Simulating physiological flexibility in the acute glucocorticoid response to stressors reveals limitations of current empirical approaches

Conor C. Taff

Cornell Lab of Ornithology and Department of Ecology & Evolution

cct63@cornell.edu

ABSTRACT

1. Wild animals often experience unpredictable challenges that demand rapid and flexible responses. The glucocorticoid mediated stress response is one of the major systems that allows vertebrates to rapidly adjust their physiology and behavior. Given its role in responding to challenges, evolutionary physiologists have focused on the consequences of between-individual and, more recently, within-individual variation in the acute glucocorticoid response. However, empirical studies of physiological flexibility are severely limited by the logistical challenges of measuring the same animal multiple times.
2. Data simulation is a powerful approach when empirical data are limited, but has not been adopted to date in studies of physiological flexibility. In this paper, I develop a simulation that can generate realistic acute glucocorticoid response data with user specified characteristics. Simulated animals can be sampled continuously through an acute response and across as many separate responses as desired, while varying key parameters.
3. Using the simulation, I develop several scenarios that address key questions in physiological flexibility. These scenarios demonstrate the conditions under which a single glucocorticoid trait can be accurately assessed with typical experimental designs, the consequences of covariation between different components of the acute stress response, and the way that context specific differences in variability of acute responses can influence the power to detect relationships between the strength of the acute stress response and fitness. I also describe how to use the simulation tools to aid in the design and evaluation of empirical studies.
4. Recently there has been a great deal of interest in understanding the causes and consequences of physiological flexibility, but empirical data are often extremely limited and traditional sampling methods are poorly designed to directly address the questions

of interest. This simulation represents a way forward by revealing critical aspects of physiological flexibility and by creating a tool for designing better empirical studies that integrate simulation and theory with data collection.

Keywords: acute stress response, physiological flexibility, glucocorticoids, evolutionary endocrinology

INTRODUCTION

Animals live in a dynamic environment in which they regularly encounter unpredictable challenges. Successfully navigating these challenges often requires the ability to rapidly adjust behavior and physiology to match current conditions. For vertebrates, the glucocorticoid mediated stress response plays a major role in coordinating these changes when stressors are encountered (Sapolsky et al., 2000; Wingfield et al., 1998) and similar rapid response systems mediate changes in other taxa (Taborsky et al., 2020). Because of the central role that this response plays in coping with challenges, a great deal of research effort over the past 15 years has focused on understanding whether between-individual differences in the magnitude of this response predict coping ability and, ultimately, fitness (Breuner et al., 2008; Schoenle et al., 2020).

More recently, a series of conceptual papers have asked whether the degree of within-individual variation in glucocorticoid modulation (i.e., endocrine flexibility) across different contexts or in response to different stressors might also be an important predictor of performance (Hau et al., 2016; Lema & Kitano, 2013; Taff & Vitousek, 2016; Wada & Sewall, 2014). Perhaps the major limit to empirical progress, especially for within-individual variation, is the logistical difficulty of accurately characterizing the functional shape of the acute physiological stress response for an individual during a single acute response and across multiple acute responses occurring under different conditions. Often these measures are strictly limited by the number of samples that can safely be taken from an animal during a single capture and the number of repeated captures that are possible (but see Koolhaas et al., 2011). Given these limitations, data simulation is a powerful tool that could complement empirical work in this area, but that has not yet been applied to studies of endocrine flexibility.

Several recent papers have suggested that physiologists interested in endocrine flexibility should adopt a within-individual reaction norm approach (e.g., Hau et al., 2016; Taff & Vitousek, 2016). This approach has been widely adopted in studies of behavioral flexibility where statistical methods and empirical progress have developed synergistically (e.g., Araya-Ajoy et al., 2015; Dingemanse et al., 2010; Westneat et al., 2015). This field has also benefited from simulation studies to evaluate optimal study design (Pol, 2012) and packages that can create artificial datasets with desired patterns of between, within, and residual variance to evaluate the consequences of different patterns of variation on the ability to detect effects (see SQuID package, Allegue et al., 2017). While these approaches are powerful, they have proven difficult to apply directly to endocrine flexibility data for two reasons. First, simulation studies suggest that many patterns may only be detectable with a level of repeated sampling that is possible for many behaviors (especially when collected autonomously), but that is currently not possible for most studies of endocrine flexibility. Second, and more fundamentally, these papers often focus on somewhat discrete measures of behavior (e.g., aggression score or activity level), whereas for acute glucocorticoid responses, the functional shape of the response itself may be the important trait and it may not be possible to summarize variation in the shape of the response with a single measure.

The function valued trait (FVT) framework is an alternative approach that explicitly considers the functional shape of a biological response (Gomulkiewicz et al., 2018; Kingsolver et al., 2015; Stinchcombe et al., 2012). While FVT approaches have been suggested for studies of endocrine flexibility (Taff & Vitousek, 2016), I am not aware of any papers that have applied this framework to empirical data on acute glucocorticoid responses, probably because sufficient data are not available. Conceptually, however, this approach is a better match to the acute glucocorticoid response, because the shape of a response curve is explicitly considered as the phenotypic trait of interest. In some cases, it may make sense to estimate particular parameters of the curve (e.g., maximum rate of increase and maximum value reached) and then treat those parameters as phenotypic values for downstream analysis, although statistical methods also exist to analyze the shape of the entire curve directly without the need to extract discrete parameters (Kingsolver et al., 2015). This approach has been used to study a variety of phenotypes where values can be measured continuously or pooled across many individuals from the same group to accurately estimate the shape of a curve (see Table 1 in Stinchcombe et al., 2012). Applying the technique to endocrine flexibility at the within-individual level faces the same empirical challenges described for within-individual reaction norms above. Note that FVT and within-individual reaction norms approaches are not necessarily incompatible, but they have largely developed separately.

The recognition that characterizing the functional shape of an acute stress response is challenging goes back to the earliest studies conducted in wild animals. Early studies often employed various control groups and sampled individual animals at a variety of time points over a long period in order to describe the full response curve for a particular group (e.g., a species or a breeding stage, Wingfield et al., 1992). These validations were considered essential to characterize key parameters of the acute response for each group being studied (i.e., baseline, rate of increase, maximum level, time of peak, and area under the curve; John Wingfield, *personal communication*). The challenge of estimating these parameters becomes much more difficult when trying to describe the response for an individual animal rather than for a group, because glucocorticoids can often only be measured at two or three time points and only a small number of times per animal (e.g., Vitousek et al., 2018). Because these studies require an estimate for each individual, the solutions used by older studies that added additional animals to allow for sampling at more time points are not available.

For individual based studies, the most common approach to this problem is to standardize measurements as much as possible by measuring animals at the same time of the day during the same context, and by taking blood samples at standard times (often <3 and 30 minutes after capture) to characterize baseline and stress-induced glucocorticoids. This standardization allows for comparison between individuals, but in some cases it may also completely obscure the ability to detect variation in certain characteristics of the acute response curve. For example, if the speed (rate of initial increase) and scope (maximum value) of the acute response vary independently, samples taken at only two time points cannot accurately capture variation in either parameter. Indeed, several discussions in recent years about methods such as the '3 minute rule' and the relative merits of 'area under the curve' versus time point measures of glucocorticoids are fundamentally related to a recognition of the importance of understanding variation in the functional shape of stress responses and whether different components of that shape covary within individuals (e.g., Cockrem & Silverin, 2002; Small et al., 2017).

One of the characteristics of both the within-individual reaction norm and FVT literature is that empirical work has proceeded in very close coordination with simulation and statistical method development. In contrast, studies of endocrine flexibility often point to these methods, but don't address the ways that the particular logistical challenges of hormone measurement might necessitate different empirical approaches. I believe this is one reason that there are currently more conceptual papers arguing for a reaction norm approach to endocrine variation than there are empirical papers actually applying the approach. While many of the tools developed in these related fields are transferable, studies of physiological flexibility would benefit from a focus on analysis development and testing that explicitly incorporates the particular details and challenges of these questions. One way to accomplish these goals is to use simulations, but to my knowledge no studies of physiological flexibility have developed simulations of the acute stress response that address the issues discussed above.

