møller, A. 2003. A survey of the statistical power of research in
588 behavioral ecology and animal behavior. *Behavioral Ecology*, 14, 438–445.
589
590 Johnson, P.C., Barry, S.J., Ferguson, H.M. & Müller, P. 2014. Power analysis for
591 generalized linear mixed models in ecology and evolution. *Methods in Ecology*
592 *and Evolution*, 6, 133–142.

593 Kizilkaya, K. & Tempelman, R.J. 2005. A general approach to mixed effects modeling of
594 residual variances in generalized linear mixed models. *Genetics Selection*
595 *Evolution*, 37, 31–56.

596 Lawson, C.R., Yngvild V., Bailey, L. & van de Pol, M. 2015. Environmental variation
597 and population responses to global change. *Ecology Letters*,
598 doi: 10.1111/ele.12437

599 Lynch, M. & Walsh, B. 1998. *Genetics and Analysis of Quantitative Traits*, 1st edn.
600 Sinauer Press, Sunderland.

601 Martin, J.G.A. & Réale, D. 2008. Temperament, risk assessment and habituation to
602 novelty in eastern chipmunks, *Tamias striatus*. *Animal Behaviour*, 75, 309–318.

603 Martin, J.G.A., Nussey, D.H., Wilson, A.J. & Réale, D. 2011. Measuring individual
604 differences in reaction norms in field and experimental studies: a power analysis
605 of random regression models. *Methods in Ecology and Evolution*, 2, 362–374.
606
607 Mathot, K.J., van den Hout, P.J., Piersma, T., Kempenaers, B., Réale, D. & Dingemanse,
608 N.J. 2011. Disentangling the roles of frequency-vs. state-dependence in
609 generating individual differences in behavioural plasticity. *Ecology letters*, 14,
610 1254–1262.

611 Miner, B.G. & Vonesh, J.R. 2004. Effects of fine grain environmental variability on
612 morphological plasticity. *Ecology Letters*, 7, 794–801.

613 Moineddin, R., Matheson, F.I. & Glazier, R.H. 2007. A simulation study of sample size
614 for multilevel logistic regression models. *BMC Medical Research Methodology*,
615 7, 1–34.
616
617 Moore, A.J., Brodie, E.D. III & Wolf, J.B. 1997. Interacting phenotypes and the
618 evolutionary process: I. Direct and indirect genetic effects of social interactions.
619 *Evolution*, 51, 1352–1362.

620 Nakagawa, S. & Schielzeth, H. 2010. Repeatability for Gaussian and non-Gaussian data:
621 a practical guide for biologists. *Biological Reviews*, 85, 935–956.

622 Nussey, D.H., Wilson, A.J. & Brommer, J.E. 2007. The evolutionary ecology of
623 individual phenotypic plasticity in wild populations. *Journal of Evolutionary*
624 *Biology*, 20, 831–844.
625
626 Pigliucci M. 2001. *Phenotypic Plasticity: Beyond Nature and Nurture*. John Hopkins
627 University Press, Baltimore, MD.
628
629 Pinheiro, J. & Bates, D. 2001. *Mixed-Effects Models in S and S-PLUS*. Springer, New
630 York.
631
632 R Development Core Team. 2014. *R: A Language and Environment for Statistical*
633 *Computing*. R Foundation for Statistical Computing, Vienna. URL http://www.R-
634 project.org
635
636 Scherbaum, C.A. & Ferreter, J.M. 2009. Estimating statistical power and required sample
637 sizes for organizational research using multilevel modeling. *Organizational*
638 *Research Methods*, 12, 347–367.
639
640 Schielzeth, H. & Nakagawa, S. 2013. Nested by design: model fitting and interpretation
641 in a mixed model era. *Methods in Ecology and Evolution*, 4, 14–24.
642
643 Stamps, J.A., Briffa, M. & Biro, P.A. 2012. Unpredictable animals: individual differences
644 in intraindividual variability (IIV). *Animal Behaviour*, 83, 1325– 1334.
645

646 Stroup, W.W. 2012. *Generalized linear mixed models: modern concepts, methods and*
647 *applications*. CRC Press, USA.

