A Reaction Norm Perspective on Reproducibility

Bernhard Voelkl & Hanno Würbel

January 3, 2019

Abstract

Reproducibility in biomedical research, and more specifically in preclinical animal research, has been seriously questioned. Several cases of spectacular failures to replicate findings published in the primary scientific literature have led to a perceived reproducibility crisis. Diverse threats to reproducibility have been proposed, including lack of scientific rigour, low statistical power, publication bias, analytical flexibility and fraud. An important aspect that is generally overlooked is the lack of external validity caused by rigorous standardization of both the animals and the environment. Here, we argue that a reaction norm approach to phenotypic variation, acknowledging gene-by-environment interactions, can help us seeing reproducibility of animal experiments in a new light. We illustrate how dominating environmental effects can affect inference and effect size estimates of studies and how elimination of dominant factors through standardization affects the nature of the expected phenotype variation. We do this by introducing a construct that we dubbed the reaction norm of small effects. Finally, we discuss the consequences of a reaction norm of small effects for statistical analysis, specifically for random effect latent variable models and the random lab model.

1 Introduction

Since the mid-17th century reproducibility, i.e. the ability to reproduce an experimental outcome by an independent study is a fundamental corner stone of the scientific method which distinguishes scientific evidence from mere anecdote. In modern research, however, such independent replication has been replaced by principles of experimental design which—in principle—should render replication by independent studies redundant. In the simplest form, the effect of a predictor (independent variable) on an outcome (dependent variable) is measured in a sample of independent replicate units (individuals). Scientific evidence generated in this way is arguably reproducible if the experimental units (i.e. individuals) are true random samples of the overall target population. Despite the general wisdom, that true random samples are practically impossible to achieve when the target population is e.g. a biological species, the potential consequences of non-independence on the reproducibility of results are usually ignored. This is mirrored by the fact that no independent replication studies

are generally required by funders for accepting grant proposals or by editors before accepting manuscripts for publication.
Over the last 10-15 years, however, reproducibility in biomedical research, and more specifically in preclinical animal research, has been seriously questioned [1]. Several cases of spectacular failures to replicate findings published in the primary scientific literature have led to a perceived reproducibility crisis [2, 3]. In 2011 researchers from the company Bayer reported that out of 67 in-house replication studies of published research in the areas of oncology, women's health and cardiovascular diseases only 14 (21 percent) could fully replicate the original findings[4]. Similarly researchers of the company Amgen have replicated 53 original research studies deemed 'landmark' studies in haemathology or oncology, recovering the original findings only in 6 cases (11 percent)[5]. These reports and a surge of meta-analyses confirming low replication rates (e.g. [6, 7, 8]) lead to a heated debate within as well as outside the scientific community about the usefulness of animal models for bio-medical research [3, 2, 9, 10, 11].

Several potential causes for poor reproducibility have been proposed, including lack of scientific rigour, low statistical power, publication bias, analytical flexibility, and perverse incentives in research—leading in some cases to outright fraud [10, 2, 3]. While all of these aspects might contribute to replication failure, we will here focus on another aspect that is all too often ignored: biological variation. Biological variation is the sum of genetic variation, environmentally induced variation and variation due to the interaction between environment and genotype (G $\times$ E interaction). As the response of an animal to an experimental treatment (e.g. a drug) depends on the phenotypic state of the animal, the response, too, is a product of the genotype and the environmental conditions. Despite attempts to standardize animal facilities, laboratories always differ in many environmental factors that affect the animals' phenotype (e.g. noise, odours, microbiota, or personnel [12, 13, 14, 15, 13, 16]. In a landmark study Crabbe and colleagues[12] investigated the confounding effects of the laboratory environment and G $\times$ E interactions on behavioural strain differences in mice. Despite rigorous standardization of housing conditions and study protocols across three laboratories, systematic differences were found between laboratories, as well as significant interactions between genotype and laboratory. Even temporal variation within a single laboratory can lead to relevant effects, as demonstrated in a recent study where researchers found considerable phenotypic variation between different batches of knockout mice tested successively in the same laboratory [17].

The reaction norm is a concept helping to explain the observation that individuals of the same genotype will produce different phenotypes if they experience different environmental conditions [18]. The reaction norm is the result of a complex environmental cue response system, which buffers the functioning of the organism against environmental and genetic perturbations [19, 20]. The consequence of such a regulatory system is that environmental influences can play an important part in shaping the phenotype. Environmental influences do

not only play a role at the time of assessment of the phenotype but throughout the ontogeny of the organism [21]. A reaction norm perspective on phenotypic traits unifies two concepts which have often been treated as opposing mechanisms: phenotype diversification due to environmental variation (plasticity) and the limitation of phenotypic variation by mechanisms that buffer development against genetic and environmental variation (canalization). Both plasticity and canalization have been considered as adaptive traits evolved as a consequence of environmental variation, though following Woltereck's [18] arguments, it is the reaction norm itself that one should consider as the evolved trait [22]. Its adaptive value is, however, limited to a certain range of environmental fluctuation: environmental situations that lie far outside the range of environments a species experienced over its evolutionary past, can overtax the organism's ability to appropriately respond to the situation and lead to maladaptive or pathological responses. With respect to reproducibility it must be emphasised that 'phenotype' is not restricted to visible differences between individuals but does equally refer to differences in physiological or behavioural responses to any sort of stimulation or treatment.

We have recently argued that a failure to recognize the implications of reaction norms might seriously compromise reproducibility in bioscience—specifically in in-vivo research [23, 24]. Laboratory experiments that are conducted with inbred animals under highly standardized conditions are testing only a very narrow range of one specific reaction norm. Independent replicate studies that fail to reproduce the original findings might not necessarily indicate that the original study was poorly done or reported, but rather that the replicate study was probing a different region of the norm of reaction. Therefore, the attampt to improve reproducibility through rigorous standardization of both genotype and environment has been referred to as "standardization fallacy" [27]. Here we will explore this proposition in more detail by providing a formal treatment of the norm of reaction, consider the case of a single dominating environmental factor, discuss special cases, and introduce a concept that we termed "the reaction norm of small effects". In practical terms this will lead us to emphasise the importance of including the laboratory environment as a factor in multi-laboratory studies and meta-analyses or to consider introducing a correction factor in the statistical model to account for predicted lab-specific variation.

