Journal of Thermal Biology 132 (2025) 104248

Contents lists available at ScienceDirect

Journal of Thermal Biology

journal homepage: www.elsevier.com/locate/jtherbio

Forum Article

On testing the tolerance-plasticity trade-off hypothesis as the change of thermal tolerance across two environments

Mauro Santos [a,b,*], José F. Fontanari [c]

[a] Departament de Genètica i de Microbiologia, Grup de Genòmica, Bioinformàtica i Biologia Evolutiva (GBBE), Universitat Autònoma de Barcelona, Spain
[b] cE3c - Centre for Ecology, Evolution and Environmental Changes & CHANGE - Global Change and Sustainability Institute, Lisboa, Portugal
[c] Instituto de Física de São Carlos, Universidade de São Paulo,São Carlos, 13560-970, São Paulo, Brazil

ARTICLE INFO

Keywords:
Phenotypic plasticity
Thermotolerance
Trade-off
Spurious correlations
Randomization tests
Trend analysis

ABSTRACT

The world is warming rapidly, threatening the extinction of much of the world's biota. Thermal tolerance plasticity has been touted as an important buffer against global warming. The temperature tolerance-plasticity trade-off hypothesis (TOH) posits that ectotherms who have adapted to high temperatures have done so at the expense of having limited plasticity to further improve their heat tolerance. Empirical evidence is mixed and inconsistencies may arise due to statistical artefacts caused by spurious correlations. This lack of consensus is problematic because an accurate evaluation of the TOH is crucial for estimating the buffering capacity of thermal plasticity in ectotherms that already live close to their upper physiological thermal limits. In this study, we demonstrate that the manner in which the statistical bias has been addressed when evaluating the TOH, as measured at the intraspecific level by the change in thermal tolerance across two environments, is erroneous. This is because there has never been a statistically robust prediction for either a trade-off or a lack of trade-off in two-environment experiments, which is surprising given the importance of such predictions in this field. Here, we derive a statistical framework to correctly test the hypothesis that the observed change in thermal tolerance is consistent with TOH predictions. To demonstrate how our approach can alter the conclusions and interpretations of the TOH, we apply it to two existing datasets. We show that the TOH may indeed be valid, despite previous claims to the contrary, highlighting the critical importance of a sound statistical approach to avoid spurious conclusions that can have significant implications for our understanding of climate change responses.

1. Introduction

It has been more than twenty years since Stillman (2003) published a seminal paper showing that acclimation capacity in upper thermal tolerance (CT_{max}) of cardiac function trades-off with basal heat tolerance in four species of porcelain crabs from different thermal habitats. As summarized by Parmesan et al. (2022, p. 225): "Some species have evolved extreme upper thermal limits at the expense of plasticity, reflecting an evolutionary trade-off between these traits. The most heat-tolerant species, such as those from extreme environments, may therefore be at a greater risk of warming because of an inability to physiologically adjust to thermal change (*low confidence*)". Although Parmesan et al. (2022) focused on the interspecific level, several studies have also tested the tolerance-plasticity trade-off hypothesis (TOH) at the lineage or intraspecific level (reviewed in van Heerwaarden and Kellermann, 2020; see also Gunderson, 2023). In any case, the generality (or lack thereof) of the TOH remains unclear. This lack of clarity is a critical problem because incorrect conclusions about the TOH could lead to a fundamental misunderstanding of which species are most vulnerable to climate change, particularly those already living close to their upper physiological thermal limits.

van Heerwaarden and Kellermann (2020) argue that the problem is not with the TOH itself, but with the way experiments and/or statistical tests have been conducted (see also van Heerwaarden et al., 2024). Gunderson (2023) analysed studies at the intraspecific level that supported the TOH but suggested that regression to the mean had led to a significant overestimation of support for the TOH, and that this should be considered in future tests of the hypothesis.

Testing the TOH can be plagued by false positives due to statistical bias, or false negatives due to inappropriate statistical analysis and/or

* Corresponding author. Departament de Genètica i de Microbiologia, Grup de Genòmica, Bioinformàtica i Biologia Evolutiva (GBBE), Universitat Autònoma de Barcelona, Spain.
E-mail address: mauro.santos@uab.es (M. Santos).

https://doi.org/10.1016/j.jtherbio.2025.104248
Received 22 January 2025; Received in revised form 26 July 2025; Accepted 15 August 2025
Available online 1 September 2025

low statistical power as we will show here. Many studies estimate thermal plasticity as the change in thermal tolerance ΔCT_{max} across two environments (21 of 30 studies reviewed by van Heerwaarden and Kellermann, 2020), where the upper thermal limit or critical thermal maximum CT_{max} is estimated by any appropriate method: (i) heat tolerance measured at a reference or basal temperature ($CT_{max(1)}$), and (ii) heat tolerance after heat hardening, where sub-lethal exposure to thermal stress can temporarily increase thermal tolerance (Bowler, 2005), or heat acclimation to improve thermoregulation ($CT_{max(2)}$; $\Delta CT_{max} = CT_{max(2)} - CT_{max(1)}$). We acknowledge that the interplay between time and temperature dosing can significantly influence conclusions about thermal tolerance. However, our focus here is on studies where experiments are designed to avoid deleterious acclimation to suboptimal temperatures, allowing us to concentrate on the relationship between basal tolerance and heat acclimation capacity. The Pearson product-moment correlation or Spearman's rank correlation coefficient between ΔCT_{max} and $CT_{max(1)}$ can be used to test for the trade-off. The problem is that the two variables, plastic response and basal or reference heat tolerance, share the common index $CT_{max(1)}$, resulting in a "spurious" correlation (Pearson, 1897; Jackson and Somers, 1991; Kronmal, 1993). Pearson (1897) defined spurious correlations as correlations caused solely by data transformations that do not reflect meaningful properties of the underlying data.

The motivation for the present work came largely from reading the works of Deery et al. (2021) and Gunderson (2023). Deery et al. (2021) studied heat tolerance plasticity in two lizard species: *Anolis carolinensis* and *Anolis sagrei*. They measured basal heat tolerance $CT_{max(1)}$ and subsequent heat hardening $CT_{max(2)}$ in a total of 97 lizards, but used a subset of animals (30 *A. carolinensis* and 35 *A. sagrei*) to test for a trade-off between heat hardening capacity and basal heat tolerance. Heat tolerance plasticity was estimated as $\Delta CT_{max} = CT_{max(2)} - CT_{max(1)}$ for each species. To avoid false positives, Deery et al. (2021) used the randomization approach suggested by Jackson and Somers (1991) to test the null hypothesis of a nonsignificant correlation between ΔCT_{max} and basal heat tolerance $CT_{max(1)}$ (Fig. 4 in Deery et al., 2021). They also reanalysed the data from Phillips et al. (2016) using the randomization approach (Figure S1 in Deery et al., 2021). In all cases, Deery et al. (2021) concluded that the null hypothesis that there was no relationship between basal heat tolerance and heat hardening capacity could not be rejected.