648 Tonsor, S.J., Elnaccash, T.W. & Scheiner, S.M. 2013. Developmental instability is
649 genetically correlated with phenotypic plasticity, constraining heritability, and
650 fitness. *Evolution*, 67, 2923–2935.
651
652 van de Pol, M. & Wright, J. 2009. A simple method for distinguishing within-versus
653 between-subject effects using mixed models. *Animal Behaviour*, 77, 753–758.
654
655 Verbeke, G. & Molenberghs, G. 2000. *Linear Mixed Models for Longitudinal Data*.
656 Springer, New York.
657
658 Westneat, D.F., Hatch, M.I., Wetzel, D.P. & Ensminger, A.L. 2011. Individual variation
659 in parental care reaction norms: integration of personality and plasticity. *The*
660 *American Naturalist*, 178, 652–667.
661
662 Westneat, D.F., Wright, J. & Dingemanse, N.J. 2014. The biology hidden inside residual
663 within-individual phenotypic variation. *Biological Reviews*. doi:10.
664 1111/brv.12131
665
666 Wolf, M. & Weissing, F.J. 2012. Animal personalities: consequences for ecology and
667 evolution. *Trends in Ecology & Evolution*, 27, 452-461.
668
669 Zuur, A.F., Leno, E.N., Walker, N., Saveliev, A.A. & Smith, G.M. 2009. *Mixed effects*
670 *models and extensions in ecology with R*. Springer, New York.

671 Zuur, A.F., Hilbe, J.M. & Leno, E.N. 2013. *A Beginner's Guide to GLM and GLMM with*
672 *R: A Frequentist and Bayesian Perspective for Ecologists*. Highland Statistics
673 Ltd, Newburgh, UK.

674

675

Figure 1. Reaction norm plots for a two treatment LMM. In all graphs bolded black lines depict treatment mean reaction norms and thin lines depict reaction norms of individuals. Grey envelopes in (C) illustrate the magnitude of within-individual intercept variation. Here among-individual variation in intercept (A), slope (B), and within-individual variation in intercept (C) is larger in treatment 2.

Figure 2. Power to detect differences by treatment in σ_0 for three effect sizes (ratio of σ_0 between treatments) and three TSS_T (total sample size per treatment). Each scenario was simulated with 5 Bernoulli observations per sampling occasion.

Figure 3. Power to detect differences by treatment in σ_1 for three effect sizes and three TSS_T. Each scenario was simulated with 5 Bernoulli observations per sampling occasion.

Figure 4. Power to detect differences by treatment in σ_v for three effect sizes and three TSS_T. Each scenario was simulated with 5 sampling occasions.

Figure 5. Power to detect differences by treatment in σ_0 (A) and σ_1 (B) under increasing Bernoulli observations per sampling occasion; σ_v (C) under increasing sampling occasions. In (A) and (B) ratios of individuals to sampling occasions follow figures 2B and 3B respectively. In (C) ratios of individuals to Bernoulli observations follows figure 4B.

Figure 6. Power to detect differences by treatment in σ_0 (A) and σ_1 (B) under increasing variation in σ_v; σ_v (C) under increasing variation in σ_0. Noise is given as the ratio of effect size to variation in the non-target variance parameter. In (A) and (B) ratios of individuals to sampling occasions follow figures 2C and 3C respectively. In (C) ratios of individuals to Bernoulli observations follows figure 4C.

Table 1(on next page)

Parameter values for all simulations

Table 1: For example, Scenario 1: Figure 2C illustrates power to detect differences in σ^2_{0k} across ratios of individuals to sampling occasions with a TSS_T of 2,400 at effect sizes of 2x, 2.5x, and 3x difference in standard deviation by treatment.

Table 1. Parameter values for all simulations. For example, Scenario 1: Figure 2C illustrates power to detect differences in σ^2_{0k} across ratios of individuals to sampling occasions with a TSS_T of 2,400 at effect sizes of 2x, 2.5x, and 3x difference in standard deviation by treatment.

