This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier’s archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright
Measurement error in heat tolerance assays

Luis E. Castañeda a,b,n, Gemma Calabria c, Luz A. Betancourt a, Enrico L. Rezende a, Mauro Santos a

a Departament de Genètica i de Microbiologia, Grup de Biologia Evolutiva (GBE), Universitat Autònoma de Barcelona, Bellaterra (Barcelona) 08193, Spain
b Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
c Departament de Genètica, Grup de Biologia Evolutiva (GBE)/Institut de Recerca de la Biodiversitat (IRBio), Facultat de Biologia, Universitat de Barcelona, Diagonal 643, Barcelona 08028, Spain

Abstract

Biological measurements inherently involve some measurement error (ME), which is a major concern because measurement accuracy (how closely a measurement reproduces the true value of the attribute being measured) and statistical power steadily decrease with increasing ME. However, ME has been largely overlooked in the thermal biology literature, which can be explained by the fact that thermotolerance estimates often involve the collapse or death of the tested individuals and measurements cannot be repeated in vivo with the same specimen under identical conditions. Here we assess inter- and intra-researcher (test-retest) reliability of heat tolerance measured as knockdown time from digital recordings of Drosophila subobscura flies individually assayed in vials with a dynamic method. We provide a summary of various estimators used to compute measurement reliability (the degree to which the measurement is affected by ME) together with their statistical properties. Our results indicate that the estimation of heat knockdown time has poor reliability: intra-researcher ME = 29% and inter-researcher ME = 34%. This difference is attributed to lack of ‘accurateness’ (the difference in the marginal distributions of the measurements taken by the two researchers) because measurement imprecision was essentially the same in both estimates (27%). In view of our results we conclude that thermal biologists should report the reliability of thermotolerance estimates and, when necessary, adopt some straightforward guidelines suggested here to improve measurement reliability.

1. Introduction

Spatiotemporal variation in environmental temperature is a key determinant of fitness because it imposes limits to organisms’ performance (Huey and Kingsolver, 1989; Huey and Hertz, 1983; Angilletta, 2009). Therefore, thermal biologists have long been concerned with estimating thermal limits; in particular, obtaining reliable estimates of upper thermal limits is currently pressing in view of the ongoing effect of global warming in biological systems (Deutch et al., 2008; Pörtner and Farrell, 2008; Huey et al., 2009). Heat tolerance can be quantified by subjecting the study organism to a constant stressful temperature and record the time to death (static method) or by gradually increasing the test temperature until the end-point is reached (dynamic method or ‘CTmax method’; Lutterschmidt and Hutchinson, 1997). As all other estimators, these estimates of heat tolerance will be affected by measurement error (ME; defined as the variability of repeated measurements of a trait taken on the same individual that is solely a function of the investigator), which is a subject of major concern in various areas of biology; for instance, in analyses of morphometric data (Bailey and Byrnes, 1990; Yezerinac et al., 1992; Muñoz-Muñoz and Perpiñán, 2010) or fluctuating asymmetry (Björklund and Merilä, 1997; Palmer and Strobeck, 2003). Conversely, little attention has been paid to the extent of ME in the thermal biology literature possibly because the estimation of heat tolerance involves the collapse or death of the tested individual and measurements cannot be repeated with the same specimen under identical conditions (see also Rezende and Santos, 2012).

This is particularly odd because, contrary to morphological characters, which tend to show the highest repeatabilities (Wolak et al., 2011), the repeatability of heat tolerance is generally low (about 0.2; Krebs and Loeschcke, 1997). Therefore, whereas ME can be negligible for morphological traits relative to the variability among individuals (with the exception of fluctuating asymmetry because differences between sides are often very small) it might be substantial for heat tolerance estimates (Santos et al., 2011).

Here we evaluate inter- and intra-researcher reliability (i.e., the extent to which measurements are inherently reproducible or, alternatively, influenced by ME) of heat tolerance, estimated as knockdown time from digital recordings of Drosophila subobscura flies individually assayed in vials with a dynamic method.
We first provide a theoretical background that summarizes several concepts and techniques used to compute reliability. We explain the intraclass correlation coefficient originally proposed by Fisher (1946) and generally applied in biology (Lessells and Boag, 1987; Sokal and Rohlf, 1995, p. 213; Falconer and Mackay, 1996, pp. 145–146), and also discuss other coefficient mainly used in the social and clinical sciences to compute measurement reliability that has better properties than Fisher’s coefficient. Finally, we show that the estimation of knockdown time in our assays with *D. subobscura* has poor reliability. This result suggests that ME is an important issue that has been unwarrantedly overlooked in the thermal biology literature.