Data simulation is a powerful approach for several reasons. Because true parameter values (e.g., maximum glucocorticoid level) are known, it is possible to evaluate how well different study designs and analytical choices perform in recovering true patterns and how sensitive those designs are to different assumptions. Thus, simulation can tell us whether the study designs we use can *in principle* detect the patterns we predict given realistic effect sizes. Simulated data can also identify conditions under which current study designs will perform well or poorly. For example, if simulations suggest that the baseline paired with stress-induced paradigm only works well when the speed and scope of responses are positively correlated, then empirical work could seek to determine the degree of correlation for a particular study system as justification for the approach. This ability to highlight key assumptions and create data sets with known properties has the potential to both provide insight into physiological flexibility directly and to guide empirical work by improving study design and identifying key areas for subsequent sampling. In the rest of this paper, I develop a simple simulation of acute physiological stress responses and then briefly illustrate several possible applications of the simulation.

MATERIALS AND METHODS

DESCRIPTION OF THE SIMULATION

I developed a set of functions in R version 4.0.2 (R Core Team, 2020) to generate acute physiological response curves. This simulation makes no assumptions about the mechanistic process that results in the shape of a glucocorticoid response. Rather, parameters are sampled to generate curves that are similar in shape and degree of variation to empirically observed responses (Figure 1). This simulation is designed to create data sets with realistic structure that can be used to better design and plan studies of physiological flexibility, to evaluate power of current study designs, and to evaluate the sensitivity of sampling regimes to any number of modifications to the shape of glucocorticoid response curves (e.g., changing covariation patterns between different features of the response). I explore a small number of scenarios in the next section, but I expect that many other scenarios can be addressed with these tools. For illustration purposes, I refer to simulated glucocorticoid responses, but the simulation applies equally well to any physiological mediator of a rapid response. The package can be installed in R using the following command.

```r
devtools::install_github("cct663/simcoRt")
```

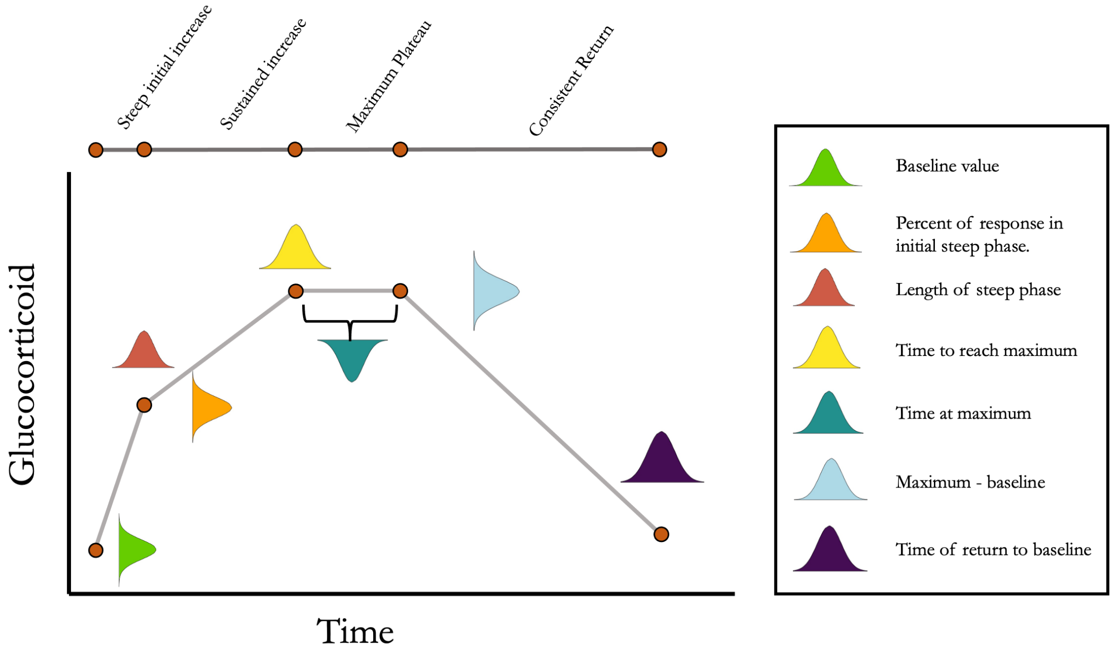

Figure 1: Conceptual illustration of the structure of the simulation. For each simulated animal, seven parameters are sampled from a multivariate normal distribution. Together, these seven parameters define the turning points in an acute response curve. The mean and standard deviation for each parameter can be set along with the degree of covariation between each pair of parameters. Note that the simulation can easily be simplified as desired by setting some parameter mean or standard deviations to zero.

The simulation is constructed as two main functions with several minor functions for downstream analysis. Detailed descriptions of the arguments to each function are included with the package documentation. Briefly, function `cort_sim1` samples the parameters shown in Figure 1 from an arbitrary number of animals. These parameters are sampled from a multivariate normal distribution with user specified mean, variance, and covariance for each parameter. I consider these values to be the ‘true,’ unobserved, phenotype of the animal (setting aside the question of whether or not a ‘true’ physiological phenotype exists).

A second function, `cort_sim2`, starts with a population of animals generated from `cort_sim1` and samples observed acute glucocorticoid responses an arbitrary number of times for each animal. Two sources of variation in the observed relative to true parameter values can be specified. First, within-individual variation in expression is represented by specifying what amount of variation in the observation of each parameter is determined by the true value and what amount is determined by an additional randomly sampled response, based on the population parameters (this additional sampling maintains the user specified covariance structure of the population). After sampling the parameters, values are interpolated for each one minute time point and a localized regression is fit to create a smoothed curve that represents the observed glucocorticoid response. From this expressed response, individual data points are then collected at user specified times that would reflect an empirical study design (e.g., 2, 30, and 60 minutes). Additional noise can be added to these data points to represent measurement error (e.g., assay error).

The function also generates a simulated performance (e.g., fitness) measure, based on the underlying true values. Data reflecting the true phenotypic values, the repeated expression of acute responses, and the observed time points can then be used in downstream analyses with any standard statistical approaches or software. For example, a user could perform an analysis to ask whether a known relationship between fitness and a particular true parameter is recovered in a study that includes only measures taken at particular time points. An additional convenience function summarizes the output of a simulation run in a multi-panel plot (Figure 2).

Finally, given recent interest in estimating the repeatability of glucocorticoid regulation (Cockrem, 2013; Hau et al., 2016; Taff et al., 2018), I also included a function that takes input from `cort_sim2` and calculates the observed repeatability of several measures using package `rptR` (Stoffel et al., 2017). Full details are included in the package documentation, but this function returns repeatability for each individual time point specified in the down