Target Variance	**σ^2_{0k}**					**σ^2_{1k}**					**$\sigma^2_{\nu k}$**				
Scenario	**1**			**2**		**1**			**2**		**1**				**2**
Figure	**2A**	**2B**	**2C**	**5A**	**6A**	**3A**	**3B**	**3C**	**5B**	**6B**	**4A**	**4B**	**4C**	**5C**	**6C**
Parameter	Sampling Occasions			Bernoulli Obs	$\sigma^2_{\nu k}$	Sampling Occasions			Bernoulli Obs	$\sigma^2_{\nu k}$	Bernoulli Observations			Sampling Occasions	σ^2_{0k}
TSS_T	600	1,200	2,400	240 - 3,600	2,400	300	600	1,200	120- 1,800	1,200	600	1,200	2,400	240 - 3,600	2,400
# Individuals	2-60	2-120	2-240	120-2	2-240	2-30	2-60	2-120	60-2	2-120	2-60	2-120	2-240	2-120	2-240
# Sampling Occasions	60-2	120-2	240-2	2-120	240-2	30-2	60-2	120-2	2-60	120-2	5	5	5	1-15	5
# Bernoulli Observations	5	5	5	1- 15	5	5	5	5	1-15	5	60-2	120-2	240-2	120-2	240-2
Effect Sizes	2; 2.5; 3	2; 2.5; 3	2; 2.5; 3	2.5	2.5	2; 2.5; 3	2; 2.5; 3	2; 2.5; 3	2.5	2.5	2; 2.5; 3	2; 2.5; 3	2; 2.5; 3	2.5	2.5

Figure 1(on next page)

Reaction norm plots for a two treatment LMM

Figure 1: In all graphs bolded black lines depict treatment mean reaction norms and thin lines depict reaction norms of individuals. Grey envelopes in (C) illustrate the magnitude of within-individual intercept variation. Here among-individual variation in intercept (A), slope (B), and within-individual variation in intercept (C) is larger in treatment 2.

