doi:10.1111/j.1558-5646.2011.01565.x

COMPARING EVOLVABILITIES: COMMON ERRORS SURROUNDING THE CALCULATION AND USE OF COEFFICIENTS OF ADDITIVE GENETIC VARIATION

Francisco Garcia-Gonzalez,[1,2] Leigh W. Simmons,[1] Joseph L. Tomkins,[1] Janne S. Kotiaho[3] and Jonathan P. Evans[1]

[1]*Centre for Evolutionary Biology, School of Animal Biology, The University of Western Australia, Nedlands, WA 6009, Australia*

[2]*E-mail: paco.garcia@uwa.edu.au*

[3]*Centre of Excellence in Evolutionary Research, University of Jyväskylä, Jyväskylä, FI-40014, Finland*

Received August 12, 2011
Accepted December 20, 2011

In 1992, David Houle showed that measures of additive genetic variation standardized by the trait mean, CV_A (the coefficient of additive genetic variation) and its square (I_A), are suitable measures of evolvability. CV_A has been used widely to compare patterns of genetic variation. However, the use of CV_As for comparative purposes relies critically on the correct calculation of this parameter. We reviewed a sample of quantitative genetic studies, focusing on sire models, and found that 45% of studies use incorrect methods for calculating CV_A and that practices that render these coefficients meaningless are frequent. This may have important consequences for conclusions drawn from comparative studies. Our results are suggestive of a broader problem because miscalculation of the additive genetic variance from a sire model is prevalent among the studies sampled, implying that other important quantitative genetic parameters might also often be estimated incorrectly. We discuss the most prominent issues affecting the use of CV_A and I_A, including scale effects, data transformation, and the comparison of traits with different dimensions. Our aim is to increase awareness of the potential mistakes surrounding the calculation and use of evolvabilities, and to compile general guidelines for calculating, reporting, and interpreting these useful measures in future studies.

KEY WORDS: CV_A, evolvability, fitness, heritability, I_A, natural selection, quantitative genetics, sexual selection.

The ability of populations to respond to natural or sexual selection, termed "evolvability" in quantitative genetics (Houle 1992; Lynch and Walsh 1998; Hansen 2006; Sniegowski and Murphy 2006; Hansen and Houle 2008; Pigliucci 2008), is contingent on the level of additive genetic variation underlying trait expression. Consequently, a common practice in quantitative genetics studies is to derive standardized measures of evolvability that allow comparisons among traits and taxa. In a landmark paper, Houle (1992) proposed that a dimensionless statistic (see also Charlesworth 1987), termed the coefficient of additive genetic variation (CV_A), is appropriate for such purposes. CV_A is simply

$$CV_A = \frac{\sqrt{V_A}}{\bar{X}}, \quad (1)$$

that is, the square root of the additive genetic variance (V_A) divided by the phenotypic mean of the trait (note that Houle (1992) expressed this quantity using a 100 multiplier). Unlike heritability, CV_A is a measure of additive genetic variation that is standardized

EVOLUTION INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

by the trait mean and therefore independent of other sources of variance. It is precisely these properties that make CV_A suitable for comparative purposes.

Houle (1992) addressed a long-standing difficulty with the interpretation of patterns of genetic variation in fitness traits. Traits closely related to fitness, such as survival or fecundity, typically exhibit lower narrow-sense heritabilities (i.e., the ratio of additive genetic variance to total phenotypic variance) than traits under weak or stabilizing selection, such as morphological traits (Gustafsson 1986; Charlesworth 1987; Mousseau and Roff 1987; Roff and Mousseau 1987; Houle 1992; Falconer and Mackay 1996; Kruuk et al. 2000; Merila and Sheldon 2000). This pattern was traditionally interpreted as resulting from the depletion of genetic variation due to strong directional selection (Fisher 1930). By contrast, Houle (1992) showed that traits closely associated with fitness generally exhibit higher CV_As, and thus higher, not lower additive genetic variability than those under weaker selection. Houle's (1992) data supported the view that traits closely associated with fitness have higher levels of residual variation (e.g., nonadditive genetic, maternal, and environmental variation, including error variation), thereby explaining their low heritabilities (Barton and Turelli 1989; Price and Schluter 1991).