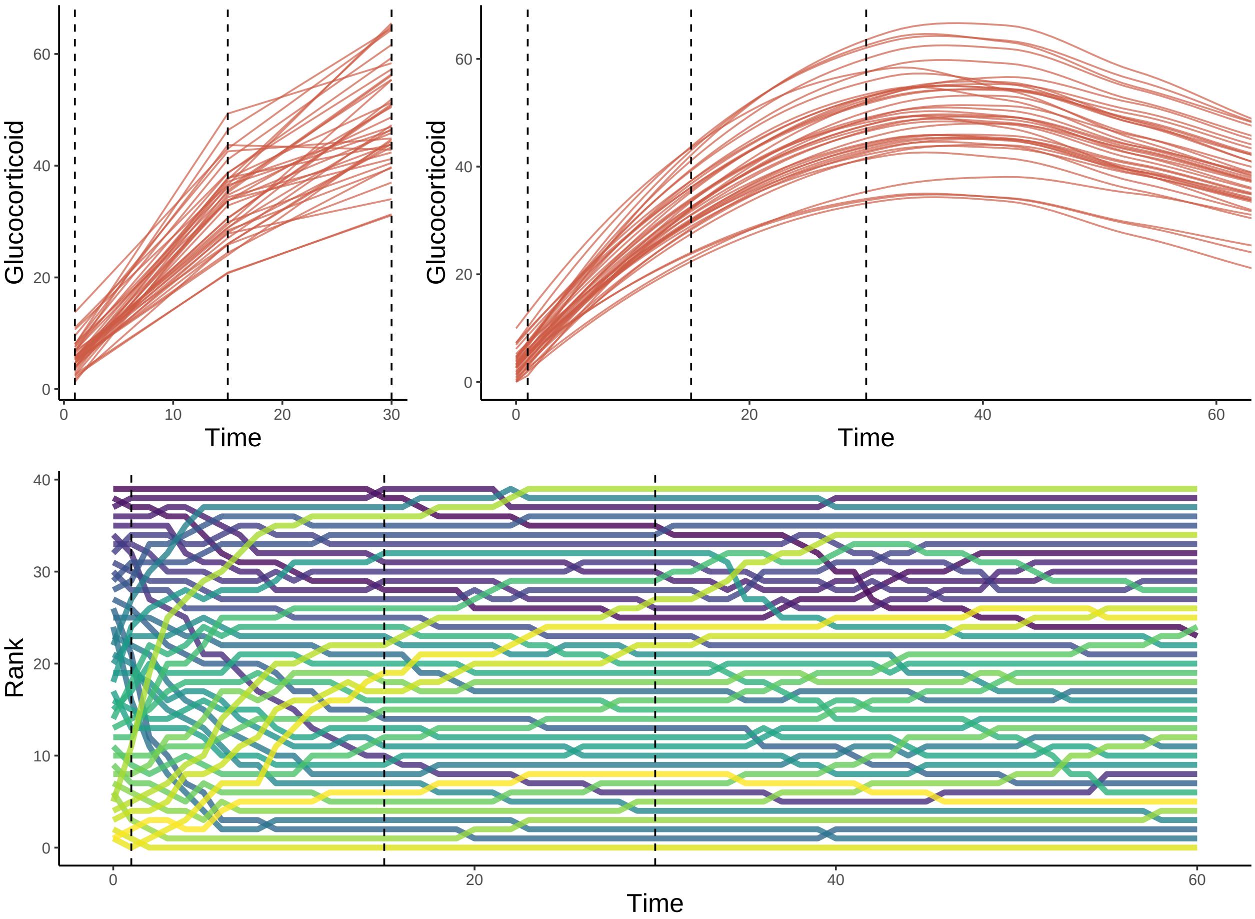

Figure 2: Example of simulation output with default settings. Panel A shows the downsampled data set for this run with samples collected at 1, 15, and 30 minutes in this case. Panel B shows the full observed response curve for each animal. Panel C shows the rank order of glucocorticoid level at each time point for each animal. In each panel, the vertical dashed lines represent the three time points that might have been measured in a typical empirical study. Note that individuals in the top panels do not match perfectly because measurement error is added to the downsampled dataset in panel A.

sampled data set, profile repeatability (Reed et al., 2019), and repeatability for area under the curve calculated as both increase ($\mathrm{AUC_I}$) and ground ($\mathrm{AUC_G}$) approaches (Pruessner et al., 2003). For each AUC measure, the function returns repeatability for the full time course, for an estimate using only the observed values in the down sampled data set, and for the full data set constrained to the time period encompassing the observed data points. Simple plots illustrating repeated samples from the same individuals are also returned by default. I do not develop an example of repeatability in this manuscript, but the functions here could be used to determine the impact of different study design choices on repeatability estimates.

APPLICATIONS OF SIMULATION

The goal of this simulation is to provide a flexible tool that can produce realistic datasets of physiological flexibility for a variety of different systems and scenarios. As such, there are many possible applications and here I briefly highlight a few possibilities. These are by no means exhaustive, and I hope the simulation will be a useful tool to guide empirical work for specific hypotheses and study systems. Within each scenario, I have illustrated how the simulation functions might be used to address the particular question of interest, but I have not fully explored all the possible permutations of parameters systematically, because these will depend to a large extent on the empirical details of the system being studied. A complete set of reproducible code to create all of the examples presented in this paper is available on GitHub (https://github.com/cct663/speed_vs_scope).

Simulating empirically parameterized data In order for simulation to be useful, we should be able to create artificial datasets that have similar characteristics to empirical data for different systems. Simulating realistic data provides a starting point for evaluating different study designs and the consequences of changes in different assumptions or parameters. Simulating realistic data is also useful because it can aid in study design or be used as a basis for pre-registered reports that demonstrate the feasibility of a planned study before data are ever collected. Simulated data can be created and entered in a complete analysis pipeline, with empirical data substituted later. In addition to helping to design better studies, this approach has the advantage of increasing the transparency and reliability for studies of physiological flexibility, by making analysis choices and predictions clear before data are collected.

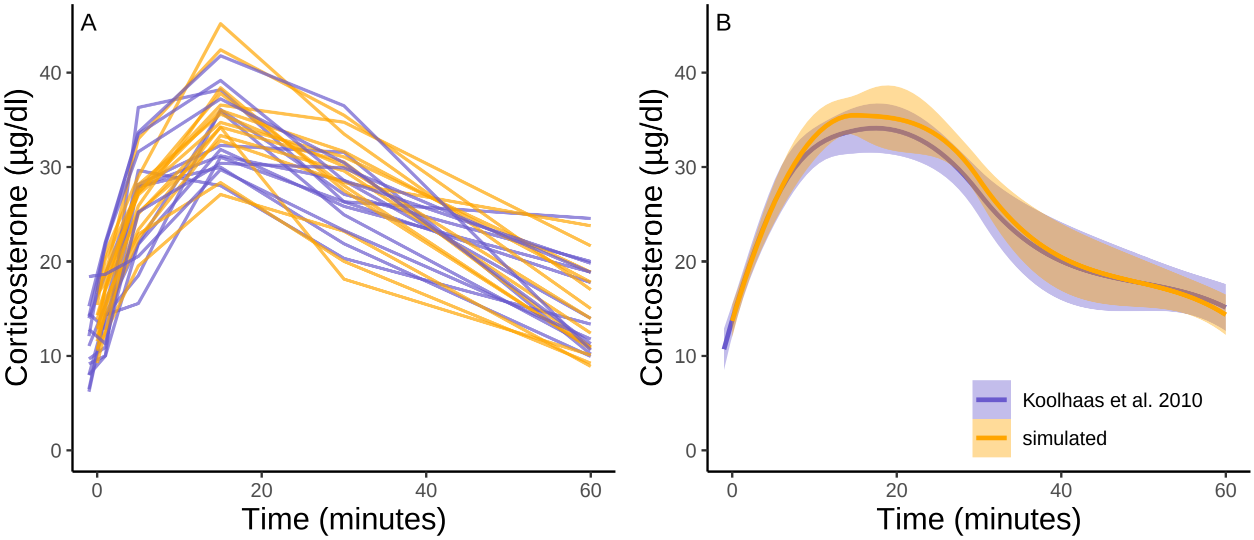

Figure 3: Panel A shows the acute corticosterone response for measured (blue) or simulated (orange) rats measured at five time points. Panel B shows the mean and standard error of the two datasets. Empirical data are extracted from Koolhaas et al. 2010 Figure 6 using the WebPlotDigitizer tool.

To demonstrate this utility, I have redrawn data from Koolhaas et al. (2010). As part of that study, a series of corticosterone measurements were collected during and after an acute stressor from 14 laboratory rats *Rattus norvegicus* using permanently implanted jugular vein canulae. I next simulated data using the functions described above starting with the input values calculated directly from the empirical data. The simulation creates a new dataset that has similar variation and patterns to the empirical data (Figure 3A) along with a population wide corticosterone response curve shape that closely matches the empirical data (Figure 3B). In this case, the plotted simulation data include the same number of animals sampled at the same time points as the empirical data, but these sampling points and total sample size can easily be changed as desired. The parameterized simulation can now be used to test the sensitivity of any number of experimental designs before additional data is collected.

Accurately measuring a single glucocorticoid trait Single time point measures of glucocorticoids are often interpreted as representing meaningful variation between individuals. For example, variation in the level of glucocorticoids after 30 minutes of standardized restraint is typically interpreted as variation in the magnitude of the stress response (Taff et al., 2019). However, this interpretation rests on assumptions that are rarely explicitly tested with empirical data. For example, the time chosen to take a stress-induced sample is often assumed to be either at the species peak or during a plateau period after the species peak. In some early studies, great care was taken to determine an average population level peak time (Wingfield et al., 1992), but many studies adopt the widely used 'standard' time of 30 minutes post capture without extensive validation (compiled in Vitousek et al., 2019). While there is a general assumption that sampling later than the peak is acceptable (and perhaps preferable) because animals will be sampled during a relatively stable high plateau, there is little empirical data to evaluate this assertion or to determine how much under or overshooting the species peak timing might influence inferences. Furthermore, even when the average peak timing is well established, differences in the amount of between-individual variation in the time to reach the peak or in peak values are common across species and even in different life history stages within species (Wingfield et al., 1992). The combinations of these patterns of variation could have major consequences on the accuracy of single point estimates taken at 30 minutes, but these questions cannot be addressed directly with empirical datasets where the true underlying values of each individual are unknown.