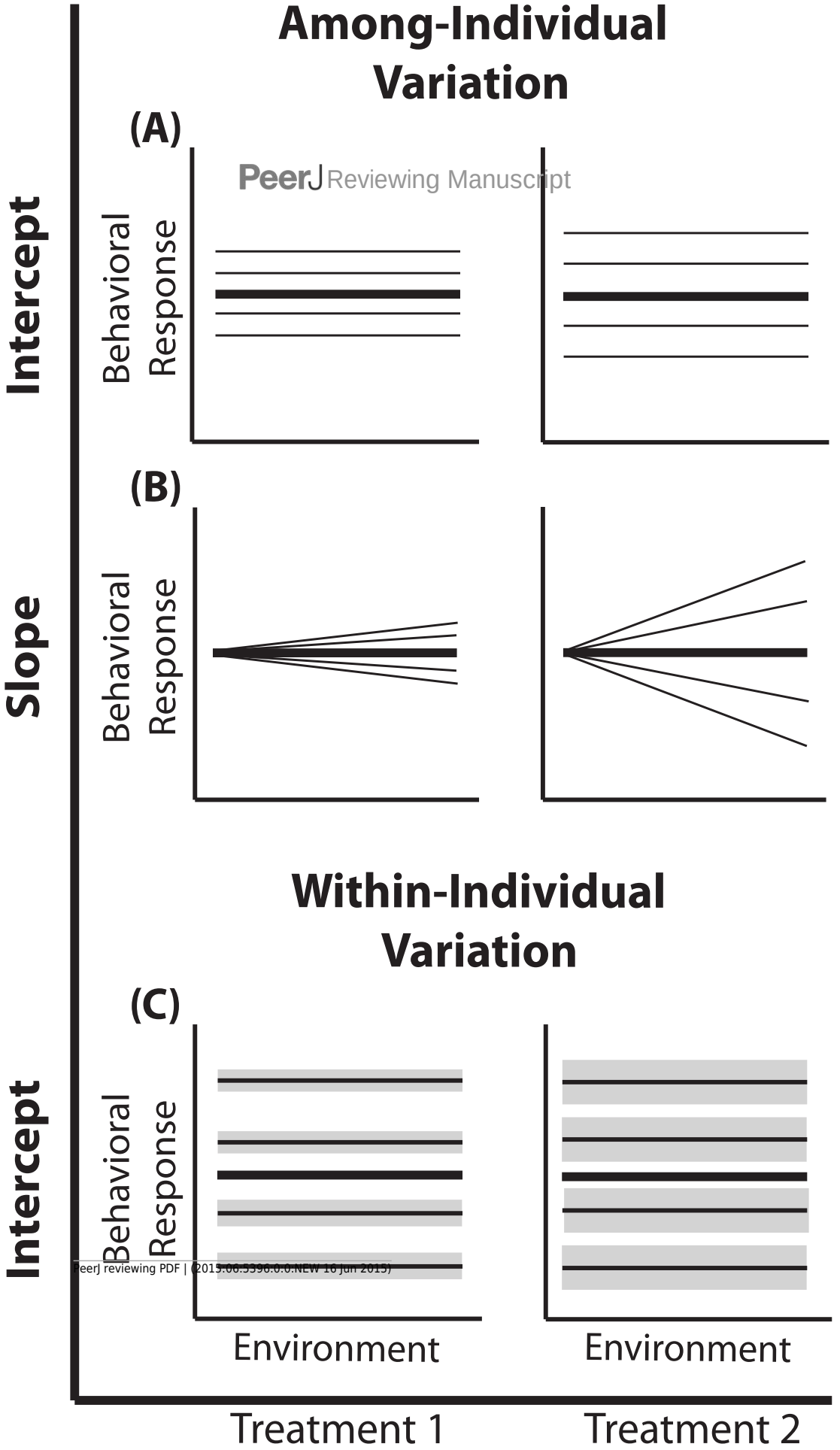


Figure 2(on next page)

Power to detect differences by treatment in among-indiviudal variation in intercept

Figure 2: Power to detect differences by treatment in σ_0 for three effect sizes (ratio of σ_0 between treatments) and three TSS_T (total sample size per treatment). Each scenario was simulated with 5 Bernoulli observations per sampling occasion.

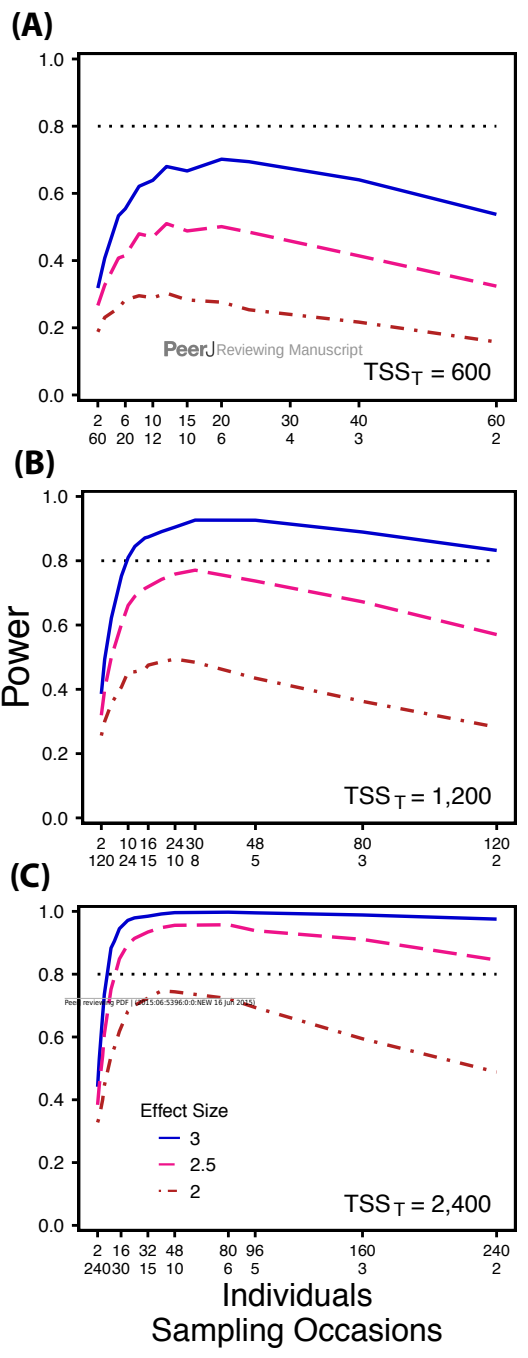

(A)
1.0
0.8
0.6
0.4
0.2
0.0
PeerJ Reviewing Manuscript
TSS$_T$ = 600
2 6 10 15 20 30 40 60
60 20 12 10 6 4 3 2
(B)
Power
1.0
0.8
0.6
0.4
0.2
0.0
TSS$_T$ = 1,200
2 10 16 24 30 48 80 120
120 24 15 10 8 5 3 2
(C)
1.0
0.8
0.6
0.4
0.2
0.0
Effect Size
3
2.5
2
TSS$_T$ = 2,400
2 16 32 48 80 96 160 240
240 30 15 10 6 5 3 2
Individuals
Sampling Occasions

Figure 3(on next page)

Power to detect differences by treatment in among-indiviudal variation in slope

Figure 3: Power to detect differences by treatment in σ_1 for three effect sizes and three TSS_T. Each scenario was simulated with 5 Bernoulli observations per sampling occasion.