Since the publication of Houle's (1992) paper, there has been a proliferation of studies reporting either mean-scaled additive genetic variances (predominantly CV_A but see below) or using these evolvability measures for comparisons, and this work has improved our understanding of the factors that contribute toward the maintenance of genetic variation. This work has led to the general consensus that traits tightly linked to fitness exhibit high levels of both genetic and residual variation (e.g., Merila and Sheldon 1999, 2000; Kruuk et al. 2000; McCleery et al. 2004; Coltman et al. 2005; Hansen et al. 2011). Furthermore, the realization that some traits harbor considerable additive genetic variance despite strong directional selection has fuelled the development of new theory, such as the genic-capture model used to address the lek paradox (Kotiaho et al. 2008), and more generally the maintenance of genetic variation in fitness traits (Rowe and Houle 1996; Tomkins et al. 2004).

Given the utility of CV_A for comparative studies of genetic variation, it is crucial that primary studies employ correct and consistent methods to estimate this parameter (or suitable alternatives; see below and Discussion). If mistakes are frequent, incorrectly calculated CV_As are likely to have been reported in reviews or studies that compile or compare these values, thus potentially biasing and/or confounding the conclusions drawn from such studies. In an attempt to determine the extent to which mistakes in the calculation of CV_A occur in the literature, and their potential consequences, we have reviewed recent quantitative genetic studies that have reported this statistic. We also review important issues in relation to the use and limitations of CV_A.

In his original paper, Houle (1992) also described another mean-standardized additive genetic variance, termed I_A, as a measure of evolvability (Houle 1992; Hansen et al. 2011). I_A equals ${CV_A}^2$ if CV_A is expressed as in equation (1)

$$I_A = \frac{V_A}{\bar{X}^2}. \quad (2)$$

Although CV_A and I_A are related, they are distinct quantities (Houle 1992), and a key advantage of I_A is that its numerical value can be interpreted as the expected proportional change under a unit strength of selection (see Hansen et al. 2003; Hereford et al. 2004; Hansen et al. 2011). For this reason, Hansen et al. (2011) recommend the use of I_A as a measure of evolvability. It is therefore likely that future research will shift focus from CV_A to I_A. Our review, however, focuses on CV_A because until now this coefficient has been used predominantly to report and compare evolvabilities. Nevertheless, given the relationship between CV_A and I_A and the fact that both involve mean scaling, both measures suffer from similar limitations and are prone to similar calculation errors. Therefore, our results and guidelines can be extended to both measures of evolvability.

Methods

We used the Web of Science to identify all articles published between 2000 and 2010 (11 years) that cited Houle (1992). Specifically, we selected papers appearing in the top 40 journals ranked according to impact factors (Journal Citation Reports 2009) within each of the following four areas: Evolutionary Biology, Genetics and Heredity, Multidisciplinary Sciences, and Biology. These filters limited the results to 346 papers published in 29 journals. During the process, we noticed that a high number of papers citing Houle (1992) were published in the journal *Genetica* ($n = 18$ between 2000 and 2010); we therefore also included this journal in our Web of Science search. Including papers from this journal does not change results. Thus, the total number of papers covered in our initial screening was 364.

We classified studies by quantitative genetic design. We decided a priori to focus exclusively on studies employing nested full-sib half-sib designs (Lynch and Walsh 1998; termed half-sib designs after Roff 1997) because this is the single most common quantitative genetic design and thus provides a simple limit to the breadth of the literature review. Our study is therefore based on a sample of quantitative genetic studies, and assumes that this sampling yields a nonbiased picture of the use of CV_A in general.