Here, I simulate a simple scenario exploring the consequences of variation in each of these parameters on the accuracy of estimating between individual differences in maximally expressed glucocorticoids during an acute response. For purposes of this illustration, I consider a single study design in which animals are sampled at 30 minutes. Using this design as a starting point, I systematically vary i) the timing of the population average peak (15, 30, or 45 minutes), ii) the amount of variation in maximum glucocorticoid levels reached, iii) and the amount of variation in the number of minutes taken to reach peak levels. All other variables in the simulation are constrained to be invariant between individuals in the population (e.g., all individuals have identical baseline glucocorticoids in this case), though I consider cases in which multiple aspects of the rapid response are correlated with each other in the next section. I included moderate within-individual variability and a small amount of assay error across all iterations. For each combination of parameters, I simulated 200 animals and estimated the R^2 value from a regression of the observed estimates of glucocorticoid levels at 30 minutes to the true known values. This simulation is likely a best case scenario because it eliminates many sources of variation or noise that would be present in real data, but it illustrates the effect of variation in these three key parameters even when the exact same sampling design is employed.

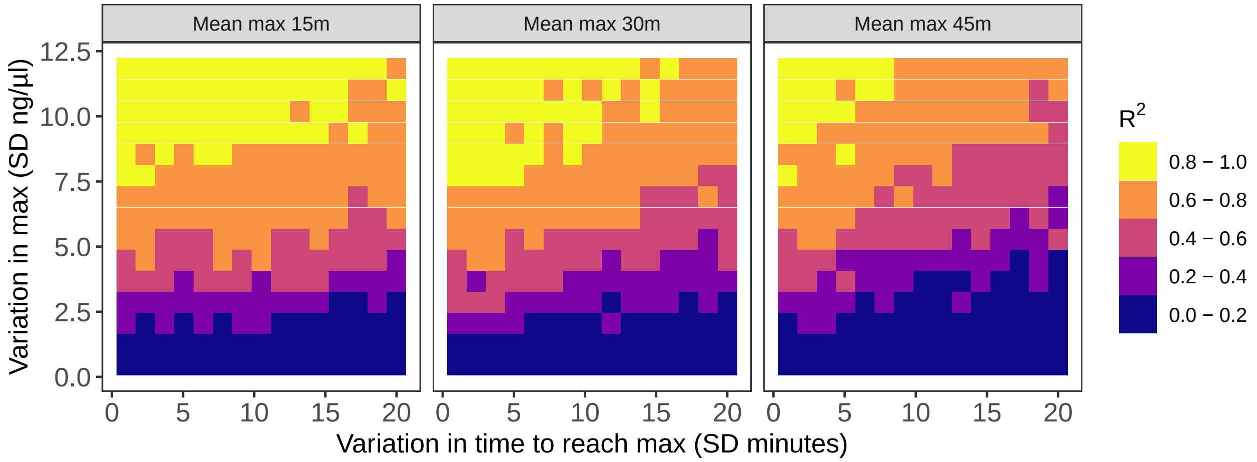

Figure 4: Results of simulation runs with different amounts of between-individual variation in the time to reach maximum glucocorticoid levels and in the maximum level reached. Simulations are run with samples taken at 30 minutes on populations with an average peak time of 15 minutes (left), 30 minutes (center), or 45 minutes (right). Each grid cell is the R^2 value from the regression of observed glucocorticoids at 30 minutes to true maximum levels in a simulation of 200 individuals.

Results of this simulation are summarized in figure 4. The amount of between-individual variation in the maximum glucocorticoid value has a profound effect on the ability to detect true maximal levels with samples taken at 30 minutes. In one sense, this result is unsurprising because it is intuitive that large differences would be easier to detect, but there are important consequences of this fact for interpreting studies that seek to link between-individual variation in the magnitude of the stress response with other traits. For example, the magnitude of the acute stress response often varies substantially across life history stages (Wingfield et al., 1992). Even if study designs are identical and maximum glucocorticoids are associated with performance, it will be easier to detect those patterns during life history stages with greater variation (see section on detecting fitness associations below). There is a weaker, but still substantial impact of variation in the time taken to reach maximum values on the accuracy of estimates in this simulation. Greater variation in the speed of the response reduces the accuracy of estimates of maximal values. Finally, the timing of sampling relative to the average population peak timing also influences accuracy. Measuring after the average peak time results in the most accurate estimates across a range of parameter values, while measuring before the average peak time produces the least accurate measures, particularly when there is also high variation in the time to reach maximum values between individuals. This simple example demonstrates clearly that the same experimental design will perform better or worse depending on the combination of glucocorticoid regulation parameters in the population being studied.

Exploring covariance between response components In reality, fully characterizing the acute glucocorticoid response requires more than identifying just the maximum value reached. Individuals may differ in baseline levels, rate of initial increase, the speed of reaching the maximum level, time spent at maximum, and the speed of return to baseline. Moreover, each of these components of the endocrine response could be positively or negatively correlated with each other within and between individuals. In these cases, measurements taken at particular time points contain information about multiple aspects of the response and without additional information it may be difficult to know what trait is being measured. The fact that each of these traits might be important and that they might covary has been discussed in a general sense (e.g., Baugh et al., 2013), but simulations are uniquely powerful for exploring under exactly what conditions time point measure of glucocorticoids can or cannot be used as indicators of these traits.

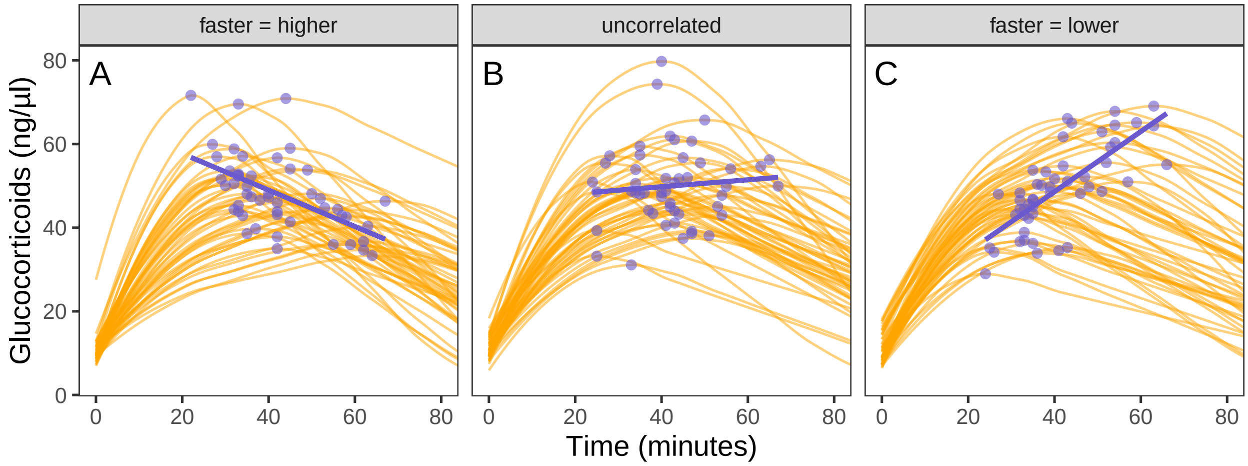

Figure 5: Simulated glucocorticoid responses in which the maximum value and response speed are positively correlated (A), uncorrelated (B), or negatively correlated (C). Orange curves show the full response for each individual. Blue points show the maximum value and time to reach maximum for each individual. Blue lines are simple linear regressions of speed and maximum value for each group. For clarity, only the first 40 individuals in each simulated dataset are plotted.

To illustrate this point, I explored the consequences of variation in the correlation between and relative amount of variation in just two aspects of the acute stress response: the maximum glucocorticoid level reached and the time required to reach the maximum level. For simplicity, I refer to the 'speed' of the response, but note that other aspects, such as the rate of initial increase, could also be considered as variation in the speed of response. When considering these two traits, a population of animals could plausibly display one of three patterns. Individuals that reach their maximum value faster might also reach higher values (figure 5A; simulation correlation = -0.6). Alternatively, the speed and maximum values might vary independently (figure 5B; correlation = 0). Finally, individuals that are faster responders might max out at lower glucocorticoid values (figure 5C; correlation = 0.6). While many researchers in this field might have intuitions about which of these scenarios is most likely to prevail, there is very little empirical data available to actually determine which is most common. Moreover, regardless of the specifics for this particular correlation, the general pattern and considerations presented here will apply in similar ways to correlations between other aspects of the acute stress response.