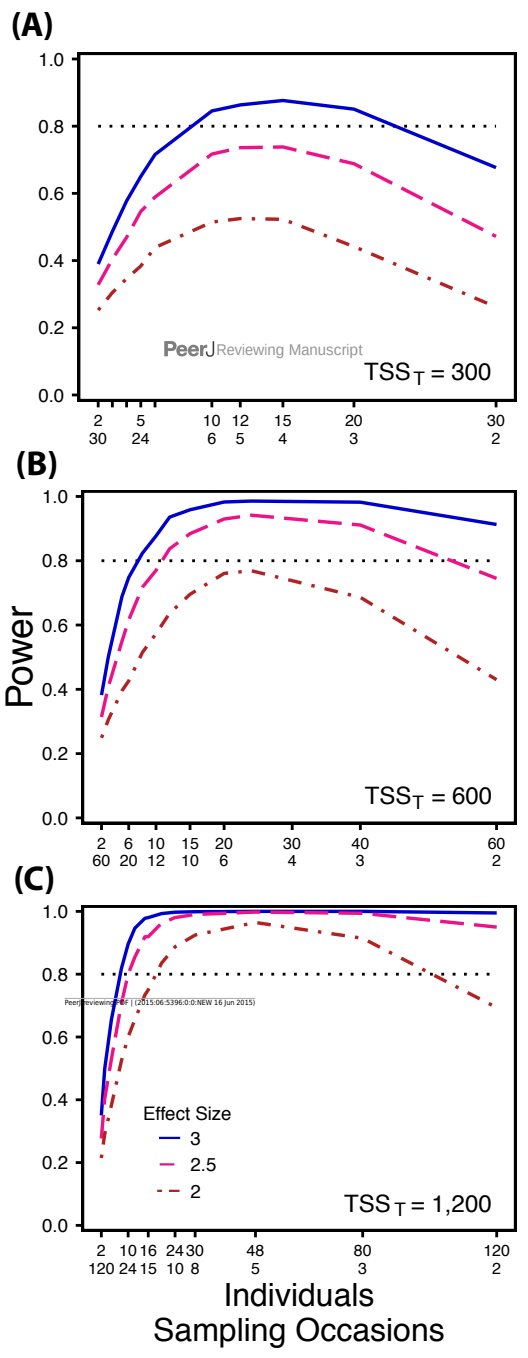

(A)
1.0
0.8
0.6
0.4
0.2
0.0
PeerJ Reviewing Manuscript
TSS$_T$ = 300
2 5 10 12 15 20 30
30 24 6 5 4 3 2
(B)
Power
1.0
0.8
0.6
0.4
0.2
0.0
TSS$_T$ = 600
2 6 10 15 20 30 40 60
60 20 12 10 6 4 3 2
(C)
1.0
0.8
0.6
0.4
0.2
0.0
Effect Size
3
2.5
2
TSS$_T$ = 1,200
2 10 16 24 30 48 80 120
120 24 15 10 8 5 3 2
Individuals
Sampling Occasions

Figure 4(on next page)

Power to detect differences by treatment in within-indiviudal variation in intercept

Figure 4: Power to detect differences by treatment in σ_v for three effect sizes and three TSS_T. Each scenario was simulated with 5 sampling occasions.