2. Background theory

Terminology in the area of measurement accuracy is not always consistent; hence here we employ the usage in Hand (2004), which distinguishes between accuracy, validity and reliability. Accuracy is a general term that describes how closely a measurement reproduces the true value of the attribute being measured. This has two aspects: the extent to which a measurement procedure measures what we want to measure (validity, which is similar to bias), and the difference between multiple measures of an attribute (reliability, which is an estimate of the proportion of variation that is not due to ME). Let us assume that heat tolerance is independently estimated by two researchers as the temperature at which each tested individual succumbs to heat stress in a dynamic method (the attribute they want to measure is $CT_{\text{max}}$ defined as ‘the maximum temperature that an organism might potentially tolerate given its physiological condition in the absence of any other hazard’; Santos et al., 2011). The records of the first and second researcher on the $i$th individual ($y_{ij}$, $i=1,2,\ldots,n; j=1,2$) can be modeled as follows:

$$y_{ij} = \tau_i + b_{ij} + e_{ij},$$  \(1\)

where $\tau_i$ is the ‘true value’ ($CT_{\text{max}}$) of the individual that would be observed with a perfect measurement procedure (assuming one exists), $b_{ij}$ is the difference between the expected value obtained from the actual experimental protocol and the true value (bias), and $e_{ij}$ is the error of measurement. However, here we are mainly concerned about ME, or how reproducible measurements between researchers are when they involve the same tested individual under identical conditions (inter-researcher reliability or agreement), and no notion of ‘true value’ is needed in this case. Reliability can be estimated in several ways, one of which involves calculating the intraclass correlation coefficient (ICC) as originally defined by Fisher (1946) (see Lessells and Boag, 1987; Sokal and Rohlf, 1995, p. 213).

We can redefine Eq. (1) as follows:

$$y_{ij} = \delta_i + e_{ij},$$  \(2\)

where $\delta_i$ is the actual thermotolerance of the $i$th individual, which is taken to be the expectation of $y_{ij}$ over the population of possible measurements for the $i$th individual. It follows that the expectation of the ME is zero: that is, $E(\epsilon_{ij})=0$. Assuming that ME between researchers is independent and have constant variance, and that ME is also independent of the actual knockdown temperature or time, we have:

$$\text{Var}(y_{ij}) = \text{Var}(\delta_i) + \text{Var}(e_{ij}).$$  \(3\)

The inter-researcher reliability can then be estimated as follows:

$$\text{ICC} = \frac{\text{Var}(\delta_i)}{\text{Var}(\delta_i) + \text{Var}(e_{ij})} = 1 - \frac{\text{Var}(e_{ij})}{\text{Var}(\delta_i) + \text{Var}(e_{ij})} = 1 - \text{ME},$$  \(4\)

(See Appendix A for the estimation of the variance and confidence intervals for ICC).

Because the intraclass covariance (i.e., the covariance between the pair of measurements in the same individual) is $\text{Cov}(y_{1i},y_{2i}) = \text{Var}(\delta_i)$, it immediately follows that ICC is equal to the Pearson correlation coefficient $\rho$. However, this assumes that the $\delta_i$ are the same within each pair of repeated measurements, otherwise ICC $\neq \rho$ (see below). Eq. (4) contains an important point that is particularly relevant for heat tolerance estimates. ICC and, hence, ME are not invariant to the particular method we use to estimate heat resistance because $\text{Var}(\delta_i)$ will depend on the conditions of the assay. For instance, Chown et al. (2009) found that different heating rates in dynamic assays result in different phenotypic variances for knockdown temperature. If we assume that $\text{Var}(e_{ij})$ will be the same whatever the heating rate, ME will be larger when the variance in knockdown temperatures is smaller.

An alternative method to measure inter-researcher agreement is to use the concordance correlation coefficient CCC (Lin, 1989, 1992; Lin et al., 2002):

$$\text{CCC} = \frac{2\text{Cov}(y_{1i},y_{2i})}{\text{Var}(y_{1i}) + \text{Var}(y_{2i}) + (\text{Var}(\delta_i))^2}$$  \(5\)

where $\text{Var}(y_{1i})$ is the variance of measurements taken by researcher 1 (2), and $\text{Var}(\delta_i)$ is the average. CCC can also be written as follows:

$$\text{CCC} = \chi_{a} \rho;$$  \(6\)