Using these three simulated populations as a starting point, I asked how well glucocorticoid

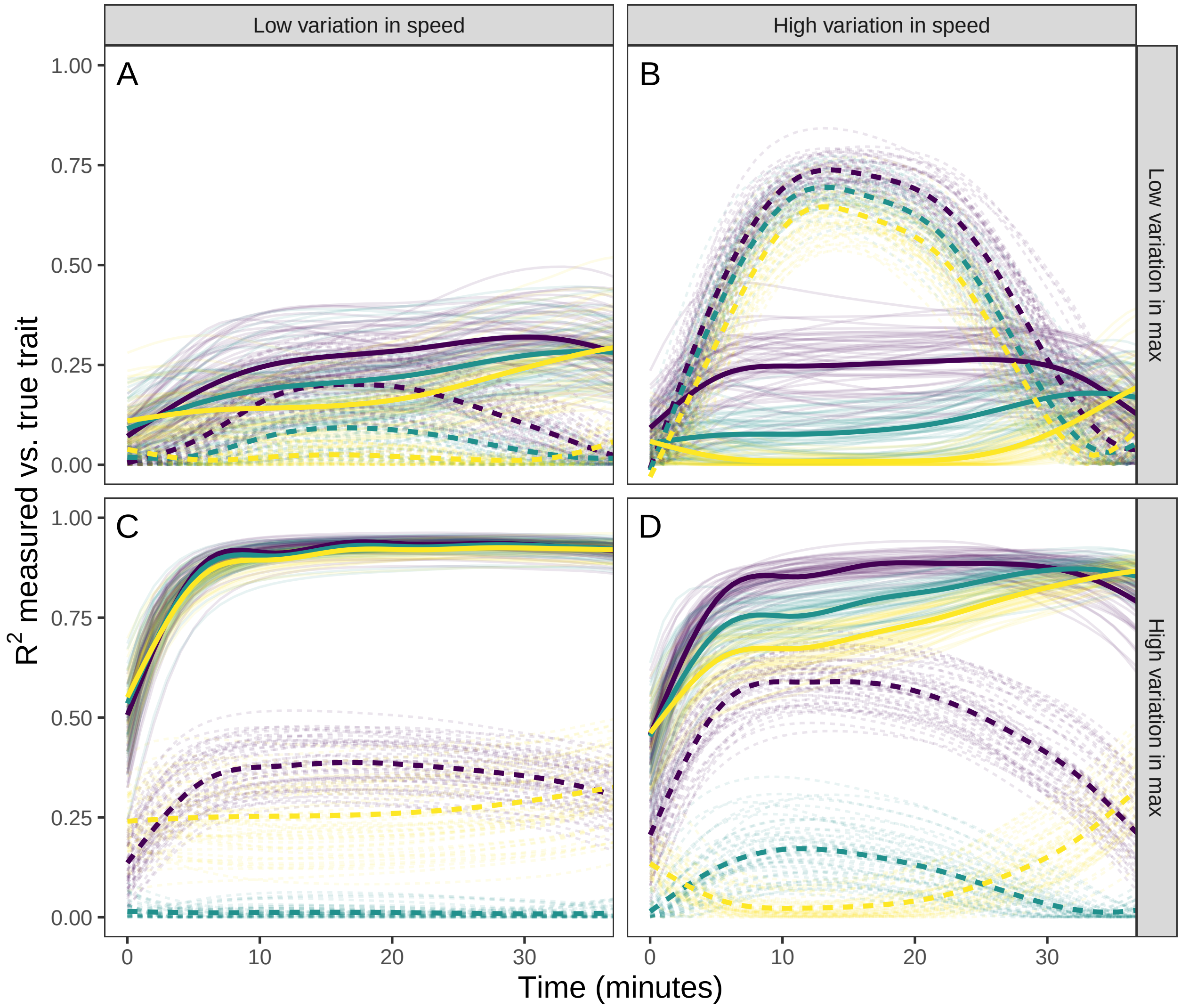

Figure 6: Relationship between single time point measures of glucocorticoids and the true value of either maximum level (solid lines) or the speed of the glucocorticoid response (dashed lines). Panels show results when the overall variation in maximum values and speed are both low (A), when one is low while the other is high (B and C), and when both are high (D). In each panel, three different simulation scenarios illustrate the patterns when speed and maximum value are positively correlated (purple), uncorrelated (teal), or negatively correlated (yellow). Faded lines show the results from each of 50 separate simulation runs and thick lines are the averages across all runs.

values measured at one timepoint reflected true trait values. For each population I set an average population level speed of 30 minutes with other values in the simulation set at their default value. For every time point from 0 to 35 minutes I fit two simple linear regressions of the measured value on the true speed and maximum value and extracted the R^2 value from the model. I repeated this simulation for all populations 50 times with 100 individuals sampled from the population each time. Finally, I repeated the entire set of simulations with each combination of low and high between-individual variation in the speed or maximum values (variation in speed: low = 2 minute SD, high = 12 minute SD; variation in maximum: low = 1ng/μl SD, high = 10ng/μl SD).

The time that samples were taken at, relative amount of variation in speed and maximum, and degree of correlation between the speed and maximum all had substantial impacts on the ability to infer true trait values from single time point glucocorticoid measures (figure 6). While these scenarios do not explore all possible parameter space, there are several clear conclusions that can be made. First, neither speed or maximum traits could be assessed accurately when between-individual variation in both traits was low (figure 6A). This is potentially important for interpreting apparent differences in glucocorticoid fitness relationships because between-individual variation is known to differ across life history stages (Wingfield et al., 1992). Second, accurately assessing variation in speed was much harder—if not impossible—with single measures.

It was only possible to accurately estimate speed when high between-individual variation in speed was coupled with low variation in maximal values, but this situation may be rare in natural populations. When speed was tightly correlated with maximum (figure 6D) it was sometimes possible to attain reasonable estimates of speed (figure 6C-D), but when speed was not correlated with maximum, single measures were not good indicators of variation in speed (figure 6A, C-D). Finally, measuring variation in maximum values was much easier under many conditions (figure 6C-D), but the accuracy of assessment of maximum values was also negative impacted by variation in speed and the degree of this impact differed depending on the correlation between the two traits (figure 6). Beyond the specifics of this particular example, what these results demonstrate clearly is that understanding what aspect of the glucocorticoid response is being measured by any particular study design depends on extensive knowledge of the overall shape and amount of variation in different aspects of the acute stress response.

Detecting links between fitness and responses A common goal of recent studies is to establish whether variation in glucocorticoids is associated with fitness or some proxy for fitness (Schoenle et al., 2020). While there has been a great deal of discussion about the extent to which these relationships might differ with life history characteristics or between breeding stages, there has been relatively little consideration of the way that methodological limitations might limit the ability to detect these relationships even when they exist.

Here, I imagine a simple scenario in which the 'true' maximum glucocorticoid level during an acute response explains 80% of the variation in fitness (clearly this is unrealistically high, but it is chosen for illustration only). I next construct a study in which researchers measure 50 individuals using a typical stress-induced (30 minute) sampling protocol. For simplicity, I set the other parameters in the simulation at their default values. Keeping the study design constant, I ask whether the glucocorticoid-fitness relationship can be recovered for two hypothetical populations that have low or high between-individual variation in maximum glucocorticoid levels. For each of these two populations, I ask how the ability to detect glucocorticoid-fitness relationships changes with different amounts of within-individual variation in acute response expression and with differing amounts of measurement error. For each combination of parameters, I simulated 50 populations and fit a simple linear regression model with observed glucocorticoid levels at 30 minutes as a predictor of fitness to ask whether the true glucocorticoid-fitness relationship was recovered.