2 Conceptualizing The Norm of Reaction

The norm of reaction can be conceptualized as a function mapping an environmental parameter to an expected value of a phenotypic trait. If we denote the environmental parameter as X and the phenotypic trait of the organism as Y, then the norm of reaction $h(\cdot)$ gives the expected value for Y given the environmental state x as $E(y|x) = h(x)$. In many cases the phenotypic trait will be a continuous valued trait. In this case we can describe the distribution of expected values for the trait by a probability density function (PDF) $f(x)$.

Integrating the PDF gives the cumulative density function (CDF), $f_c(x)$. The environmental parameter is assumed to be a characteristic that can be measured on a continuous scale. Environments differ in the environmental parameter and the probability of finding the environment in a specific state regarding this parameter can be given by a probability density function $g(x)$. Again, integrating $g(x)$ gives the respective CDF $g_c(x)$ [1]. Hence, with the help of the reaction norm, we can describe the relationship between the expected trait value and the distribution of the environmental states with the composite function

$$f_c(x) = (g_c \circ h)(x) = g_c(h(x)). \tag{1}$$

The reaction norm $h(\cdot)$ is usually an unknown biological property that can be found if data have been collected that allow estimating both $f_c(\cdot)$, the distribution function for expected values, and $g_c(\cdot)$, the distribution function for the environmental states, as

$$h(x) = (g_c^{-1} \circ f_c)(x). \tag{2}$$

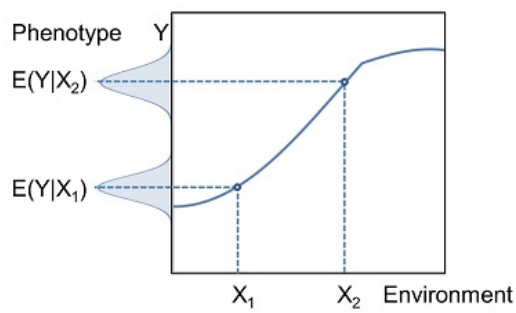

Figure 1: The norm of reaction allows describing the relationship between the expected value of a phenotypic trait ($E(Y)$) and an environmental parameter (X) for a specific genotype. The observed values of the phenotypic trait (indicated by the Gaussian bell curves) will vary due to both biological variation induced by variation in other environmental parameters and measurement error.

2.1 Dominating Factors

Originally Woltereck [18] referred to the relationship between a specific environmental variable and the phenotype as *Phänotypenkurve* (phenotype curve), while he used the term *Reaktionsnorm* (reaction norm) for specifying the collective influence of all environmental variables. However, later Woltereck widened the use of the term reaction norm to include also small subsets of phenotype

[1] We will henceforth use the CDF representation of the distributions because the PDFs of the relevant distributions are not invertible—a property that will be required in the next step. However, the PDF can always be regained as the derivative of the CDF.

curves or even phenotype curves of a single environmental variable. Today the term norm of reaction is usually used to describe the relationship between a single environmental parameter on the expected phenotype of the organism [25, 26]. In evolutionary ecology reaction norms are often the target of the study. Reaction norms are studied experimentally by systematically varying one environmental parameter while all other environmental parameters are kept constant. Usually researchers are focusing on dominating environmental parameters—i.e. parameters that contribute much more to the total environmentally induced trait variance than most other parameters. If one wants to describe the combined effect of two or more environmental parameters on the phenotype, the norm of reaction takes on the form of a plane or a hyperplane. Conceptually, there is no bound for the number of dimensions included, though limits of human imagination sets constrains as the heuristic value of the model quickly decreases with increasing dimensionality. Furthermore, collecting empirical data becomes very cumbersome when combinations of several parameters need to be varied systematically. For these two reasons defining high-dimensional norms of reaction is an approach rarely taken or advised.

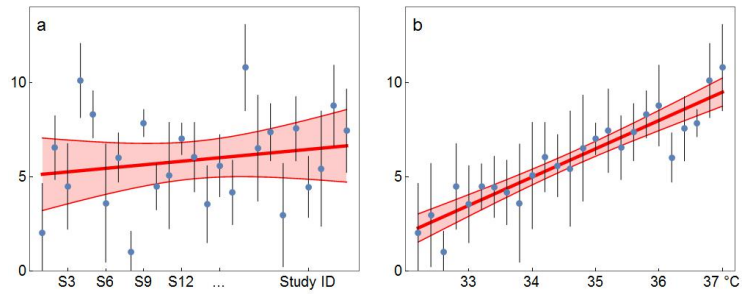

Figure 2: The effect of dominating factors on effect size estimates and reproducibility. Panel (a) shows the hypothetical results of 25 studies, where between-study variability is relatively large in comparison to within study variability and several studies would not cover the CI for the combined effect size estimate, suggesting "replication failure". In panel (b), however, studies are sorted by an environmental gradient (ambient temperature) on the y-axis, suggesting that this environmental factor has a linear influence on the effect size of the experimental treatment. In this case, an inclusion of this factor, would allow giving predicted values with respect to the environmental variable and all those studies capturing the predicted value for the respective ambient temperature should be considered successful replications. In the case of a specific environmental factor that was reliably measured and reported for all studies, such a regression approach would, indeed, be the best option for both estimating the conditional effect size and estimating replication success. If, however, the nature of this environmental variable is not known, we cannot proceed that way.