where $w$ is the scale shift that causes an angle deviation from the expected 45° line through the origin (assuming ME is absent) in the $x$-$y$ coordinate plot of measurements, $\rho$ is the location shift relative to the scale that places the expected 45° above or below the origin, and $\rho$ is the Pearson correlation coefficient (Lin, 1989; Lin et al., 2002). CCC is closely related to ICC but has the advantage over ICC that CCC clearly decomposes the inter-researcher agreement in two components: inter-researcher ‘accuracy’ $\chi_{a}$ (i.e., the extent to which the marginal distributions of the measurement in the two researchers are equal) and inter-researcher precision (i.e., the Pearson correlation coefficient between the measurements of the two researchers). Inter-researcher accuracy $\chi_{a}$ is equal to one if and only if both means and variances are equal. In this case ICC=CCC=ρ. Overall ME is estimated as 1-CCC, but we can also have an idea where the main ME comes from (i.e., inter-researcher lack of accuracy or lack of precision). Expressions to estimate the variance and confidence intervals of CCC are given in Appendix A.

We can now understand why the use of ICC to compute reliability or agreement has been objected (Lin, 1989): whenever there is a shift in scale and/or location ICC captures these deviations and considers them as imprecision. However, we still retain the use of ICC because (i) its performance is very similar to that of CCC; (ii) in the biological sciences it is customary to employ ICC and required sample sizes to obtain an estimate with the confidence interval width as small as possible are known (Bonett, 2002; see also figure 3 in Wolak et al. (2011)); and (iii) one problem with CCC is that it cannot evaluate agreements among more than two readings on the same individuals, whereas the standard one-way ANOVA used to compute ICC (Sokal and Rohlf, 1995, p. 213) is readily generalized to any number of researchers.
3. Material and methods

The D. subobscura flies used in this study originated from a laboratory stock established from wild flies collected near Barce-
lona (41° 43′N, 2° 13′E) in October 2007. The stock was maintained in bottle culture at 18 °C (12:12 light/dark cycle). Heat tolerance was measured on 120 flies individually placed in capped 5-ml glass vials and randomly allocated in two 4-flies ×15 columns (columns) racks. Each rack was immersed in a 45 × 35 × 35 cm3 Plexiglas tank filled with ~40 L of water. The temperature of the water in the tank was controlled by a programmable heating unit (JULABO ED, JULABO Labortechnik GmbH, Seelbach, Germany) that also ensured proper water circulation. The start temperature for the dynamic assay was T0 = 24 °C. After an equilibration period of 10 min, temperature was increased at a rate ΔT = 0.6 °C min⁻¹. A thermocouple was placed in the water tanks and the flies were video-recorded during the assays with a digital HD video camera (SONY HDR-CX110E, Tokyo, Japan).

To assess inter-researcher reliability we first asked two coau-
thors (R1 and R2) to estimate knockdown times from the video-
recorded files. The coauthors did not know each other’s identity and were instructed to estimate knockdown times as the time at which flies lost righting ability and ceased to move. Knockdown time provides a comparable estimate to knockdown temperature because they are co-linear in the dynamic method (see Rezende et al., 2011). Five weeks after the first measurements we asked R1 to estimate knockdown times again to assess intra-researcher reliability or test-retest reliability.

4. Results and discussion

4.1. Inter- and intra-researcher reliability or agreement

Table 1 (flies 1–60) and Table 2 (flies 61–120) give the knockdown times estimated by R1 (at times t₁ and t₂) and R2 (fly 113 was removed from the data set because it was apparently damaged during handling). The results are plotted in Fig. 1(a) (inter-researcher reliability; R1 at time t₁ vs. R2) and (b) (intra-
researcher reliability; R1 at time t₁ vs. R1 at time t₂). Mean knockdown times estimated by R1 were on average higher than those estimated by R2 (Var(yinterR1) = 21.26 min, Var(yinterR2) = 20.68 min), and the difference was statistically significant (F1,236 = 10.28, P = 0.002). This implies that the inter-researcher ICC will capture this difference as imprecision and will be lower than the Pearson correlation coefficient (see Table 3), which ignores the inter-
researcher accurateness component. For this reason, the use of CCC is preferable here. On the other hand, R1 estimated approxi-
mately the same average knockdown time in the two records (Var(yintraR1) = 21.54 min, F1,236 = 1.75, P = 0.187). No heterogeneity of variances was detected between inter-Var(yinterR1) < 10 min², Var(yinterR2) = 1.81 min²; Levene’s F1,236 = 0.50, P = 0.481; and intra-
researcher Var(yintraR1) = 2.89 min²; Levene’s F1,236 = 0.26, P = 0.611) measurements.