Our point here is that while the randomization approach used by Deery et al. (2021) correctly generates the null correlation histograms, there is no clear prediction of what to expect from the TOH. For example, the null correlation histograms plotted by Deery et al. (2021) show vertical lines indicating the one-sided 95th percentile threshold of the permuted values, so they are testing $H_0 : \rho = \rho_0$ versus $H_1 : \rho < \rho_0$; i.e. a one-tailed test where the decision rule is that the observed empirical correlation is lower than the null expectation. We believe that their reasoning was based on the following (mistaken) intuition. In Fig. 1A, we recast Jackson and Somers (1991) by showing $N = 200$ simulated data for two variables X and Y derived from a bivariate normal distribution with parameters $\mu = [100 \quad 100]$ and covariance matrix $\Sigma = \begin{bmatrix} 625 & 0 \\ 0 & 625 \end{bmatrix}$; i.e. both variables have the same mean and variance, but they are uncorrelated. In Fig. 1B, we plot the null correlation histogram after 1000 random permutations of the raw data, and the blue line shows the empirical Pearson correlation $r_{X,Y} = 0.01$. The TOH posits that there is a negative correlation between basal heat tolerance and heat acclimation/hardening capacity; however, due to the spurious correlation, we cannot longer use $H_0 : \rho = 0$ versus $H_1 : \rho < 0$ as the hypotheses to be tested and have to derive the null correlation histogram that results after the transformation $\Delta = Y - X$. Fig. 1C shows the scatterplot after the

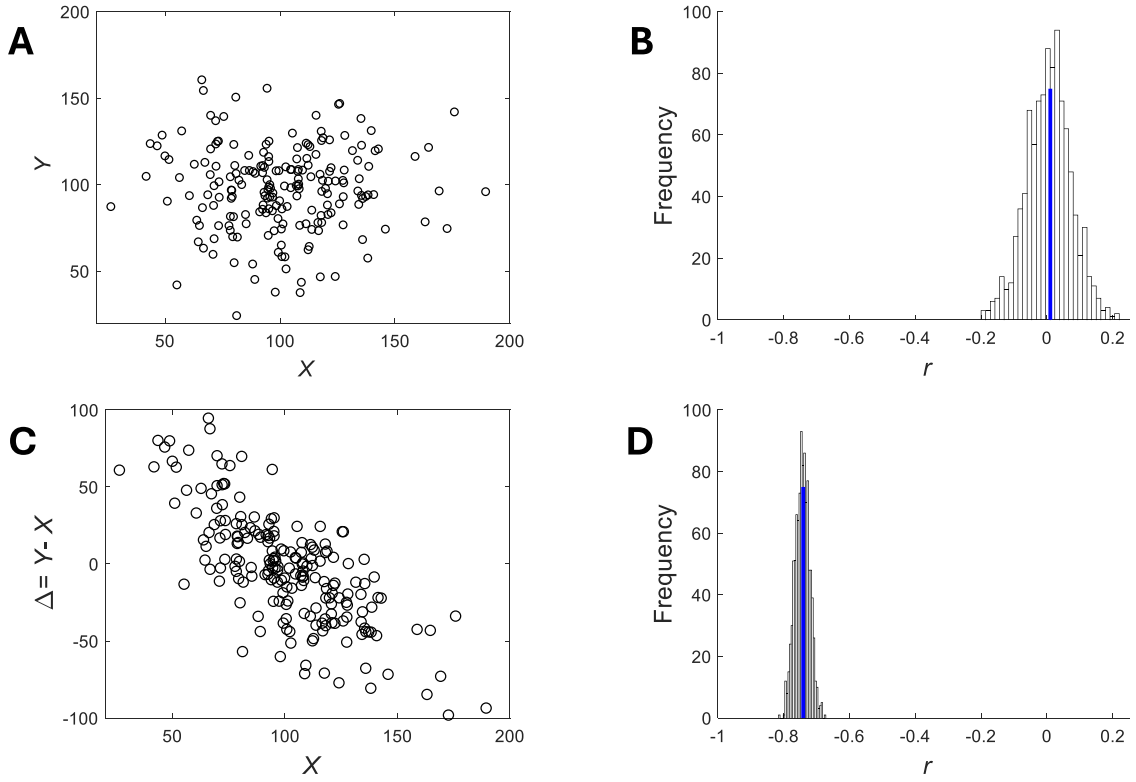

Fig. 1. (A) Scatterplot of $N = 200$ simulated data points for two variables X and Y with Pearson correlation $r_{X,Y} = 0.01$. (B) Null correlation histogram obtained after 1000 random permutations of the data in A. The blue line shows the empirical correlation $r_{X,Y}$. (C) Scatterplot of $\Delta = Y - X$ and X showing a strong negative (spurious) correlation $r_{X,Y-X} = -0.74$. (D) Null correlation histogram obtained after 1000 random permutations of the transformed data in C. The blue line shows the empirical correlation $r_{X,Y-X}$.

transformation, and Fig. 1D shows the null correlation histogram after 1000 random permutations of the transformed data. This transformation shifted the whole null correlation histogram in Fig. 1B to the left, and the blue line shows the empirical (spurious) correlation $r_{X,Y-X} = -0.74$. Therefore, it seems intuitively reasonable to use $H_0 : \rho = \rho_0$ versus $H_1 : \rho < \rho_0$ as the hypotheses to be tested, where the usual null correlation of 0 has been replaced by the expected null (spurious) correlation ρ_0 resulting from the data transformation. Unfortunately, this intuition is disastrously wrong and can lead to many false negatives (failing to reject the null hypothesis when the TOH is true), thereby resulting in the erroneous conclusion that thermal phenotypic plasticity can facilitate population persistence. This is a critical error, particularly for ectotherms that are already living close to their upper physiological thermal limits. The reason for this is that the TOH may still hold even if the correlation between basal heat tolerance and heat acclimation/hardening capacity is significantly higher than the null expectation after transforming the raw data, as long as the correlation is negative, i.e. in Fig. 1D there is plenty of room between the spurious correlation of the transformed raw data and 0.