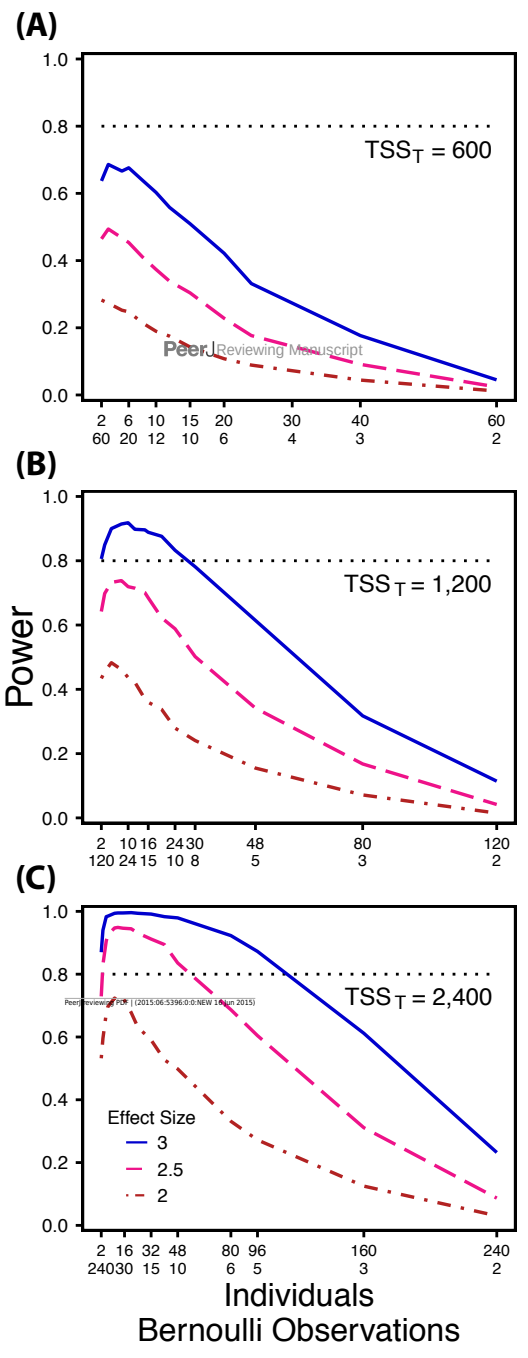


Figure 5(on next page)

Power under increasing Bernoulli observations or sampling occasions

Figure 5: Power to detect differences by treatment in σ_0 (A) and σ_1 (B) under increasing Bernoulli observations per sampling occasion; σ_v (C) under increasing sampling occasions. In (A) and (B) ratios of individuals to sampling occasions follow figures 2B and 3B respectively. In (C) ratios of individuals to Bernoulli observations follows figure 4B.