In most cases of biomedical research, environmentally induced trait variation is not of interest and considered as unwanted noise. The predominant

approach taken to deal with environmentally induced variation is to identify potential dominating environmental parameters and keep them constant (standardization). In those cases, where a dominating factor can be identified but not controlled, it might be recorded and added to the analysis as co-variate or nuisance factor. The very idea of environmental standardization is, thus, to reduce environmentally induced trait variation by reducing variation of all those environmental factors that are known to—or are suspected to—cause trait variation. The list of factors standardized in most pre-clinical studies with rodent model organisms includes (but is not limited to) cage size, cage content (nesting material, shelter, enrichment devices), housing temperature, humidity, light regime, stocking density, food and water supply, handling techniques and cage maintenance routines. In fact, even many more environmental factors are standardized, though some of them seem to be so self-evident or trivial that they are hardly ever mentioned and easily overlooked (e.g. all laboratory environments are free of catastrophic events like hailstorms or feline predators). Thus, rigorous standardization is presumed to eliminate most or all dominating factors and, hence, lead to a substantial reduction of environmental variation and arguably also to a reduction in environmentally induced trait variation.

2.2 Reaction Norm of Small Effects

If all environmental factors with dominating contributions to trait variation have been "neutralized" in a big sweep—together with many other factors that had no or only minor effects—, one might believe that the remaining environmentally induced variation is of little interest. This, however, might not necessarily be the case, because in addition to standardizing environmental conditions, the genetic background of the laboratory animals is also highly standardized when experiments are conducted with inbred mouse strains. Mice used in a single study will be delivered from the same breeding facility and stem from the same breeding line. That is, individual genetic variation, too, has been largely reduced with the result that environmentally induced variation and G $\times$ E interactions might still make up most of the total biological variation of the organism [27]. Environmental effects should, therefore, still be taken into account. Yet, the nature of the combined environmental influences has changed. Originally, we were confronted with the situation of many environmental parameters, each having a small effect on the trait variation and one or a small number of dominating parameters, contributing much more to the trait variability. Under such circumstances, those dominating parameters are best accounted for by adding them as co-variates to the statistical model. After standardization, however, we should be left only with a large number of factors, each having a small effect on the total variance. This situation requires a different treatment. For now, we assume that those factors are additive and independent of each other. Recalling the central limit theorem, we can expect that under those assumptions the limiting distribution for the environmental states with respect to their effect on the trait value can be described by a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma)$.

Before we continue with exploring the consequences of this conjecture for the reaction norm, we should briefly pause and contemplate to what extent the assumptions made above are justified. The central limit theorem has been formulated repeatedly in slightly different forms with different degrees of generality [28, 29, 30]. The essence of these formulations is that if a random variable X is the sum of a large number of independent and identically distributed sequences of random vectors, then X will be approximately normally distributed. Ronald Fisher used it in his seminal 1918 paper [31] in order to reconcile the observations of biometricians, who found many continuously varying traits, with Mendelian genetics. His explanation, that the cumulative effect of the presence or absence of many different alleles—each having a small effect of the trait and each being independently inherited—will lead to a Normal distribution of the trait value, became a corner stone of quantitative genetics, where it referred to as "Fisher's infinitesimal model" [32]. The application of the central limit theorem to the distribution of biological traits has been questioned [33, 34, 35], mainly on the grounds that effects of genes can vary substantially, with one or a few genes dominating. This problem is not an issue when we investigate the case, where the dominating factors have already been accounted for. The question of non-independence can be resolved by accommodating interaction terms and treating them as factors in their own right. Yet, we have to consider a third requirement for the central limit theorem in its original formulation: that all random variables should have the same distribution. At this point the analogy between genetic and environmental factors ends. In the case of genes, the atomic random variables are all Bernoulli variables: an allele can either be present or absent. In the case of environmental parameters, this is not the case and the generating functions might be of different nature. However, refined versions of the central limit theorem by Lyapunov [36], Lindeberg [37] and Feller [38] relax this assumption to a certain extent—though as a minimal requirement the variances of all contributing variables must be finite. In a thorough review Frank [39] has pointed out that most patterns observed in nature can be traced back to a small number of generating processes, all of which leading to distributions with finite variance. Thus, invoking the Lindeberg-Feller limiting conditions we argue that the assumptions required to employ the central limit theorem for small environmental effects on expected trait values are sufficiently met.

Finding the reaction norm $h(\cdot)$ as outlined in equ. 2 becomes now a question of combining empirical observations, allowing to infer the distribution of expected trait values $E(y)$, with the assumed distribution for X of combined effects of many environmental factors, each having a small effect. In principle, the parametrisation of the distribution for X is arbitrary and one could settle for any values, like a mean of 0 and standard deviation of 1. However, if the response variable itself is not centred and standardised, this can lead to rather odd looking reaction norms easily betraying the eye of the observer. Graphing such reaction norms would do a poor job in aiding our understanding of the process. We therefore suggest to first re-scale $g(x)$ without loss of generality,

employing an entropy minimizing approach for finding those parameter values μ and σ that give the closest fit of $g(x)$ to $f(x)$. This can be done using the Kullback-Leibler divergence,

$$D_{KL}(f|g) = \int_{-\infty}^{\infty} f\,(x) \log \frac{f(x)}{g(x)}\,\mathrm{d}x, \tag{3}$$

where we have to find numeric approximations so that $D_{KL}(f|g) \longrightarrow min$. After having found approximate values for μ and σ we can derive $h(\cdot)$ using equ. 2. There is no closed form expression for the inverse CDF of a Normal distribution, but numerical approximations can easily be calculated (e.g. [40]).