These problems arise because, to our knowledge, no one has calculated how the expected correlation between heat acclimation/hardening capacity and basal heat tolerance $[\rho(\Delta CT_{max}, CT_{max(1)})]$ that tests for the TOH changes as a function of the statistical properties of the original variables $CT_{max(1)}$ and $CT_{max(2)}$. This is an example of a wider problem in ecology and evolutionary biology, where "data are typically collected without any pre-study determination and justification of reasonable null and alternative hypotheses or of decision rules and decision costs" (Berner and Amrhein, 2022, p. 778). In what follows, we first derive the expected values of $\rho(\Delta CT_{max}, CT_{max(1)})$ as a function of the statistical properties of the original variables. This section is necessarily a bit technical, but it is important to understand all the statistical subtleties involved after the seemingly simple transformation $\Delta = Y - X$ $(\Delta CT_{max} = CT_{max(2)} - CT_{max(1)})$. Another issue we address is statistical power. It is unfortunately common practice not to formally calculate statistical power at the design stage of an experiment (Ellis, 2010), which directly impacts our ability to discern true biological effects. Without adequate power, there is a high risk of committing a Type II error ($\beta = 1 -$ power), incorrectly accepting a false null hypothesis when a true effect exists. This leads to the fundamental problem that many studies testing the TOH are likely underpowered, which severely impacts the validity of their statistical conclusions and interpretations. Finally, we vindicate Phillips et al.'s (2016) suggestion of a trade-off between basal heat tolerance and heat hardening in the lizard *L. coggeri*, contradicting the claims of Deery et al. (2021) (see also Gunderson, 2023), and discuss the issue of regression to the mean in the context of the randomization approach to hypothesis testing.

All numerical data reported here have been independently double-checked. M.S. performed analyses in the MATLAB (2024) algebra environment using tools provided by the Statistics Toolbox. For trend analysis, we also used the nonparametric Mann-Kendall tau function 'ktaub' (Burkey 2005) implemented in MATLAB. J.F.F performed analyses in Fortran.

2. Derivation of expected values

To simplify the notation, let $X = CT_{max(1)}$, $Y = CT_{max(2)}$, and $\Delta = \Delta CT_{max} = Y - X$. From the standard expression for the variance of a difference of two random variables we have:

$$\begin{aligned}\mathbb{V}(\Delta) &= \mathbb{V}(X) + \mathbb{V}(Y) - 2\mathrm{Cov}(X,Y)\\ &= \mathbb{V}(X) + \mathbb{V}(Y) - 2\rho(X,Y)\sqrt{\mathbb{V}(X)\mathbb{V}(Y)},\end{aligned} \tag{1}$$

where $\mathrm{Cov}(X,Y)$ is the covariance and $\rho(X,Y)$ is the Pearson product-moment correlation. We also have:

$$\begin{aligned}\mathrm{Cov}(\Delta,X) &= \mathrm{Cov}(X,Y) - \mathbb{V}(X)\\ &= 2\rho(X,Y)\sqrt{\mathbb{V}(X)\mathbb{V}(Y)} - \mathbb{V}(X).\end{aligned} \tag{2}$$

Therefore, it can be shown that:

$$\begin{aligned}\rho(\Delta,X) &= \frac{\mathrm{Cov}(X,Y) - \mathbb{V}(X)}{\sqrt{\mathbb{V}(\Delta)\mathbb{V}(Y)}}\\ &= \sqrt{\frac{\mathbb{V}(Y)}{\mathbb{V}(\Delta)}}\rho(X,Y) - \sqrt{\frac{\mathbb{V}(X)}{\mathbb{V}(\Delta)}}.\end{aligned} \tag{3}$$

This expression shows that the correlation between heat acclimation/hardening capacity $(\Delta = \Delta CT_{max})$ and basal heat tolerance $(X = CT_{max(1)})$ that tests for the TOH $[\rho(\Delta CT_{max}, CT_{max(1)})]$ is a complex function of the variance of the original variables $[\mathbb{V}(X) = \mathbb{V}(CT_{max(1)});\ \mathbb{V}(Y) = \mathbb{V}(CT_{max(2)})]$, their correlation $[\rho(X,Y) = \rho(CT_{max(1)}, CT_{max(2)})]$, and the variance of the heat acclimation/hardening capacity $[\mathbb{V}(\Delta) = \mathbb{V}(\Delta CT_{max})]$. The TOH is supported if $\rho(\Delta,X) < 0$ or equivalently:

$$\rho(X,Y) < \sqrt{\frac{\mathbb{V}(X)}{\mathbb{V}(Y)}}. \tag{4}$$

This inequality reveals the main difficulty in defining the TOH, since it is satisfied when X and Y are independent, i.e., $\rho(X,Y) = \rho(CT_{max(1)}, CT_{max(2)}) = 0$. Thus, finding a negative $\rho(\Delta,X)$ correlation is not sufficient to conclude that the TOH is valid: the correlation must be substantially different from the null correlation, i.e., the correlation when X and Y are independent. This means that the concept of statistical hypothesis testing already appears in the attempt to explicitly formalize the TOH. In the words of Pearson (1897, p. 491): "A part of the correlation he discovers between organs is undoubtedly organic, but another part is solely due to the nature of his arithmetic, and as a measure of organic relationship is spurious."

The previous treatment makes it possible to analyse the behaviour of the expected correlation $\rho(\Delta CT_{max}, CT_{max(1)})$ as a function of the statistical properties of the original variables $CT_{max(1)}$ and $CT_{max(2)}$. Fig. 2 shows this behaviour as a function of the value of the correlation between $CT_{max(1)}$ and $CT_{max(2)}$, $\rho(CT_{max(1)}, CT_{max(2)})$, and the ratio of their variances $\Gamma = \mathbb{V}(CT_{max(2)})/\mathbb{V}(CT_{max(1)})$. The arrow points to the null expectation $\rho_0 = -\sqrt{\mathbb{V}(CT_{max(1)})/[\mathbb{V}(CT_{max(1)}) + \mathbb{V}(CT_{max(2)})]}$, which is obtained by setting $\rho(X,Y) = 0$ in equation (3). Note that in Fig. 2 $\rho(\Delta CT_{max}, CT_{max(1)})$ is always less than 0, so inequality (4) holds. Parenthetically, note also that in Fig. 1 the variables X and Y were assumed to be uncorrelated with variances $\mathbb{V}(X) = \mathbb{V}(Y) = 625$ (and thus $\Gamma = 1$). For this specific simulation, the null expected spurious correlation is $\rho_0 = -\sqrt{625/(625+625)} = -0.707$, and our random sample yielded $r_{X,Y-X} = -0.74$. This uncorrelated scenario is crucial for our framework as it establishes a critical baseline for the statistical test. Our approach fundamentally involves comparing observed correlations when X and Y are uncorrelated [i.e., $\rho(X,Y) = 0$] against cases where they are correlated [i.e., $\rho(X,Y) \neq 0$]. This comparison is essential for distinguishing between spurious correlations and genuine biological relationships, thereby providing a robust method for testing the TOH.