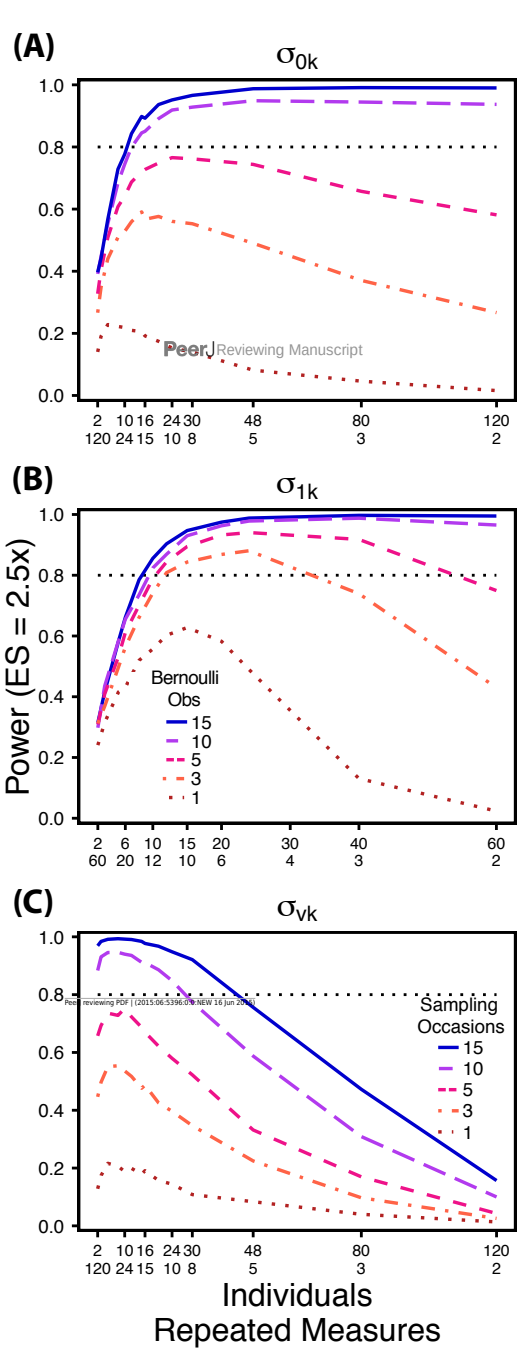
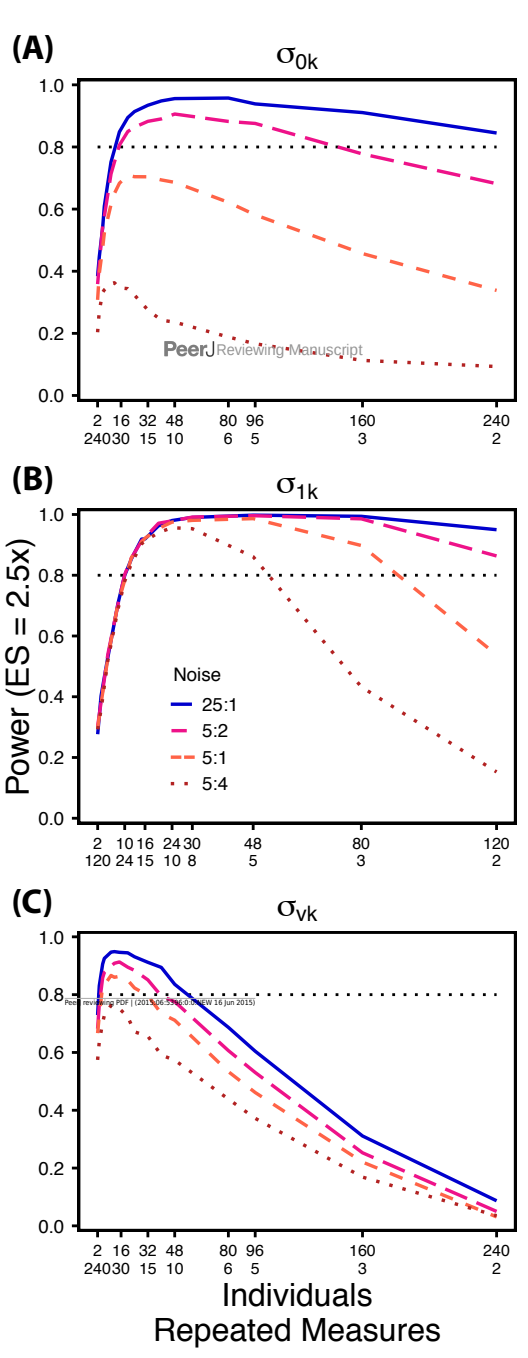

(A)
σ0k
1.0
0.8
0.6
0.4
0.2
0.0
2 10 16 24 30 48 80 120
120 24 15 10 8 5 3 2
PeerJ Reviewing Manuscript
(B)
σ1k
Power (ES = 2.5x)
Bernoulli
Obs
15
10
5
3
1
2 6 10 15 20 30 40 60
60 20 12 10 6 4 3 2
(C)
σvk
Sampling
Occasions
15
10
5
3
1
2 10 16 24 30 48 80 120
120 24 15 10 8 5 3 2
Individuals
Repeated Measures

Figure 6(on next page)

Power under increasing non-target variation

Figure 6: Power to detect differences by treatment in σ_0 (A) and σ_1 (B) under increasing variation in σ_v; σ_v (C) under increasing variation in σ_0. Noise is given as the ratio of effect size to variation in the non-target variance parameter. In (A) and (B) ratios of individuals to sampling occasions follow figures 2C and 3C respectively. In (C) ratios of individuals to Bernoulli observations follows figure 4C.


(A)
σ0k
1.0
0.8
0.6
0.4
0.2
0.0
PeerJ Reviewing Manuscript
2 16 32 48 80 96 160 240
24030 15 10 6 5 3 2
(B)
σ1k
Power (ES = 2.5x)
Noise
25:1
5:2
5:1
5:4
2 10 16 24 30 48 80 120
120 24 15 10 8 5 3 2
(C)
σvk
Peer reviewing PDF | (2015:06:5530:0:0:NEW 16 Jun 2015)
2 16 32 48 80 96 160 240
24030 15 10 6 5 3 2
Individuals
Repeated Measures