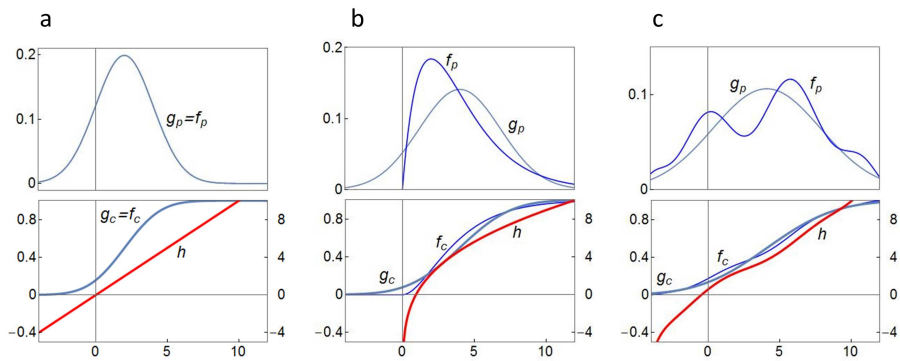

Figure 3: Graphical representation of reaction norms h of small effects for the case that the distribution of the expected trait follows (a) a normal distribution with $m = 2$ and $s = 2$, (b) a gamma distribution with $k = 2$ and $\Theta = 2$ and (c) a bimodal distribution with a PDF approximated by fitting a higher order polynomial; f_p denotes the probability density function for the expected trait values and g_p denotes the probability density function for the distribution of combined environmental effects, which is always assumed be normal with μ and σ being chosen to fit f_p most closely in terms of minimizing entropy. g_c and f_c denote the respective cumulative density functions. Function values for h are given on the right hand side of the panels.

In figure 3 we have plotted $h(\cdot)$ for three different distributions of expected trait values. Figure 3a depicts the case where the expected trait value is normally distributed: in this case $h(\cdot)$ becomes a straight line. The slope of the line indicates how strong the environment affects the trait value. A steep slope indicates a strong environmental effect while a slope of $s = 0$ would indicate that the environment has no effect on the trait value. Figure 3b depicts another potentially interesting scenario, where the expected trait value shows a skewed distribution. Finally, Figure 3c depicts the case where a higher order polynomial was fitted to a frequency distribution of expected trait values. While this is generally not advised (there is a danger of over-fitting and the heuristic value of higher order polynomials is usually rather low), we have used this here only in

order to demonstrate that it is—in principle—possible to find a reaction norm for any sort of trait value distribution.

2.3 Acceptance Region

Having found a method for relating environmental variation to expected variation of trait values Y, we might ask, whether this can help us in defining an acceptance region $[L_{y(lower)}, L_{y(upper)}]$, in which the effect size estimate of a replicate study has to fall, in order to be considered a 'successful' replication. Traditionally, the discussion how to find this region has focused almost exclusively on the domain of Y by partitioning the observed variation in the trait value in variance attributed to laboratory (i.e. environmental) variation and variance attributed to individual variation and measurement error. Here, we suggest a conceptually different approach: instead of defining the acceptance region based on observed trait variation, we want to define the acceptance region based on the range of the environments—more specifically the strength of the environmental stimulus—that should be deemed relevant. That is, we first define a region $[x_{min}, x_{max}]$ on the domain of X, and by projecting the boundaries of this region for the environmental variable onto the expected trait values with

$$L_{y(lower)} = h(x_{min}) \quad \text{and} \quad L_{y(upper)} = h(x_{max}), \tag{4}$$

we can arrive at an acceptance region for estimated trait values, where we would consider all studies with effect sizes falling within this range as successful replicates. In the presence of a single dominating factor the selection of the range $[x_{min}, x_{max}]$ can be based on biological relevance or practical considerations. We can restrict the environmental parameter range to values as the animals might encounter under natural conditions, to values where we have reasons to believe that we are not overtaxing the animal's adaptive capacity, or to values that might be relevant for practical applications or translation to clinical studies. For reaction norms of small effects one might define the acceptance region based on the assumed normal distribution of combined effects

$$[\mu - q \times \sigma < x < \mu + q \times \sigma]. \tag{5}$$

Projecting the boundaries of this region for the environmental variable onto the expected trait values we get

$$L_{y(lower)} = h(\mu - q \times \sigma) \quad \text{and} \quad L_{y(upper)} = h(\mu + q \times \sigma). \tag{6}$$

If a replicate study with sample size n_r delivers an estimate for Y of $\bar{y}_r$ with σ_r, then this study can be considered a successful replication if

$$\bar{y}_r + \frac{z \times \sigma_r}{\sqrt{n_r}} > L_{y(lower)} \quad \wedge \quad \bar{y}_r - \frac{z \times \sigma_r}{\sqrt{n_r}} < L_{y(upper)}. \tag{7}$$

At this point we have to issue a caveat: The visual similarity of equ 5 with a confidence interval should not entice the practitioner setting $q = 1.96$, as this would basically mean that by definition we declare 95 percent of replicate studies being successful replicates. This is clearly not desirable if we wish to gauge reproducibility. Instead, q must be based on the range of interest as mentioned above. Admittedly this is easier said (or written) than done. At the moment we cannot give any well-grounded guidance how to choose q, but we see this question as a potentially interesting issue for further investigation. In figure 4 we present one illustrative example, where q was set to 1.28, with the effect that the acceptance region covers 80 % of all possible study environments. In this case the reaction norm (the same as in figure 3b) has two effects: first, the percentage of expected trait values falling into the acceptance region is considerably higher (89 % in this case) and second, the rejection regions (blue shaded areas) are markedly asymmetric with the area under the curve of the left-hand tail being much smaller than the one for the right-hand tail.

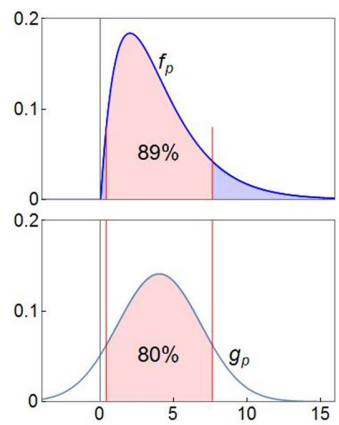

Figure 4: Example for an acceptance region (red shaded area) based on the reaction norm approach. Assuming that a non-linear reaction norm as in figure 3b leads to a skewed distribution for expected trait values (f_p, top panel), an acceptance region defined on the prevalence of small environmental effects (assumed to follow a normal distribution, g_p, bottom panel) will lead to an asymmetric and broadened acceptance region for expected trait values. Red lines indicate chosen threshold values for the acceptance region, assuming that 10 % of the most extreme environments on each side of the distribution can be considered as not relevant.