As an illustrative numerical example to show that by testing $H_0 : \rho = \rho_0$ versus $H_1 : \rho < \rho_0$ we will always get a false negative if the correlation between $CT_{max(1)}$ and $CT_{max(2)}$ is positive but lower than the upper bound in equation (4), i.e. $0 < \rho(CT_{max(1)}, CT_{max(2)}) < \sqrt{\mathbb{V}(CT_{max(1)})/\mathbb{V}(CT_{max(2)})}$, assume $N = 200$ simulated data for a given species derived from a bivariate normal distribution with parameters $\mu = [40.1 \quad 42.6]$ for $CT_{max(1)}$ and $CT_{max(2)}$ (in degrees Celsius), respectively, and covariance matrix $\Sigma = \begin{bmatrix} 1 & 0.4243 \\ 0.4243 & 0.5 \end{bmatrix}$; i.e. $\rho(CT_{max(1)}, CT_{max(2)}) = 0.6 < \sqrt{1/0.5}$, so inequality (4) is satisfied. Fig. 3A shows the change in heat tolerance ΔCT_{max} as a function of basal heat tolerance

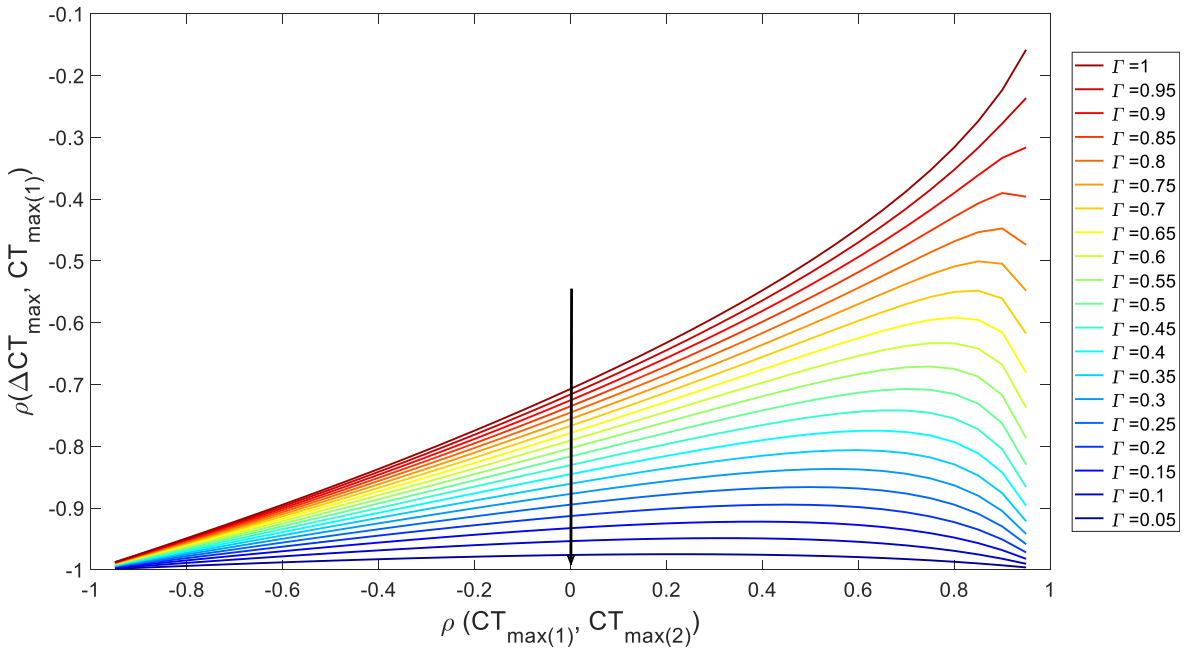

Fig. 2. Expected correlation of heat hardening capacity and basal heat tolerance $\rho(\Delta CT_{max}, CT_{max(1)})$ as a function of the correlation between basal and induced heat tolerance $\rho(CT_{max(1)}, CT_{max(2)})$, and the ratio $\Gamma = \mathbb{V}(CT_{max(2)})/\mathbb{V}(CT_{max(1)})$. We set $\mathbb{V}(CT_{max(1)}) = 1$ without loss of generality. The arrow points to the null expectation $\rho_0(\Delta CT_{max}, CT_{max(1)}) = -\sqrt{\mathbb{V}(CT_{max(1)})/[\mathbb{V}(CT_{max(1)}) + \mathbb{V}(CT_{max(2)})]}$ when $\rho(CT_{max(1)}, CT_{max(2)}) = 0$ (see text for details).