Several patterns can be identified by examining the results of this simulation. First, the correlation between the true maximum glucocorticoid value and fitness does not differ for populations simulated with high or low between-individual variation (figure 7A-D). In all cases, however, the observed correlation is lower than the true correlation and always lowest in the population with low between-individual variation. The ubiquity of this pattern is a product of the simulation structure, because adding measurement error or within-individual variation effectively adds noise to the true correlation. It is important to note that in the

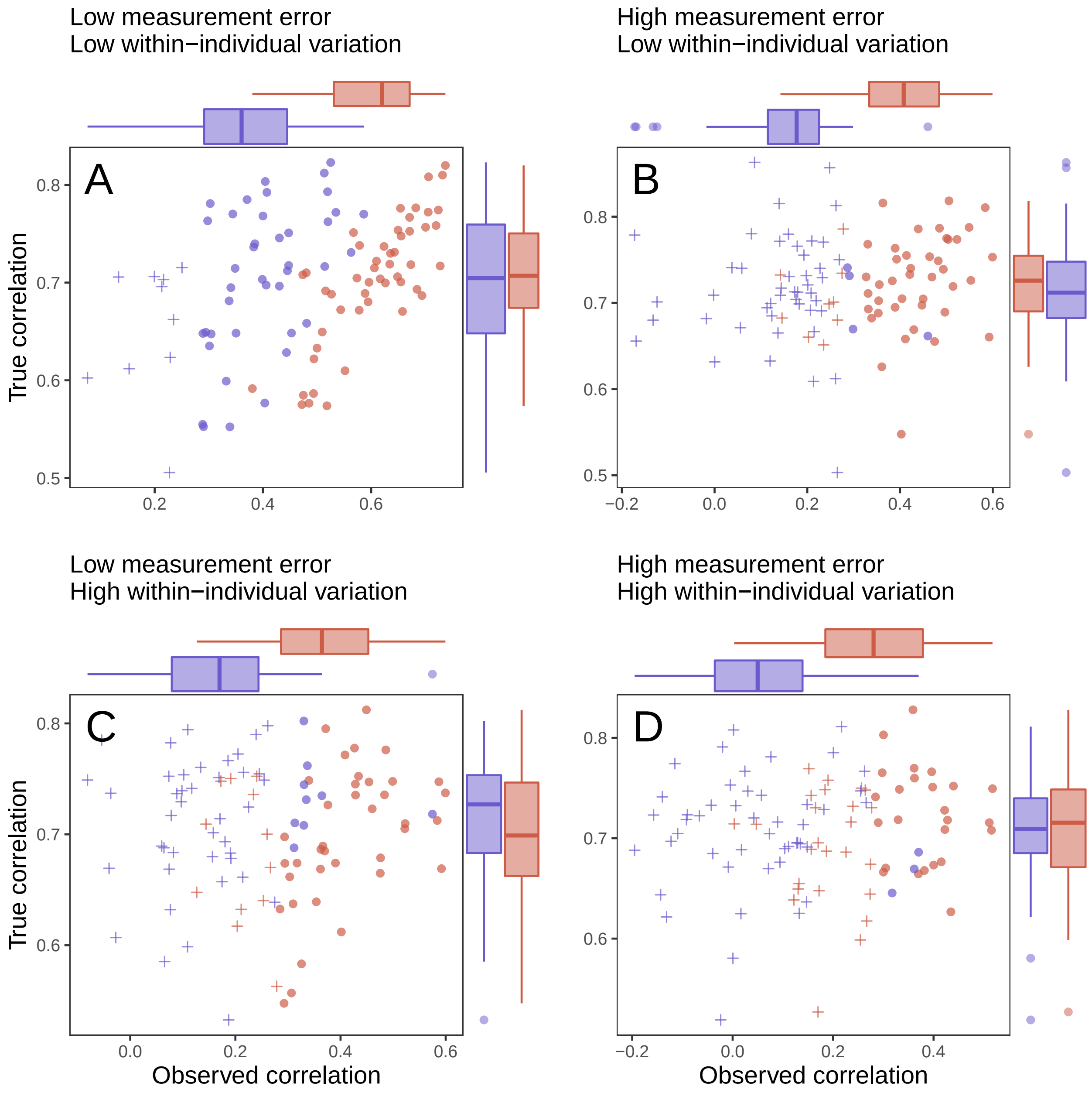

Figure 7: Relationship between observed maximum glucocorticoid values and fitness for simulated populations that have low between-individual variation (blue) or high between-individual variation (red). Each point is the result of a separate simulation of 50 individuals using the settings described in the text. Filled circles are simulations in which observed glucocorticoid values at 30 minutes were significantly correlated with fitness and crosses are simulations in which the relationship was not significant. Panels illustrate conditions with low measurement error (A, C) versus high measurement error (B, D) and low within-individual variation (A, B) versus high within-individual variation (C, D). For each simulation, the correlation between true maximum glucocorticoids fitness is plotted on the y-axis and the correaltion with observed values is plotted on the x-axis.

real world, it is unlikely that this pattern would be so universal, because unmeasured variables could influence both fitness and glucocorticoids. For example, if habitat quality directly alters fitness and glucocorticoids, the observed correlation could be stronger than the 'true' correlation. Thus, interpretation of these results should be made cautiosly in light of the simplicity of the simulation compared to real world conditions.

Nevertheless, general patterns illustrated by the simulation are likely to pertain across a wide range of conditions. In this case, it is easiest to detect significant glucocorticoid-fitness relationships when both measurement error and within-individual variation are low (figure 7A). It becomes harder to detect these true relationships when either measurement error (figure 7B) or within-individual variation (figure 7C) are high, but even in these more challenging situations the relationship can be detected the majority of the time if between-individual variation in maximum levels is high. When both measurement error and within-individual variation are high, it is nearly impossible to detect glucocorticoid-fitness relationships with low-between individual variation, but in populations with high between-individual variation the relationship is still detected in about half of the simulations.

The fact that low between-individual variation in maximum glucocorticoids makes it harder to detect true glucocorticoid-fitness relationships across a wide range of conditions has important consequences for interpreting empirical results. Many studies have demonstrated different relationships (or lack thereof) between corticosterone and fitness at different life history stages (Bonier et al., 2009; Vitousek et al., 2018), but it is also well known that the absolute amount of between individual variation in glucocorticoid traits varies considerably at different stages (Wingfield et al., 1992). Our simulation demonstrates that the power to detect true relationships will differ drastically across these conditions even with identical study designs and samples sizes, suggesting that great care is needed to conclusively differentiate true differences in glucocorticoid-fitness relationships across contexts from statistical artefacts.

Designing optimal sampling strategies One of the major benefits of simulating glucocorticoid response curves will be the ability to design optimal sampling strategies before data are collected. A simulation can be constrained to match any real world limitations (e.g., maximum number of samples possible per individual) and then explored to determine how to best allocate sampling resources. The specifics of this task will vary considerably with the study system and question being addressed, but here I illustrate one possible application. Consider an experiment in which the acute glucocorticoid response of a treatment group and control group are compared after some experimental manipulation. The details of the manipulation are unimportant here, but suppose that the prediction is that this manipulation should result in a difference in the speed of the corticosterone response between our two groups, such that the treatment group will reach it's maximum glucocorticoid value faster than the control group, but will not differ in the maximum value itself. I have implemented this difference by simulating two populations in which the treatment group has a steeper initial slope and also reaches the maximum value faster (figure 8). Any number of possible hypotheses for a particular study system could be specified following a similar approach.

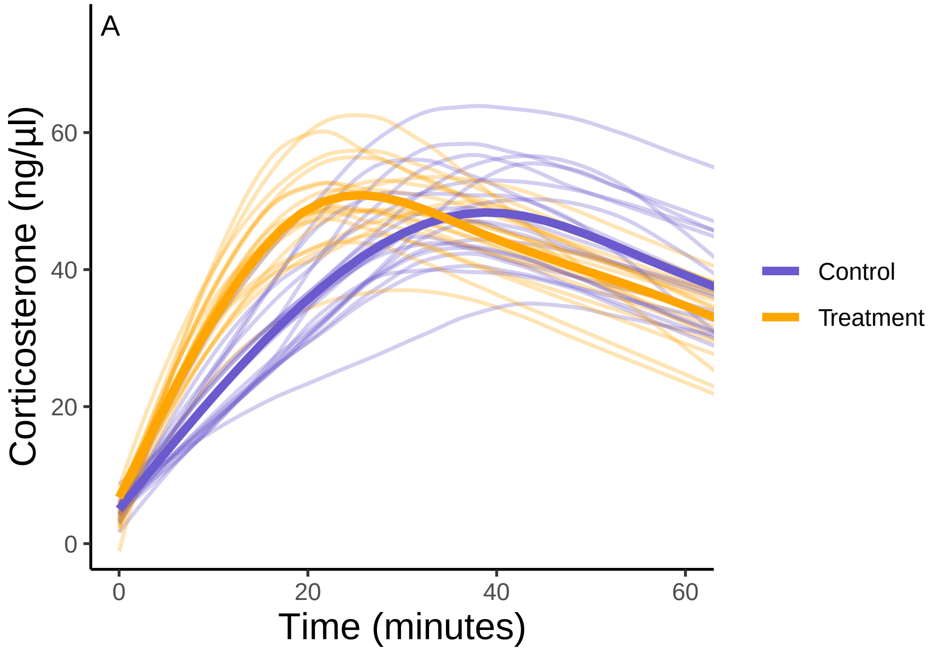

Figure 8: Simulated data for a hypothetical control (blue) and treatment (orange) group. Faded thin lines show the acute response for each individual simulated (20 per group) and thick lines show the average response curve for each group.