2.4 Latent Variable Models

From the previous treatment we learn two important things. First, as soon as the slope for the reaction norm of small effects is not flat, the environment affects the expected trait value and should be incorporated in any explanatory model

as latent random variable. In analyses of multi-lab studies and in meta-analyses this is done by treating the laboratory, the study site, or the study as random factor of a mixed effect model. Indeed, over the last decades several authors have emphasised and diligently advocated the use of mixed effect models for multi-centre studies [41, 42] and meta analyses [43]. Their efforts have not been in vain and today mixed effect models can be considered the standard approach to dealing with lab-to-lab or clinic-to-clinic variation. However, while those recommendations for the use of mixed effect models were based on statistical arguments (non-independence and the observation that adding a random factor for lab or clinic can reduce the unexplained error term), we arrived at the same suggestion from—what we would call—first principles of biology: the norm of reaction as a cogent product of stabilizing selection. Second, and perhaps more importantly, we can make the observation that the expected trait value should be normally distributed if and only if the norm of reaction is a straight line with a non-zero slope. This might, indeed, be the case from time to time, though we have to note that a straight line is just a special case of all potential reaction norms and that in most cases the reaction norm should not lead to a normal distribution for expected trait values. For example, figure 5 shows that a concave reaction norm, like in this case a negative exponential function, will lead to a skewed distribution with a fat tail.

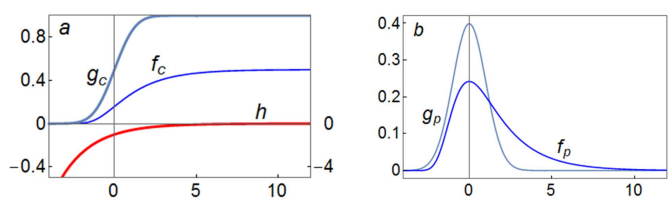

Figure 5: Assuming a Gaussian distribution for the combined small effects (red), a negative exponential reaction norm $h(x) = -e^{-\frac{x}{2}}$ results in a skewed, heavy-tailed distribution (blue) for the expected value.

The generic form of a linear mixed effect model can be expressed as

$$E(y) = g^{-1}(X\beta + Z\gamma + \delta), \tag{8}$$

where $E(y)$ is the expected trait value, $g^{-1}(\cdot)$ is the inverse link function, the term $X\beta + Z\gamma$ is the linear predictor, δ is the model offset vector, Z is the design matrix for the random effects, and γ is a column vector of the random effects. For statistical modelling it is usually assumed that γ_k is a random variable $\gamma_k \sim \mathcal{N}(0, G)$, where G is the variance-covariance matrix of the random effect. This assumption (implicit in basically all analyses based on linear mixed models) is mainly made out of convenience [44, 45], but—strictly—it is only justified in the case that the reaction norm is neatly linear with a slope $s \neq 0$.

As we have argued before, this might be the exception rather the rule.

Does this mean that almost all multi-centre studies or meta-analyses implicitly assuming a normally distributed latent variable are wrong? From a conceptual viewpoint, this might indeed be the case. There are however two reasons why this might not matter too much for practical purposes. First, for most statistical models it is sufficient that normality is only approximately met as the algorithms might be rather robust against moderate deviations from normality [46, 47, 48, 49]. That is, if the reaction norm does not lead to a heavily skewed or distorted distribution of the latent variable, then the effect on the model outcome might be negligible. Second, even though it is unlikely that reaction norms are strictly linear over the entire domain, it seems plausible that they are approximately linear over a certain range—which might coincide with the range of interest. Sigmoid functions can, for example, result in an almost linear relationship for intermediate domain values. Whether sigmoid functions are common for reaction norms is currently unknown, because examples, where researchers have collected data for a large number of different values for an environmental parameter, are still scarce. This is true for animal ecology but even more so for lab-based animal science. As a consequence we can, right now, only speculate about the shape of reaction norms for specific traits. Yet, if speculating we do, sigmoid shapes do seem rather plausible, as they might describe situations where effects on the trait value are bounded and where the increase or decrease in the trait value becomes smaller the closer one comes to the limiting boundaries. Thus, for these two reasons, mixed effect models with the assumption of a normally distributed latent variable might, after all, be sufficiently suitable models for most cases. If one has reason to believe that the assumption is substantially violated, then a non-parametric modelling approach based on mixture-models [44, 50] or Markov chain Monte Carlo methods [51] might offer suitable alternatives.

2.5 Discussion

We started off with the observation that the phenotype of an organism is always a product of its genotype and the environmental circumstances under which it developed. Thus, a phenotypic trait should not be considered as a fixed entity but as a conditional property of the organism. Experimenters have long identified environmental clustering—be it as sites, laboratories, batches, racks, cages—as potential sources for covariation. The seemingly logical solution to this problem is, to add shared environment as a random effect in the statistical model. For example, if a large biomedical intervention study is carried out at several laboratories, then a joint analysis would include the identity of the laboratory as a random factor in the analysis. In single-lab studies batch or cage are often added as random factors. These random factors are by default assumed to be normally distributed random variables. As several authors have noted (e.g. [50, 44, 45, 46]) this assumption is made for computational convenience and not because of compelling empirical evidence. From a conceptual

viewpoint it is clearly not justified: here we should assume that the environmental influence is a sum of several different underlying processes. For one, results from different laboratories might differ because of between-lab differences in how measurements are made, different levels of accuracy and precision, or sampling variation. Assuming a normal distribution for these kinds of 'measurement error' or noise might, indeed, be appropriate. However, as we have pointed out, the forming effect of the environment will also lead to increased variation of phenotypes at the grouping level. In this case, we should not necessarily assume a normal distribution. Thus, the conclusions one should draw from the conceptual reaction norm approach differs from the statistically justified mixed effect approach in one important aspect: While the latter assumes usually that the latent variable is normally distributed, the reaction norm approach suggests that a normally distributed latent variable should be the exception rather than the rule—occurring only in the case of strictly linear reaction norms. That is, we should consider adding two environmental factors to our statistical model: one as a Gaussian random variable capturing between-site random 'noise', and one acknowledging the variation in the expected trait value as a result of the reaction norm. Importantly, while the former can be considered unwanted 'noise' one wants to control for, the latter is usually not 'noise' but biologically relevant information that we do not want to spirit away. If well-nourished organisms (raised in lab environment A) respond stronger to a specific treatment that organisms that grew up under severe food restriction (in lab environment B), then this is biologically relevant information and we do not want to put this variation in the same bucket as variation resulting from differently calibrated instruments in different labs. In our opinion, this conceptual difference is even more important than differences concerning the form of the limiting distributions.