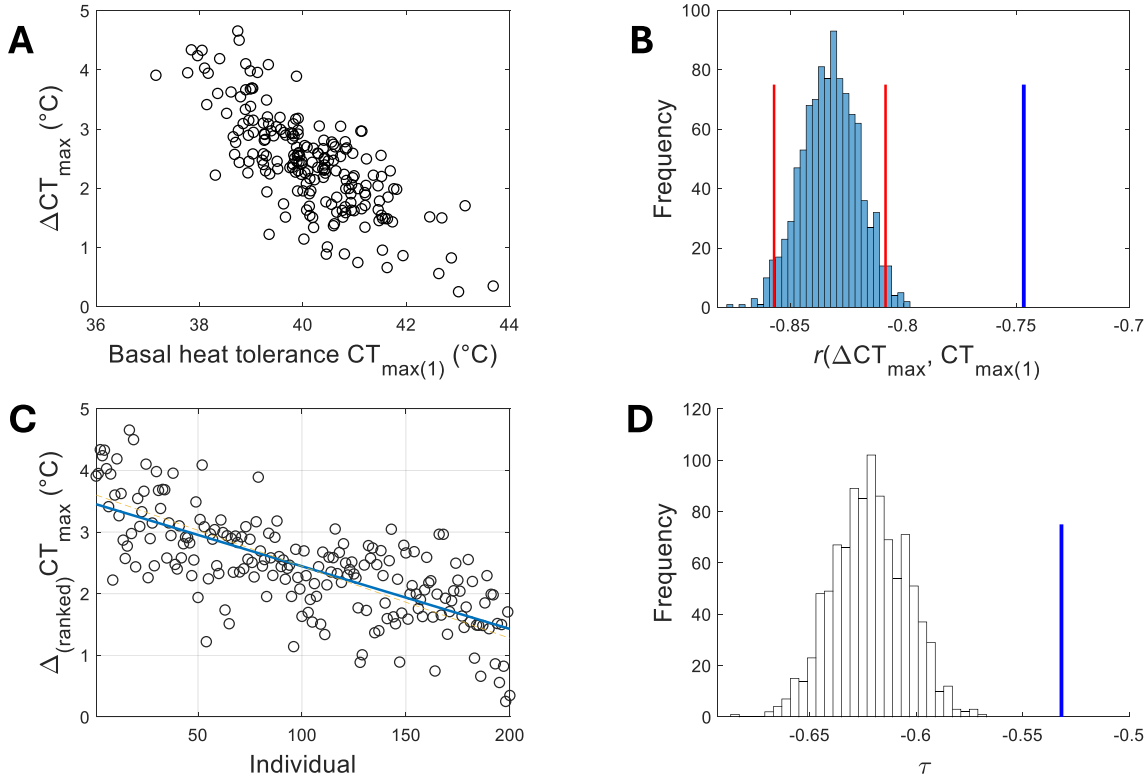

Fig. 3. (A) Change in heat tolerance (ΔCT_{max}) as a function of basal heat tolerance. The plot is based on a random sample of $N = 200$ animals derived from a bivariate normal distribution with parameters $\mu = [40.1 \quad 42.6]$ for $CT_{max(1)}$ and $CT_{max(2)}$ (in degrees Celsius), respectively, and covariance matrix $\Sigma = \begin{bmatrix} 1 & 0.4243 \\ 0.4243 & 0.5 \end{bmatrix}$. (B) Null distribution of Pearson product-moment correlation coefficients calculated after 1000 random permutations. The red lines indicate the two-sided 95th percentile threshold of permuted values, and the blue line the empirical correlation coefficient. (C) Same data but now plotted as the ranked heat tolerance plasticity ($\Delta_{(ranked)}$ CT_{max}), namely, the difference $CT_{max(2)} - CT_{max(1)}$ ranked according to $CT_{max(1)}$ in ascending order, against the individual number (according to the ranked $CT_{max(1)}$). (D) Null distribution of the nonparametric Mann-Kendall τ for trend after 1000 random permutations of the ranked change in heat tolerance. The blue line shows the empirical $\tau(\tau = -0.53)$.

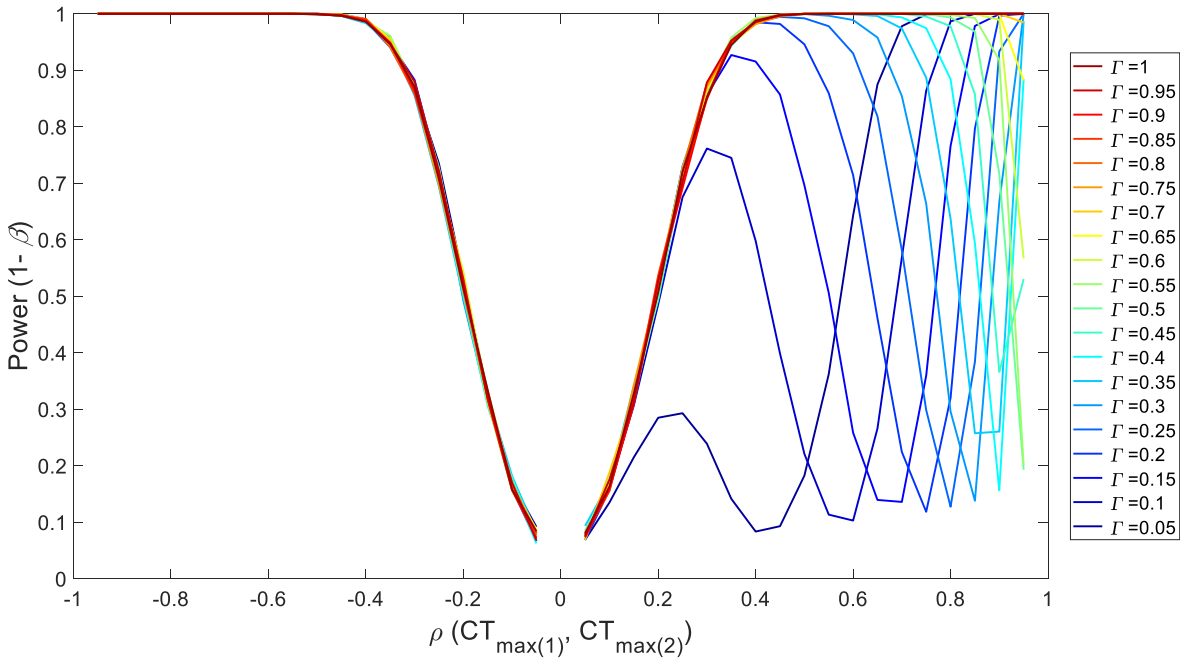

Fig. 4. Power analysis assuming $N = 100$ individuals. To estimate power, we obtained synthetic data from bivariate normal distributions with parameters $\mu = [40.1 \quad 42.6]$ and covariance matrices Σ according to $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)})$ and $\Gamma = \mathbb{V}(\mathrm{CT}_{\mathrm{max}(2)})/\mathbb{V}(\mathrm{CT}_{\mathrm{max}(1)})$. In each case, we tested the TOH after 1000 random permutations of the data and assigned a value of 1 if the null hypothesis was rejected and a 0 otherwise. This was repeated 2000 times, so that for each combination of parameter values, the figure shows the average of 2000 independent runs.