Next, we can ask how well different study designs can detect this difference. Here we can impose any logistical constraints relevant to the study system. As an example, in this case we can only sample a maximum of 20 individuals per group, we can only sample each individual once post-treatment, and during that single sampling event we can take a blood sample at a maximum of two different time points, resulting in a total of 80 data points. Given these constraints, I compare three different sampling designs: i) a study in which every animal is sampled at 1 minute, 30 minutes, and 60 minutes, ii) a study in which two sampling times between 1 and 60 minutes are randomly chosen for every animal, iii) a study in which two sampling times are randomly chosen for each animal, but weighted more heavily around the range of times when maximum levels are expected to be reached for the population.

Note that the first sampling scheme closely mirrors the most common empirical design and in this case I have allowed an extra, third sample at 60 minutes, such that it includes 120, rather than 80, data points. For illustration purposes I sampled directly from the 'true' response curves in this example so that there is no additional measurement error added. To evaluate these schemes I compare estimates of the acute response curve for each group to the 'true' known curves shown in figure 8. Note that a more complete analysis of a sampling schemes performance should include many more iterations and full statistical comparisons, but the details here will be highly dependent on the study system and goals, so I provide this simple example to illustrate the approach rather than to make any more widely applicable conclusions.

In this case, the standard sampling scheme performs very poorly (figure 9A), with no differences detectable between the two groups, despite the fact that the treatment group reaches it's maximum value on average 12 minutes (~40%) faster than the control group. In contrast, both the random sampling and weighted sampling schemes detect differences in the shape of the acute response (figure 9 B & C). In this particular scenario, there is no clear difference between these two approaches. A few clear takeaways can be derived from these results. First, while strict standardization of the timing of samples has some clear advantages,

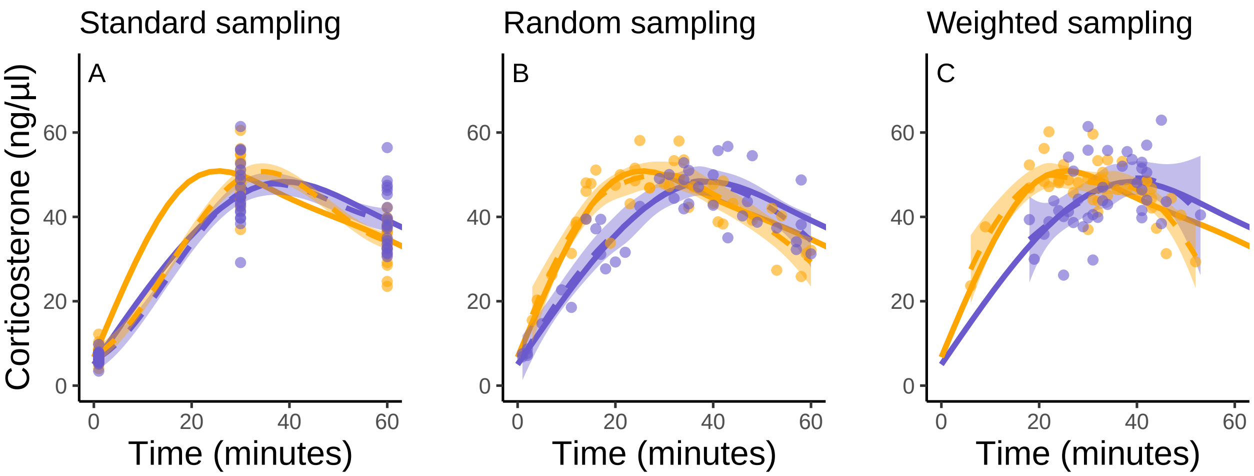

Figure 9: Three possible sampling schemes to compare two groups. For standard sampling (A), every individual is sampled at exactly 2, 30, and 60 minutes. For random sampling (B) each individual is sampled at two random points between 1 and 60 minutes. For weighted sampling (C) two sampling times are chosen for each individual from a normal distribution with mean of 32 and sd of 9 minutes. In all three panels, solid lines are the true group averages, dashed lines are the estimates based on samples, and points are individual samples collected.

it also comes with costs and likely makes it nearly impossible to detect certain types of variation between groups or individuals. In this case, standardized sampling performed much worse than the other two approaches despite the fact that the analysis included 50% more data; it should be clear that no amount of additional sampling would allow that approach to detect this particular pattern of between group differences. Second, while it may be very difficult to accurately estimate the full shape of the acute stress response for *individuals*, the sampling schemes shown here demonstrate that it should be possible to describe these shapes accurately for groups (e.g., treatments, species, different contexts) even without extraordinarily large sample sizes. A similar argument about the power of randomly timed sampling has been put forward in the function valued trait literature (Gomulkiewicz et al., 2018), but this type of sampling scheme is rarely used in evolutionary endocrinology research. It is perhaps unsurprising that the few empirical papers that have emphasized the importance of different time courses (rather than only maximum) of the stress response have often focused on between group comparisons or investigated variation in the exact sampling time between individuals (e.g., Baugh et al., 2013; Small et al., 2017)

This simulation is particular to a single very specific scenario, but a similar scenario could be designed for any number of studies and any number of predictions about how the speed, scope, or other attributes of the glucocorticoid response are expected to change with a treatment or between different groups or species. Clearly, when estimating the timing of peak glucocorticoids, a simple baseline plus induced sampling scheme is sub optimal, but this scheme may be perform well in other situations where the maximum value is the target and there is relatively little variation in response time. Creating simulations like this before studies are conducted has the potential to increase the efficient use of researches time and funds, but also forces researches to think explicitly about quantitative predictions ahead of time. These simulations could be included as part of a study pre-registration, grant application, or registered report to demonstrate exactly what data collection and analysis approaches are planned and to justify those decisions.

DISCUSSION

While there has been increasing interest in understand within- and between-individual variation in the acute glucocorticoid response in recent years (Hau et al., 2016; Lema & Kitano, 2013; Taff & Vitousek, 2016; Wada & Sewall, 2014), the methods and data available to tackle these questions have changed relatively little. Many sophisticated statistical tools are now available and clear arguments have been made about the need to apply these approaches to endocrine traits, but relatively few empirical studies have effectively used these tools. Arguably, the biggest roadblock at the moment is the limited availability of empirical data needed to test hypotheses. Simulation offers one way forward, by allowing for more efficiently designed studies and by allowing researches to identify when the question of interest can *in principle* be answered with a given study design. Ideally, conceptual papers, empirical work, and simulation will proceed together to make progress in this field. The tools presented here only scratch the surface of the ways that data simulation can be applied to address pressing questions in evolutionary endocrinology.

Nevertheless, even the simple demonstrations included in this paper suggest several ways that simulation could help move the field forward. One of the main benefits of simulating datasets is identifying unmeasured properties and assumptions of currently available data that can become targets for empirical work. For example, I demonstrated that the covariation between different components of the acute stress response and the relative amount of variation in each of these can have profound effects on the ability to accurately measure any single component. Empirical work specifically designed to assess covariation and variance at different times could help to understand what conclusions we can reasonably draw from available data. One takeaway from these simulations is that variation in glucocorticoid-fitness relationships across seasons or life history stages can easily arise as a statistical artefact when between-individual variation in hormones also varies across the contexts. The simulation exploring different sampling designs also suggests that there are potentially gains to be made by considering more diverse sampling designs tailored to the particular research question and study system. While standardized sample collection timing has allowed for large scale comparisons in this field (Vitousek et al., 2019), it also creates clear blind spots to certain types of variation between groups.

In addition to providing insight in its own right, simulation has great potential to hone the design of future empirical studies by allowing for a principled analysis of various study design options and choices before costly data are collected and before animals are needlessly disturbed. For example, I showed that one of the most common sampling design schemes has essentially no ability to detect a difference in the speed of increase between two groups if they do not also differ in maximum values. It is perhaps not surprising to find that there is little published evidence for differences in the speed of the acute response when most study designs employed to date cannot *in principle* detect those differences. Across a wide range of disciplines there has been an increasing push for pre-registration, reproducible research, and transparent research practices (O'Dea et al., 2021). Simulation provides an opportunity for evolutionary endocrinologists to embrace these best practices by improving the quality of study design, allowing for more quantitative hypotheses and predictions, and providing a clear justification for experimental choices.