Next, we have noted that reaction norms come in two flavours: dominating factors and factors of small effect. Given the usually continuous nature of environmental effects on trait values, this is a rather arbitrary distinction that would defy any attempt of operationalization. Dominating factors are environmental factors that contribute much more to the overall trait variation, than other environmental factors, but for practical purposes we can simply define dominating factors as factors where we can see clear effects on the trait variation given realistic (reasonably small) sample sizes. We assume that if such effects exist, vigilant experimenters will either control the environmental parameter (keeping it constant) or incorporate it in the analysis by systematically varying it and adding it to the model. Thus, our proposition that experimenters are vigilant allowed us keeping the discussion of dominating effects short and, for the larger part of our study, we focused on the reaction norm of small effects. Here, we argued that we can expect a large number of environmental parameters having a small effect on the expected phenotype value. Employing a relaxed version of the central limit theorem, we suggested summarizing the effects of all those parameters in a single one-dimensional reaction norm. The question arises, whether those small environmental effects can have an effect on the reproducibility of a study result. We argue that this can, indeed, be the case for two reasons. First, even if the

effect of a single environmental parameter might be rather small, the combined effects of many such parameters can become substantial. Second, what we see in biomedical research is a tendency for standardizing many aspects of experimental studies. Standardizing instruments and measurement protocols means reducing measurement error. Standardizing housing conditions and testing conditions means eliminating most dominating environmental factors and, hence, reducing the overall variation. At the same time, standardizing the genotype by working with highly inbred lines means that also the genetic variation is largely reduced—leading again to a reduction of variance of the phenotype. Thus, while the overall phenotypic variation is reduced through standardization, the relative proportion of the phenotypic variation contributed by the remaining environmental factors will consequently increase [27]. As the reduction of measurement error and genetic variation results in a larger proportion of phenotype variation that can be attributed to the reaction norm of small effects, we have to consider what consequences this has for the distribution of the expected trait value.

Recently, an extensive meta-analysis of 5580 medical studies of over 300 different quantities showed that outliers were much more frequent than expected assuming normally distributed error terms [52]. Observed distributions were better described by heavy-tailed Student's t-distributions or Cauchy distributions than by Normal distributions. Various explanations for the frequent deviations from normality have been put forward in the literature including various forms of bias [11, 9] and underestimation of the variance [53, 54]. Yet, Bailey [52] suggests a different explanation, arguing that most measurements are in fact the result of complex systems of interacting components. In complex systems power-law behaviour of distributions is rather common [55], with the cumulative distributions of observed effects (x) declining at $1/x^{\alpha}$. The observation that the parameter v of a t-distribution (which equals the cumulative tail exponent α) has similar values for medical studies as reported for well researched complex systems like software or power grids [56, 57], is seen both as corroboration that the Student's t-distribution (with $v \sim 2$—3) might be a better fit for error terms than a Gaussian Normal distribution, and as indication that outcomes of medical tests behave as the outcomes of complex systems [52]. As any biologist will support the notion that traits of organisms are the results of complex systems, it is very tempting to subscribe to this explanation for the over-abundance of highly significant findings in biomedical research. Yet, we want to caution against accepting this explanation as the sole explanation for the high proportion of very low p-values. While we do not want to challenge Bailey's explanation in principle, we argue that non-linear reaction norms can also lead to heavy tails, even if the underlying process (environmental variation) has a Gaussian limiting distributions.

A statistical approach incorporating the reaction norm into estimates of individual studies requires that the specific paradigm is well researched and a substantial body of studies, where the same parameters were investigated in individuals raised under diverse environmental conditions, already exists. Such

rich treasure troves of empirical data are very rare, making it usually impossible to give plausible estimates for the reaction norm. As a consequence, Kafkafi and colleagues [58] have suggested an alternative approach, dubbed the random lab model (RLM), ascribing a random effect to each laboratory. This model adds 'noise' for the presumed variation contributed by the $G \times E$ interaction term to the individual noise, generating an 'adjusted yardstick' [58] for inference and parameter estimates. The RLM is, thus, raising the benchmark for finding significant results by trading statistical power for increased reproducibility through wider confidence intervals of the effect size estimates [24]. The effect of this $G \times E$ adjustment is technically achieved by adding a penalizing $G \times E$ term to the variance. The standard error for the effect size estimate of a simple contrast of two groups (e.g. 'test' and 'control') can, then, be calculated as

$$SE = \sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2}) + 2s^2_{G\times E}}, \tag{9}$$

where s^2 is the observed variance and $2s^2_{G\times E}$ is the added '$G \times E$ noise' [58]. The latter term cannot be estimated from data from a single experiment, but it is suggested—or hoped for—that large data bases or meta analyses will allow giving rough approximate values for specific fields of research and specific types of interventions.