$\mathrm{CT}_{\mathrm{max}(1)}$, and Fig. 3B shows the distribution of Pearson correlation coefficients after 1000 random permutations of the data. The red lines define the two-tailed 95 % percentile interval of the permuted values, and the blue line shows the empirical correlation, which is clearly outside the 95 % percentile interval. Perhaps it is easier to see that the TOH holds in this numerical example by plotting the data as in Fig. 3C, which shows the ranked heat tolerance plasticity ($\Delta_{(\mathrm{ranked})}\ \mathrm{CT}_{\mathrm{max}}$), namely, the difference $\mathrm{CT}_{\mathrm{max}(2)} - \mathrm{CT}_{\mathrm{max}(1)}$ ranked according to $\mathrm{CT}_{\mathrm{max}(1)}$ in ascending order, against the individual number (according to the ranked $\mathrm{CT}_{\mathrm{max}(1)}$). The essence of the TOH is that individuals that have evolved the greatest basal heat tolerance (those in the lower right of Fig. 3C) will be most vulnerable to a further increase in habitat temperature due to their lower acclimation capacity of $\mathrm{CT}_{\mathrm{max}}$ (lower $\Delta_{(\mathrm{ranked})}\ \mathrm{CT}_{\mathrm{max}}$). If TOH holds, we expect a negative and statistically significant trend for $\Delta_{(\mathrm{ranked})}\ \mathrm{CT}_{\mathrm{max}}$, and we can use the nonparametric Mann-Kendall statistical test for trend to assess whether $\Delta_{(\mathrm{ranked})}\ \mathrm{CT}_{\mathrm{max}}$ is increasing or decreasing, and whether the trend in either direction is statistically significant (Helsel et al. 2020). In Fig. 3D we show the null distribution histogram of the Mann-Kendall τ after 1000 random permutations of the ranked change in heat tolerance, and the blue line shows the empirical $\tau (\tau = \ -0.53)$, which is clearly outside the null distribution histogram.

To summarize, if we use (e.g.) the randomization approach suggested by Jackson and Somers (1991) to test the null hypothesis of a nonsignificant correlation between the change in heat tolerance $\Delta\mathrm{CT}_{\mathrm{max}}$ as a function of basal heat tolerance $\mathrm{CT}_{\mathrm{max}(1)}$, the following hypotheses should be tested: $\mathrm{H}_0 : \rho = \rho_0$ versus $\mathrm{H}_1 : \rho \neq \rho_0$, i.e. a two-tailed test. This avoids obtaining false negatives when $0 < \rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)}) < \sqrt{\mathbb{V}(\mathrm{CT}_{\mathrm{max}(1)})/\mathbb{V}(\mathrm{CT}_{\mathrm{max}(2)})}$ and $\mathrm{H}_1 : \rho < \rho_0$ as used by Deery et al. (2021).

3. Power analysis

Ideally, a statistical power analysis should be performed when planning a TOH experiment across two environments (Ellis, 2010, pp. 59–60). This involves calculating the probability of correctly rejecting a false hypothesis when a particular alternative hypothesis is true. The challenge has been that a prospective or *a priori* power analysis has not been possible, simply because the statistical properties of the correlation $\rho(\Delta\mathrm{CT}_{\mathrm{max}}, \mathrm{CT}_{\mathrm{max}(1)})$ as a function of the original variables $\mathrm{CT}_{\mathrm{max}(1)}$ and $\mathrm{CT}_{\mathrm{max}(2)}$ were not fully understood. Therefore, researchers did not have a good sense of what constitutes a sensible sample size and the associated level of power.

Here, our aim is to illustrate how the statistical power of testing the TOH across two environments changes as a function of $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)})$ and the ratio $\Gamma = \mathbb{V}(\mathrm{CT}_{\mathrm{max}(2)})/\mathbb{V}(\mathrm{CT}_{\mathrm{max}(1)})$. We hope this will also help researchers determine the necessary sample size to achieve a specified level of statistical power when planning a study to test the TOH across two environments. Fig. 4 presents the statistical power assuming $N = 100$, a sample size considered reasonably large for this type of experiments (e.g., Deery et al., 2021). Note that Fig. 4 skips the power value when $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)}) = 0$. This is because in this case we are not talking about power, but type I error, which here is around 5 % (average 0.051) for all values of Γ, as expected from a type I error $\alpha = 0.05$. For $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)}) < 0$ power is basically the same regardless of Γ, but for $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)}) > 0$ power is a complex function of $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)})$ and Γ. These surprising results can be understood by comparing Figs. 4 and 2. The bending down function $\rho(\Delta\mathrm{CT}_{\mathrm{max}}, \mathrm{CT}_{\mathrm{max}(1)})$ when $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)}) > 0$ and $\Gamma \leq 0.5$ or so in Fig. 2 makes the correlation coefficient between heat tolerance $\Delta\mathrm{CT}_{\mathrm{max}}$ and basal heat tolerance $\mathrm{CT}_{\mathrm{max}(1)}$ similar to the null expectation ρ_0, which dramatically reduces power for some values of the parameters. In these cases, we will never get a test powerful enough to reject the null hypothesis unless N is unrealistically large (several thousand).

Our analysis reveals a crucial point: while statistical power naturally decreases as the population correlation $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)})$ approaches 0, we found that power can also be greatly reduced, even for relatively large $\rho(\mathrm{CT}_{\mathrm{max}(1)}, \mathrm{CT}_{\mathrm{max}(2)})$, depending significantly on the ratio of the variances of $\mathrm{CT}_{\mathrm{max}(1)}$ and $\mathrm{CT}_{\mathrm{max}(2)}$. This highlights a previously unappreciated factor that critically impacts the ability to robustly test the

TOH, often increasing the risk of false negatives (Type II errors) by failing to reject the null hypothesis ($H_0 : \rho = \rho_0$) when it should be.

4. Empirical data

In Fig. 5, we reanalyse the data for *A. carolinensis* (Deery et al., 2021) and for *L. coggeri* (Phillips et al., 2016) independently for the two observers (Observer 1: VL; Observer 2: AH), as in Deery et al. (2021, their Figure S1). Fig. 5A shows the change in heat tolerance as a function of basal heat tolerance for *A. carolinensis*, and Fig. 5B shows the null correlation histograms after 1000 random permutations. The empirical correlation, $r = -0.65$ (blue line), falls within the two-tailed 95 % percentile interval (red lines), consistent with the conclusions of Deery et al. (2021) that the null hypothesis of no relationship between basal heat tolerance and heat hardening capacity cannot be rejected. However, it is important to note that the relatively small sample ($N = 30$) and the weak correlation observed between $CT_{max(1)}$ and $CT_{max(2)}$ ($r = 0.12$) in this experiment suggest that the study may have been underpowered. While the data are consistent with a null relationship, these factors mean we cannot definitively rule out the existence of a true underlying relationship that the experiment lacked the power to detect.