This package and paper is meant only as an initial exploration of the ways that simulation can be applied to evolutionary endocrinology. I have no doubt that many more scenarios and complications could be added on to each of the simple examples presented here. Furthermore, there is ample room to create more sophisticated simulations that incorporate realistic mechanistic processes or interactions with other molecules and other components of the stress response system. I hope that this work will be a starting point to build and improve on as we work to understand the importance of variation in these flexible response systems.

ACKNOWLEDGEMENTS

I would like to thank Michaela Hau and her lab for discussions about the ideas presented in this project. John Wingfield also provided an insightful overview of the history of study design in field based studies of the acute stress responses and pointed me to key references. Finally, I thank Maren Vitousek and her lab members for feedback and discussion on early versions of this project.

REFERENCES

Allegue, H., Araya-Ajoy, Y. G., Dingemanse, N. J., Dochtermann, N. A., Garamszegi, L. Z., Nakagawa, S., Reale, D., Schielzeth, H., & Westneat, D. F. (2017). Statistical quantification of individual differences (SQuID): An educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. *Methods in Ecology and Evolution*, *8*(2), 257–267.

Araya-Ajoy, Y. G., Mathot, K. J., & Dingemanse, N. J. (2015). An approach to estimate short-term, long-term and reaction norm repeatability. *Methods in Ecology and Evolution*, *6*(12), 1462–1473.

Baugh, A. T., Oers, K. van, Naguib, M., & Hau, M. (2013). Initial reactivity and magnitude of the acute stress response associated with personality in wild great tits (parus major). *General and Comparative Endocrinology*, *189*, 96–104.

Bonier, F., Moore, I. T., Martin, P. R., & Robertson, R. J. (2009). The relationship between fitness and baseline glucocorticoids in a passerine bird. *General and Comparative Endocrinology*, *163*(1-2), 208–213.

Breuner, C. W., Patterson, S. H., & Hahn, T. P. (2008). In search of relationships between the acute adrenocortical response and fitness. *General and Comparative Endocrinology*, *157*(3), 288–295.

Cockrem, J. F. (2013). Individual variation in glucocorticoid stress responses in animals. *General and Comparative Endocrinology*, *181*, 45–58.

Cockrem, J. F., & Silverin, B. (2002). Variation within and between birds in corticosterone responses of great tits (parus major). *General and Comparative Endocrinology*, *125*(2), 197–206.

Dingemanse, N. J., Kazem, A. J., Réale, D., & Wright, J. (2010). Behavioural reaction norms: Animal personality meets individual plasticity. *Trends in Ecology & Evolution*, *25*(2), 81–89.

Gomulkiewicz, R., Kingsolver, J. G., Carter, P. A., & Heckman, N. (2018). Variation and evolution of function-valued traits. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 139–164.

Hau, M., Casagrande, S., Ouyang, J. Q., & Baugh, A. T. (2016). Glucocorticoid-mediated phenotypes in vertebrates: Multilevel variation and evolution. *Advances in the Study of Behavior*, *48*, 41–115.

Kingsolver, J., Diamond, S., & Gomulkiewicz, R. (2015). Curvethinking: Understanding reaction norms and developmental trajectories as traits. *Integrative Organismal Biology. Hoboken (NJ): Wiley Blackwell*, 39–54.

Koolhaas, J. M., Bartolomucci, A., Buwalda, B., Boer, S. F. de, Flügge, G., Korte, S. M., Meerlo, P., Murison, R., Olivier, B., Palanza, P.others. (2011). Stress revisited: A critical evaluation of the stress concept. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1291–1301.

Koolhaas, J. M., De Boer, S., Coppens, C., & Buwalda, B. (2010). Neuroendocrinology of coping styles: Towards understanding the biology of individual variation. *Frontiers in Neuroendocrinology*, *31*(3), 307–321.

Lema, S. C., & Kitano, J. (2013). Hormones and phenotypic plasticity: Implications for the evolution of integrated adaptive phenotypes. *Current Zoology*, *59*(4), 506–525.

O'Dea, R. E., Parker, T. H., Chee, Y. E., Culina, A., Drobniak, S. M., Duncan, D. H., Fidler, F., Gould, E., Ihle, M., Kelly, C. D.others. (2021). Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biology*, *19*(1), 1–5.

Pol, M. van de. (2012). Quantifying individual variation in reaction norms: How study design affects the accuracy, precision and power of random regression models. *Methods in Ecology and Evolution*, *3*(2), 268–280.

Pruessner, J. C., Kirschbaum, C., Meinlschmid, G., & Hellhammer, D. H. (2003). Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change. *Psychoneuroendocrinology*, *28*(7), 916–931.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reed, J. M., Harris, D. R., & Romero, L. M. (2019). Profile repeatability: A new method for evaluating repeatability of individual hormone response profiles. *General and Comparative Endocrinology*, *270*, 1–9.

Sapolsky, R. M., Romero, L. M., & Munck, A. U. (2000). How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions. *Endocrine Reviews*, *21*(1), 55–89.

Schoenle, L. A., Zimmer, C., Miller, E. T., & Vitousek, M. N. (2020). Does variation in glucocorticoid concentrations predict fitness? A phylogenetic meta-analysis. *General and Comparative Endocrinology*, 113611.

Small, T. W., Bebus, S. E., Bridge, E. S., Elderbrock, E. K., Ferguson, S. M., Jones, B. C., & Schoech, S. J. (2017). Stress-responsiveness influences baseline glucocorticoid levels: Revisiting the under 3 min sampling rule. *General and Comparative Endocrinology*, *247*, 152–165.

Stinchcombe, J. R., Kirkpatrick, M., Group, F. T. W.others. (2012). Genetics and evolution of function-valued traits: Understanding environmentally responsive phenotypes. *Trends in Ecology & Evolution*, *27*(11), 637–647.

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *8*(11), 1639–1644.

Taborsky, B., English, S., Fawcett, T. W., Kuijper, B., Leimar, O., McNamara, J. M., Ruuskanen, S., & Sandi, C. (2020). Towards an evolutionary theory of stress responses. *Trends in Ecology & Evolution*.

Taff, C. C., Schoenle, L. A., & Vitousek, M. N. (2018). The repeatability of glucocorticoids: A review and meta-analysis. *General and Comparative Endocrinology*, *260*, 136–145.

Taff, C. C., & Vitousek, M. N. (2016). Endocrine flexibility: Optimizing phenotypes in a dynamic world? *Trends in Ecology & Evolution*, *31*(6), 476–488.

Taff, C. C., Zimmer, C., & Vitousek, M. N. (2019). Achromatic plumage brightness predicts stress resilience and social interactions in tree swallows (tachycineta bicolor). *Behavioral Ecology*, *30*(3), 733–745.

Vitousek, M. N., Johnson, M. A., Downs, C. J., Miller, E. T., Martin, L. B., Francis, C. D., Donald, J. W., Fuxjager, M. J., Goymann, W., Hau, M.others. (2019). Macroevolutionary patterning in glucocorticoids suggests different selective pressures shape baseline and stress-induced levels. *The American Naturalist*, *193*(6), 866–880.

Vitousek, M. N., Taff, C. C., Hallinger, K. K., Zimmer, C., & Winkler, D. W. (2018). Hormones and fitness: Evidence for trade-offs in glucocorticoid regulation across contexts. *Frontiers in Ecology and Evolution*, *6*, 42.

Wada, H., & Sewall, K. B. (2014). Introduction to the symposium—uniting evolutionary and physiological approaches to understanding phenotypic plasticity. *American Zoologist*, *54*(5), 774–782.

Westneat, D. F., Wright, J., & Dingemanse, N. J. (2015). The biology hidden inside residual within-individual phenotypic variation. *Biological Reviews*, *90*(3), 729–743.

Wingfield, J. C., Maney, D. L., Breuner, C. W., Jacobs, J. D., Lynn, S., Ramenofsky, M., & Richardson, R. D. (1998). Ecological bases of hormone—behavior interactions: The “emergency life history stage.” *American Zoologist*, *38*(1), 191–206.

Wingfield, J. C., Vleck, C. M., & Moore, M. C. (1992). Seasonal changes of the adrenocortical response to stress in birds of the sonoran desert. *Journal of Experimental Zoology*, *264*(4), 419–428.