At this point we can note that the effect of adding this penalizing $G\times E$ term is effectively equivalent to a reaction norm approach assuming a linear reaction norm of small effects with positive slope. In the previous section we conceded that a stumbling block, impeding the practical application of our approach of defining an acceptance region, is the question of how to arrive at sensible values for q. We believe that this question is similar to the question of how to find reasonable $G \times E$ terms for the RLM. A potential solution for the latter is to derive this quantity from large meta-analyses for individual research fields or individual experimental paradigms. If such values can be found based on a large-enough sample of studies, it might be worth exploring what we can learn from taking the penalising term from the RLM in order to deduce q in equ. 5 through reverse engineering. We do not want to suggest that this should become a practicable way for finding q, because this would basically pervert out own argument—that the acceptance region should be based on information about the domain X of the environment—, but it might be still a useful exercise for better understanding the link between environmental variation and trait variation.

3 Conclusion

When studying living organisms we are faced with inherent biological variation which is distinct from random noise or measurement error and which is fundamental to the correct interpretation of experimental results. Fully acknowledging this requires adopting a reaction norm perspective on physiological and

behavioural responses. This will lead to a re-thinking of parameter estimation and inference, it will let us see reproducibility in a new light and it can even help gaining new insights into adaptive responses and gene-by-environment interactions. Here, we have tried to dissect its implications for the reproducibility debate and, more generally, what it means for the interpretation of experimental results in biomedical research.

References

[1] J. D. Bailoo, T. S. Reichlin, and H. Würbel, "Refinement of experimental design and conduct in laboratory animal research," *ILAR Journal*, vol. 55, pp. 383–391, dec 2014.

[2] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe, "The economics of reproducibility in preclinical research," *PLOS Biology*, vol. 13, p. e1002165, jun 2015.

[3] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, vol. 2, p. e124, aug 2005.

[4] F. Prinz, T. Schlange, and K. Asadullah, "Believe it or not: how much can we rely on published data on potential drug targets?," *Nature Reviews Drug Discovery*, vol. 10, pp. 712–712, sep 2011.

[5] C. G. Begley and L. M. Ellis, "Drug development: Raise standards for preclinical cancer research," *Nature*, vol. 483, no. 7391, p. 531, 2012.

[6] E. S. Sena, H. B. van der Worp, P. M. W. Bath, D. W. Howells, and M. R. Macleod, "Publication bias in reports of animal stroke studies leads to major overstatement of efficacy," *PLoS Biology*, vol. 8, p. e1000344, mar 2010.

[7] E. D. M. Rooke, H. M. Vesterinen, E. S. Sena, K. J. Egan, and M. R. Macleod, "Dopamine agonists in animal models of parkinson's disease: A systematic review and meta-analysis," *Parkinsonism & Related Disorders*, vol. 17, pp. 313–320, jun 2011.

[8] E. Dumas-Mallet, K. S. Button, T. Boraud, F. Gonon, and M. R. Munafò, "Low statistical power in biomedical science: a review of three human research domains," *Royal Society Open Science*, vol. 4, p. 160254, feb 2017.

[9] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis, "A manifesto for reproducible science," *Nature Human Behaviour*, vol. 1, p. 0021, jan 2017.

[10] E. Loken and A. Gelman, "Measurement error and the replication crisis," *Science*, vol. 355, pp. 584–585, feb 2017.

[11] E. Sena, H. B. van der Worp, D. Howells, and M. Macleod, "How can we improve the pre-clinical development of drugs for stroke?," *Trends in Neurosciences*, vol. 30, pp. 433–439, sep 2007.

[12] J. C. Crabbe, "Genetics of mouse behavior: interactions with laboratory environment," *Science*, vol. 284, pp. 1670–1672, jun 1999.

[13] E. J. Chesler, S. G. Wilson, W. R. Lariviere, S. L. Rodriguez-Zas, and J. S. Mogil, "Influences of laboratory environment on behavior," *Nature Neuroscience*, vol. 5, pp. 1101–1102, nov 2002.

[14] D. Wahlsten, P. Metten, T. J. Phillips, S. L. Boehm, S. Burkhart-Kasch, J. Dorow, S. Doerksen, C. Downing, J. Fogarty, K. Rodd-Henricks, R. Hen, C. S. McKinnon, C. M. Merrill, C. Nolte, M. Schalomon, J. P. Schlumbohm, J. R. Sibert, C. D. Wenger, B. C. Dudek, and J. C. Crabbe, "Different data from different labs: Lessons from studies of gene-environment interaction," *Journal of Neurobiology*, vol. 54, pp. 283–311, dec 2002.

[15] H. Würbel, "Behavioral phenotyping enhanced - beyond (environmental) standardization," *Genes, Brain and Behavior*, vol. 1, pp. 3–8, jan 2002.

[16] R. E. Sorge, L. J. Martin, K. A. Isbester, S. G. Sotocinal, S. Rosen, A. H. Tuttle, J. S. Wieskopf, E. L. Acland, A. Dokova, B. Kadoura, P. Leger, J. C. S. Mapplebeck, M. McPhail, A. Delaney, G. Wigerblad, A. P. Schumann, T. Quinn, J. Frasnelli, C. I. Svensson, W. F. Sternberg, and J. S. Mogil, "Olfactory exposure to males, including men, causes stress and related analgesia in rodents," *Nature Methods*, vol. 11, pp. 629–632, apr 2014.

[17] N. A. Karp, A. O. Speak, J. K. White, D. J. Adams, M. H. de Angelis, Y. Hérault, and R. F. Mott, "Impact of temporal variation on design and analysis of mouse knockout phenotyping studies," *PLoS ONE*, vol. 9, p. e111239, oct 2014.

[18] R. Woltereck, "Weitere experimentelle untersuchungen über artveränderung, speziell über das wesen quantitativer artunterschiede bei daphniden," *Verh. D. Tsch. Zool. Ges.*, vol. 1909, pp. 110–172, 1909.

[19] I. I. Schmalhausen, "Factors of evolution: the theory of stabilizing selection.," 1949.

[20] C. H. Waddington, "Canalization of development and the inheritance of acquired characters," *Nature*, vol. 150, pp. 563–565, nov 1942.

[21] C. D. Schlichting and M. Pigliucci, *Phenotypic Evolution: A Reaction Norm Perspective*. Sinauer Associates, 1998.