Fig. 5C–E and 5D-5F show similar plots for *L. coggeri* (observer 1 and observer 2). The empirical correlation between heat hardening plasticity and initial heat tolerance is $r = -0.77$ for observer 1 and $r = -0.52$ for observer 2, and in both cases these correlations are less negative than the null expectation: the blue lines in Fig. 5D and F pointing to the empirical correlations are well outside the two-tailed 95 % percentile interval (red lines). Therefore, for the lizard *L. coggeri*, the TOH appears to hold for both observers, contrary to the claims of Deery et al. (2021) after reanalysis of these data. We note that the correlation between $CT_{max(1)}$ and $CT_{max(2)}$ was always positive in *L. coggeri*: $r(CT_{max(1)}, CT_{max(2)}) = 0.70$ for observer 1 and $r(CT_{max(1)}, CT_{max(2)}) = 0.73$ for observer 2. These examples provide an empirical illustration of false negatives by testing $H_0 : \rho = \rho_0$ versus $H_1 : \rho < \rho_0$, i.e. a left-tailed test as the alternative

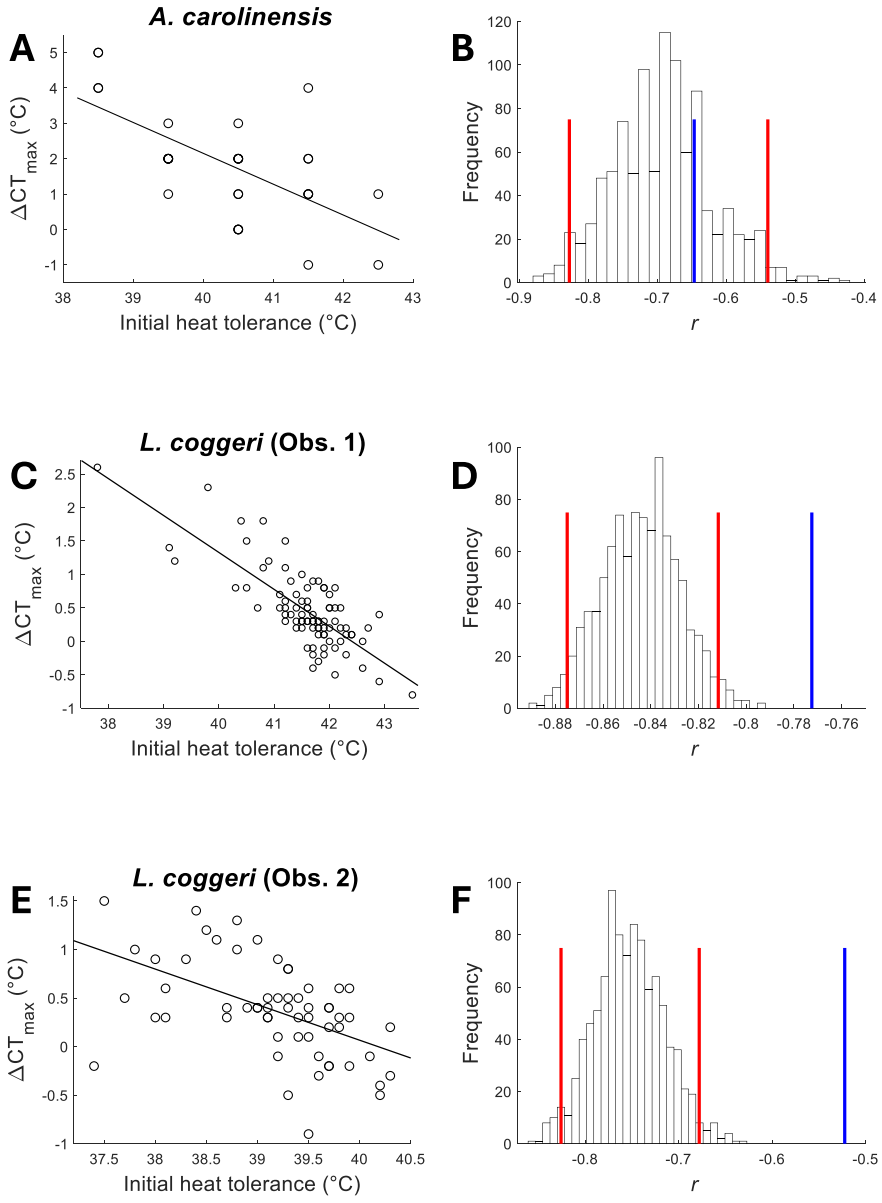

Fig. 5. Re-analysis of data for *Anolis carolinensis* (Deery et al., 2021) and for *Lampropholis coggeri* (Phillips et al., 2016) based on data collected by two observers. (A, C, E) Change in heat tolerance (ΔCT_{max}) as a function of basal heat tolerance. (B, D, F) Null distribution of Pearson product-moment correlation coefficients calculated after 1000 random permutations of the data in A, C, and E. The red lines indicate the two-sided 95th percentile threshold of permuted values, and the blue lines the empirical correlation coefficients.

hypothesis.

Note that the negative correlation between an individual's initial heat tolerance and the change in that value between the first and second measurements is a classic example of a problem where regression to the mean can be mistaken for causal factors (Kelly and Price, 2005), such as TOH (Gunderson, 2023). In well-designed experimental studies, this effect is removed because it affects both experimental and control groups. This observation motivated a theoretical method to correct for the regression effect based on resampling to generate a null hypothesis distribution that would describe the control group (Cichoń et al., 1999). This is the method used in our analysis, and the assumption that the initial and final values of an individual's heat tolerance are uncorrelated is equivalent to complete regression to the mean. Although this assumption has been heavily criticized in the context of mass loss in animals where it was originally applied (Ruf, 2000), we see no fundamental problem with the assumption that $CT_{max(1)}$ and $CT_{max(2)}$ are uncorrelated. Although there are alternative methods to correct for regression effects (Kelly and Price, 2005; Gunderson, 2023), the choice of an appropriate null hypothesis seems more natural and flexible (Cichoń et al., 1999; Deery et al., 2021), as it is the theoretical analogue of control groups in experimental studies.

In summary, our application of the proposed statistical framework to existing empirical datasets yielded contrasting results: for *A. carolinensis*, while we found no significant evidence to reject the null hypothesis of no relationship, this finding should be interpreted cautiously given the likely underpowered nature of that study (due to small sample size and weak observed correlation). In contrast, for *L. coggeri*, the data strongly supported the rejection of the null hypothesis, indicating a clear trade-off.