Agreement statistics together with the 95% confidence limits are given in Table 3 (equations employed to calculate confidence intervals for ICC and CCC are provided in Appendix A). Inter-researcher ME (estimated as 1-ICC) was 34%, which is not substantially higher than the intra-researcher ME (29%). Importantly, inter- and intra-researcher precision was essentially the same (ρ = 0.73) and the main difference between researchers was in the accurateness component, which reflects the difference between the average knockdown times estimated by both researchers (location shift relative to scale: v = –0.4162).

To sum up, our analyses suggest that the imprecision of heat tolerance estimates is substantial and will be about the same no
provides an index of reliability for a single measurement. The intraclass correlation coefficient (ICC) is the coefficient that on each fly, where ICC is the intraclass correlation coefficient that ME? It is evident that measurements should take

4.2. Obtaining more accurate estimates

Suppose the first researcher (R1) is worried by the low precision of his/her estimates of heat tolerance and wants to increase measurement ‘accuracy’ (i.e., to obtain better estimates of the ‘true’ knockdown times). The immediate question is: how many measurements k should R1 take from each individual to minimize ME? It is evident that \( \delta_i \) in Eq. (2) will be better estimated the higher is the number of measurements on each individual, but a proper determination of k is important to avoid a wasteful use of resources or time. An answer to this question is provided by recurring to the so-called ‘Spearman—Brown prophecy formula’ (Lord and Novick, 1968; Winer, 1971; Hand, 2004), which allows obtaining the reliability of the mean of k replicated measurements as follows:

\[
\mu_k = \frac{k \times ICC}{(k-1)ICC+1}
\]  

Eq. (7) is an adjusted upward for the number of measurements on each fly, where ICC is the intraclass correlation coefficient that provides an index of reliability for a single measurement.

More specifically, assume R1 wants measurement error to be ME \( \leq 0.10 \) (i.e., \( \rho_k \geq 0.90 \)). Solving Eq. (7) for k we have:

\[
k \geq \frac{0.90(1-ICC)}{ICC(1-0.90)}
\]  

As ICC=0.71 (intra-researcher agreement in Table 3), R1 should take \( k=4 \) measurements for each fly. If R1 adopts this strategy he/she can report that the reliability of heat tolerance (estimated as the average of \( k=4 \) measurements) is \( \rho_k=0.91 \). However, it will always be prudent to conduct a reliability analysis to verify that this is indeed the case (Lachin, 2004).

Although increasing the number of measurement per fly is an obvious approach to improve reliability, it is always convenient to check for outliers due to erroneous measurements or data entry mistakes (see text for further details).

5. Conclusions

Assessing measuring reliability is important because inferences from statistical analysis and estimates of underlying causal variance components depend directly on the reliability coefficient. For instance, if narrow sense heritability is estimated from

![Fig. 1. Comparison of knockdown times (min) estimated by (a) R1 at time \( t_1 \) vs. R2 to assess inter-researcher reliability or agreement and (b) R1 at time \( t_1 \) vs. R1 at time \( t_2 \) to assess intra-researcher (test-retest) reliability. The dotted line in each plot indicates equality (perfect agreement) between estimates of knockdown times.](image1)

![Fig. 2. Bland and Altman (1986) agreement test to assess 95% limits of agreement (mean \( \pm 2SD \)). The test is applied to the repeated measurements taken by R1 at two time intervals, and the labels refer to those flies in Tables 1 and 2 that appear as outliers due to erroneous measurements or data entry mistakes (see text for further details).](image2)
the regression of offspring on parents the least square regression coefficient would be attenuated toward zero as a function of the reliability by which the parental phenotype is measured (Fuller, 1987). The issue of reliability is even more pressing when dealing with physiological or behavioral traits that apparently have low repeatabilities (Wolak et al., 2011). The variance resulting from ME will render estimates of repeatability that are also downwardly-biased (Lynch and Walsh, 1998) and the question is to what extent these traits have an inherently low repeatability or are measured with a higher ME than morphological characters. Furthermore, even if there was a perfect measurement procedure that estimates the ‘true’ value of the trait we want to measure, it can be readily shown that measurement validity cannot be higher than measurement reliability (Lachin, 2004). Therefore, estimates that are unreliable are also invalid.