[22] S. C. Stearns, "The evolutionary significance of phenotypic plasticity," *BioScience*, vol. 39, pp. 436–445, jul 1989.

[23] B. Voelkl and H. Würbel, "Reproducibility crisis: Are we ignoring reaction norms?," *Trends in Pharmacological Sciences*, vol. 37, pp. 509–510, jul 2016.

[24] B. Voelkl, L. Vogt, E. S. Sena, and H. Würbel, "Reproducibility of preclinical animal research improves with heterogeneity of study samples," *PLOS Biology*, vol. 16, p. e2003693, feb 2018.

[25] M. Pigliucci, "Evolution of phenotypic plasticity: where are we going now?," *Trends in Ecology & Evolution*, vol. 20, pp. 481–486, sep 2005.

[26] S. Sarkar, "From the reaktionsnorm to the adaptive norm: the norm of reaction, 1909–1960," *Biology & Philosophy*, vol. 14, pp. 235–252, apr 1999.

[27] H. Würbel, "Behaviour and the standardization fallacy," *Nature Genetics*, vol. 26, pp. 263–263, nov 2000.

[28] F. Galton, "Statistics by intercomparison, with remarks on the law of frequency of error," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 49, pp. 33–46, jan 1875.

[29] A. De Moivre, *The doctrine of chances: or, A method of calculating the probabilities of events in play*, vol. 1. Chelsea Publishing Company, 1756.

[30] G. Pólya, "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem," *Mathematische Zeitschrift*, vol. 8, no. 3-4, pp. 171–181, 1920.

[31] R. A. Fisher, "The correlation between relatives on the supposition of mendelian inheritance.," *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433, 1919.

[32] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: from polygenic to omnigenic," *Cell*, vol. 169, pp. 1177–1186, jun 2017.

[33] H. Cramer, "Mathematical methods of statistics (princeton: Princeton universitypress, 1946).?," *CramérMathematical Methods of Statistics1946*, 1946.

[34] A. Lyon, "Why are normal distributions normal?," *The British Journal for the Philosophy of Science*, vol. 65, pp. 621–649, sep 2013.

[35] J. H. Gillespie, *Population genetics: a concise guide*. JHU Press, 2010.

[36] A. M. Lyapunov, "Une proposition générale du calcul des probabilités," *Comptes rendus hebdomadaires de l'Académie des Sciences de Paris*, vol. 132, no. 814, pp. 173–174, 1901.

[37] J. W. Lindeberg, "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 15, pp. 211–225, 1922.

[38] W. Feller, "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 40, pp. 521–559, 1935.

[39] S. A. Frank, "The common patterns of nature," *Journal of Evolutionary Biology*, vol. 22, pp. 1563–1585, aug 2009.

[40] E. T. Jaynes, *Probability theory: The logic of science*. Cambridge university press, 2003.

[41] A. R. Localio, J. A. Berlin, T. R. T. Have, and S. E. Kimmel, "Adjustments for center in multicenter studies: An overview," *Annals of Internal Medicine*, vol. 135, p. 112, jul 2001.

[42] B. C. Kahan and T. P. Morris, "Assessing potential sources of clustering in individually randomised trials," *BMC Medical Research Methodology*, vol. 13, apr 2013.

[43] P. R. Freeman, L. V. Hedges, and I. Olkin, "Statistical methods for meta-analysis.," *Biometrics*, vol. 42, p. 454, jun 1986.

[44] M. Aitkin, "A general maximum likelihood analysis of variance components in generalized linear models," *Biometrics*, vol. 55, pp. 117–128, mar 1999.

[45] C. E. McCulloch and J. M. Neuhaus, "Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter," *Statistical Science*, vol. 26, pp. 388–402, aug 2011.

[46] C. E. McCulloch and J. M. Neuhaus, "Prediction of random effects in linear and generalized linear models under model misspecification," *Biometrics*, vol. 67, pp. 270–279, may 2010.

[47] C. J. M. Maas and J. J. Hox, "Robustness issues in multilevel regression analysis," *Statistica Neerlandica*, vol. 58, pp. 127–137, may 2004.

[48] L. Grilli and C. Rampichini, "Specification of random effects in multilevel models: a review," *Quality & Quantity*, vol. 49, pp. 967–976, jul 2014.

[49] A. Bell, M. Fairbrother, and K. Jones, "Fixed and random effects models: making an informed choice," *Quality & Quantity*, aug 2018.

[50] J. Einbeck, J. Hinde, and R. Darnell, "A new package for fitting random effect models.," *R news.*, vol. 7, no. 1, pp. 26–30, 2007.

[51] J. D. Hadfield, "MCMC methods for multi-response generalized linear mixed models: TheMCMCglmmRPackage," *Journal of Statistical Software*, vol. 33, no. 2, 2010.

[52] D. C. Bailey, "Not normal: The uncertainties of scientific measurements," *Royal Society Open Science*, vol. 4, p. 160600, jan 2017.

[53] A. I. Shlyakhter, "An improved framework for uncertainty analysis: Accounting for unsuspected errors," *Risk Analysis*, vol. 14, pp. 441–447, aug 1994.

[54] M. Henrion and B. Fischhoff, "Assessing uncertainty in physical constants," *American Journal of Physics*, vol. 54, pp. 791–798, sep 1986.

[55] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, pp. 661–703, nov 2009.

[56] L. Hatton, "Defects, scientific computation and the scientific method," in *IFIP Advances in Information and Communication Technology*, pp. 123–138, Springer Berlin Heidelberg, 2012.

[57] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, "Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization," *Chaos*, vol. 17, p. 026103, jun 2007.

[58] N. Kafkafi, I. Golani, I. Jaljuli, H. Morgan, T. Sarig, H. Wrbel, S. Yaacoby, and Y. Benjamini, "Addressing reproducibility in single-laboratory phenotyping experiments," *Nature Methods*, vol. 14, pp. 462–464, apr 2017.