5. Conclusions

Because of the inherent link between plasticity and basal thermal tolerance, the approach of Deery et al. (2021) appears to be gaining momentum for testing the TOH (Gunderson and Revell, 2022). They estimated thermal plasticity as the change in thermal tolerance as a function of basal heat tolerance and used the randomization method proposed by Jackson and Somers (1991) to evaluate hypotheses confounded by spurious correlations. However, we show here that that their one-sided alternative hypothesis $H_1 : \rho < \rho_0$ is incorrect, leading to spurious conclusions, specifically an increased risk of false negatives (Type II errors) where a trade-off exists. For instance, their approach would have incorrectly failed to reject the null hypothesis for *L. coggeri*, despite our robust analysis revealing a clear trade-off for this species. This can result in the erroneous belief that thermal phenotypic plasticity provides a greater buffer against warming than it actually does, particularly for species already near their upper thermal limits. We also suggest an alternative trend analysis, which is more visually appealing, to appreciate the subtleties between the predictions of the TOH and the effect of regression to the mean.

Finally, our attempt to formalize the TOH revealed that the mere finding of a negative $\rho(\Delta CT_{max}, CT_{max(1)})$ correlation is not sufficient to support this hypothesis, since this negative correlation is obtained when the basal and plastically induced heat tolerances are independent random variables. Thus, we suggest that the null hypothesis must be part of the definition of the TOH, highlighting that its rigorous testing is uniquely susceptible to statistical artefacts and therefore fundamentally requires a robust statistical framework.

In summary, our study provides a robust statistical framework, utilizing a well-chosen two-tailed randomization test, that offers a solution to both the pervasive statistical bias and regression-to-the-mean issues in testing the tolerance-plasticity trade-off hypothesis (TOH). By correctly comparing observed results against a relevant null hypothesis, our approach allows for a far more accurate assessment of the TOH. This directly changes our interpretation of existing empirical findings. As demonstrated with *L. coggeri*, what was previously considered a lack of trade-off may, in fact, be a significant one when rigorously tested. This is profoundly important for climate change responses: correctly evaluating the TOH is crucial for accurately estimating the buffering capacity of thermal plasticity in ectotherms. Without a sound statistical basis, we risk generating misleading conclusions that could lead to overoptimistic predictions about species' resilience, potentially misguiding conservation strategies and underestimating the true threat of warming for species already living close to their upper physiological thermal limits. Our work provides the necessary methodological rigor to ensure that our understanding of thermal tolerance and plasticity is built on solid ground, enabling more precise predictions of how biological systems will respond to a rapidly changing climate.

CRediT authorship contribution statement

Mauro Santos: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **José F. Fontanari:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Data accessibility statement

This research does not include original datasets. All data analysed and discussed are derived from previously published studies, which are cited accordingly within the manuscript.

Declaration of competing interest

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *Physics of Life Reviews* and was not involved in the editorial review or the decision to publish this article.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: None.

Acknowledgments

MS is funded by grant PID2021-127107NB-I00 from Ministerio de Ciencia e Innovación (Spain) and grant 2021 SGR 00526 from Generalitat de Catalunya, Spain. JFF is partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil), grant number 305620/2021-5.

Data availability

The research described in the article did not use any original data.

References

Berner, D., Amrhein, V., 2022. Why and how we should join the shift from significance testing to estimation. J. Evol. Biol. 35, 777–787.
Bowler, K., 2005. Acclimation, heat shock and hardening. J. Therm. Biol. 30, 125–130.
Burkey, J., 2025. Mann-Kendall Tau-b with Sen's method (enhanced). MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/11190-mann-kendall-tau-b-with-sen-s-method-enhanced.
Cichoń, M., Merilä, J., Hillström, L., Wiggins, D., 1999. Mass-dependent mass loss in breeding birds: getting the null hypothesis right. Oikos 87, 191–194.
Deery, S.A., Rej, J.E., Haro, D., Gunderson, A.R., 2021. Heat hardening in a pair of *Anolis* lizards: constraints, dynamics and ecological consequences. J. Exp. Biol. 224, jeb240994.
Ellis, P.D., 2010. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge University Press, New York.
Gunderson, A.R., 2023. Trade-offs between baseline thermal tolerance and thermal tolerance plasticity are much less common than it appears. Glob. Change Biol. 29, 3519–3524.
Gunderson, A.R., Revell, L.J., 2022. Testing for genetic assimilation with phylogenetic comparative analysis: conceptual, methodological, and statistical considerations. Evolution 76, 1942–1952.

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., Gilroy, E.J., 2020. Statistical Methods in Water Resources. U.S. Geological Survey Techniques and Methods, p. 458. https://doi.org/10.3133/tm4a3 book 4, chap. A3.

Jackson, D.A., Somers, K.M., 1991. The spectre of 'spurious' correlations. Oecologia 86, 147–151.

Kelly, C., Price, T.D., 2005. Correcting for regression to the mean in behavior and ecology. Am. Nat. 166, 700–707.

Kronmal, R.A., 1993. Spurious correlation and the fallacy of the ratio standard revisited. J. Roy. Stat. Soc. 156, 379–392.

MATLAB, 2024. Version R2024a Update 6. The MathWorks Inc, Natick, MA.

Parmesan, C., Morecroft, M.D., Trisurat, Y., Adrian, R., Anshari, G.Z., Arneth, A., Gao, Q., Gonzalez, P., Harris, R., Price, J., Stevens, N., Talukdarr, G.H., 2022. Terrestrial and freshwater ecosystems and their services. In: Pörtner, H.-O., Roberts, D.C., Tignor, M., Poloczanska, E.S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B. (Eds.), *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 197–377.

Pearson, K., 1897. Mathematical contributions to the theory of evolution. —On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. Roy. Soc. Lond. 60, 489–497.

Phillips, B.L., Munoz, M.M., Hatcher, A., Macdonald, S.L., Llewelyn, J., Lucy, V., Moritz, C., 2016. Heat hardening in a tropical lizard: geographic variation explained by the predictability and variance in environmental temperatures. Funct. Ecol. 30, 1161–1168.

Ruf, T., 2000. Mass-dependent mass loss: how to get the null hypothesis wrong. Oikos 90, 413–416.

Stillman, J.H., 2003. Acclimation capacity underlies susceptibility to climate change. Science 301, 65.

van Heerwaarden, B., Kellermann, V., 2020. Does plasticity trade off with basal heat tolerance? Trends Ecol. Evol. 35, 874–885.

van Heerwaarden, B., Sgrò, C., Kellermann, V.M., 2024. Threshold shifts and developmental temperature impact trade-offs between tolerance and plasticity. Proceedings of the Royal Society B 291, 20232700.