A search in ISI Web of Science using as a search criterion several combinations of the key words ‘measurement error’, ‘reliability’, ‘thermotolerance’, ‘heat tolerance’, ‘heat resistance’, ‘cold tolerance’, ‘cold resistance’ or ‘cold hardiness’ only retrieved one relevant paper that reports repeatability of heat tolerance in Drosophila buzzatii but not measurement reliability (Krebs and Loeschcke, 1997). Thus, it seems that no single paper to date has computed the reliability or measurement error of thermotolerance estimates (for estimates of measurement error on other traits, see Wolak et al., 2011). Our results have demonstrated that the estimation of knockdown time has poor reliability even when the investigators are provided with digital video recordings of assayed individuals, which could be expected to be more precise than the direct ‘in vivo testing’. Some readers might object that the low reliability reported here can be due to a poor training of the people involved in assessing knockdown times. This is not the case. We unveal that RI in Tables 1 and 2 is L. E. Castrañeda, who is a well-trained ecophysiologist that has published several papers working with aphids, crickets, isopods, tadpoles, swans and flies (http://www. researcherg.com/rid/G-5340-2011). It seems safe to conclude that measurement error in studies of thermotolerance as performed here is in the order of 30% for D. subobscura (and possibly for other Drosophila and insects measured with a similar procedure). Note however that %ME depends on the relative size of the variance of ‘true’ knockdown times (see Table 3).

Finally, for laboratory measurements it seems that a good practice would be to have a reliability coefficient of 0.90 or higher. If the reliability of single thermotolerance estimates is low, then sample size requirements will need to be larger to assess meaningful differences between groups because statistical power decreases with decreasing reliability (figure 3 in Lachin (2004)). We suggest that it should be common practice to include an accuracy or measurement error in studies of thermotolerance as performed here is in the order of 30% for D. subobscura (and possibly for other Drosophila and insects measured with a similar procedure). Note however that %ME depends on the relative size of the variance of ‘true’ knockdown times (see Table 3).

### Table 3

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Estimate</th>
<th>95% confidence limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-researcher agreement</td>
<td>ICC</td>
<td>0.6614</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.6732</td>
</tr>
<tr>
<td></td>
<td>accurateness $\hat{x}$</td>
<td>0.9179</td>
</tr>
<tr>
<td></td>
<td>precision $\rho$</td>
<td>0.7334</td>
</tr>
<tr>
<td>Intra-researcher agreement</td>
<td>ICC</td>
<td>0.7095</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.7096</td>
</tr>
<tr>
<td></td>
<td>accurateness $\hat{x}$</td>
<td>0.9732</td>
</tr>
<tr>
<td></td>
<td>precision $\rho$</td>
<td>0.7291</td>
</tr>
</tbody>
</table>

despite the fact that measurement reliability could be expected to be more precise than the direct ‘in vivo testing’. Some readers might object that the low reliability reported here can be due to a poor training of the people involved in assessing knockdown times. This is not the case. We unveil that RI in Tables 1 and 2 is L. E. Castrañeda, who is a well-trained ecophysiologist that has published several papers working with aphids, crickets, isopods, tadpoles, swans and flies (http://www. researcherg.com/rid/G-5340-2011). It seems safe to conclude that measurement error in studies of thermotolerance as performed here is in the order of 30% for D. subobscura (and possibly for other Drosophila and insects measured with a similar procedure). Note however that %ME depends on the relative size of the variance of ‘true’ knockdown times (see Table 3).

### Appendix A. Confident limits for reliability coefficients

In the special case of duplicate measurements considered here the variance of Fisher’s (1946) intraclass correlation coefficient ICC is estimated as follows (Lachin, 2004):

$$\text{Var}(\text{ICC}) = \frac{1 - \text{ICC}^2}{2n} + \frac{(1 - \text{ICC}^2)^2}{n-1}$$

(A.1)

where $n$ is the sample size. From Eq. (A.1) confidence intervals (ICC$\text{ICC}_L$, ICC$\text{ICC}_U$) can be estimated as follows (Lachin, 2004; a SAS macro is provided by Lu and Shara (2007)):

$$\begin{align*}
\text{ICC}_L &= Z_1 - Z_{1-\alpha/2} \sqrt{\text{Var}(\text{ICC})} \\
\text{ICC}_U &= Z_1 + Z_{1-\alpha/2} \sqrt{\text{Var}(\text{ICC})}
\end{align*}$$

(A.2)

where $Z_{1-\alpha/2}$ is the point on a standard normal distribution exceeded with probability $\alpha/2$ (e.g., $Z_{1-0.05} = 1.96$ for $\alpha = 0.05$). The variance of the concordance correlation coefficient (CCC) is as follows (Lin, 1989):

$$\text{Var}(\text{CCC}) = \frac{1}{n-2} \left( \frac{1 - \rho^2}{\rho^2} \text{CCC}^2 (1 - \text{CCC}^2) + 4\text{CCC}^2 (1 - \text{CCC}^2) \right)$$

(A.3)

where $\rho$ is the Pearson correlation coefficient. Confidence intervals for CCC can also be approximated from Eq. (A.2) after replacing ICC by CCC, and Var(ICC) by Var(CCC).

### References