Of the 364 papers screened, 49 were empirical studies using nested full-sib half-sib designs and 38 of these reported coefficients of additive genetic variation. From these 38 papers, we recorded whether the study provided sufficient information to

calculate CV_A. This information is, as recommended by Houle (1992), the phenotypic mean of the trait ($\bar{X}$), and either the sire variance component (V_{sire}) or the additive genetic variance, V_A, which is four times V_{sire} for a full-sib half-sib design on a diploid organism (Becker 1984; Falconer and Mackay 1996; Roff 1997; Lynch and Walsh 1998). In the absence of these statistics, a study can report other parameters (which we refer to as "cryptic" information) that can be used to calculate CV_As:

(1) The output of the analysis of variance. In these models, the mean squares (MS) and degrees of freedom are informative, because V_{sire} and therefore V_A can be calculated from such data, provided that the output refers to analyses using untransformed data (CV_As have little relevance if they are calculated using transformed data; see Discussion).

(2) CV_As can also be calculated when narrow sense heritability (h^2) and the mean of the trait are provided together with a measure of dispersion such as the standard deviation (SD) or the variance. This is because, in general, V_A can be calculated as $V_A = h^2 \times V_P$, where V_P is the total phenotypic variance. Caution is needed, however, because h^2 estimates can be dependent on the structure of the quantitative genetic model used to infer them; the inclusion of fixed effects in the model reduces the phenotypic variance that is partitioned thereby increasing the heritability value (Wilson 2008).

(3) Likewise, if the standard error (SE) and the sample size for the phenotypic measurements are given, V_P can be calculated and from here V_A can be inferred from the formula above. In some cases, h^2 might not exactly correspond to $\frac{V_A}{V_P}$, for instance when dealing with threshold traits, and so caution needs to be taken when inferring CV_As from V_A values obtained from V_P and heritabilities.

(4) V_A can be calculated if the coefficient of residual variation is provided

$$CV_R = \frac{\sqrt{V_R}}{\bar{X}}, \tag{3}$$

on the assumption that $V_R = V_P - V_A$.

(5) Finally, if a study uses log-transformation of data, it is still possible to calculate CV_A on the untransformed scale. This is because the additive genetic variance calculated on the log-transformed scale is an estimate of I_A for the trait on the original scale as long as $I_A << 1$ (Hansen et al. 2011). CV_A (as in eq. 1) can be then calculated as the square root of I_A.

When possible, we calculated CV_A and contrasted this value with the reported CV_A. If the study reported genetic parameters for several traits, we focused on the first trait reported in the paper, unless the calculation of CV_A for this trait was not straightforward (e.g., if V_A was zero, or if the trait had zero mean or had been otherwise transformed).

Results

Of the 38 studies reporting CV_A that we scrutinized, nearly half (44.7% or 17 studies) miscalculated CV_A, and 36.8% of studies lacked information on either V_{sire} or V_A. All results can be extracted from Table S1. Error rates were not higher in the studies for which "cryptic" information (see Methods) was used (43.75% of studies that provided the mean and V_{sire} or V_A were incorrect; 44.75% of studies were wrong among those that failed to provide these parameters).

The number of studies miscalculating CV_A is probably an underestimation for several reasons. First, for some studies categorized as calculating CV_A correctly, we had to assume that the reported V_{sire} or V_A was correct. Second, we avoided looking at traits for which it was obvious that residuals were used (see Discussion). Third, some mistakes were only obvious after reanalyzing the original data of those studies (we were able to do this in papers authored by us; see Table S1). Thus, it is possible that similar mistakes would be uncovered in other studies assumed to be correct, if reanalysis of the raw data in such studies was possible.

Our review reveals a general lack of consistency surrounding the calculation and reporting of evolvabilities, the most common being the incorrect use of V_{sire} rather than V_A (which is four times V_{sire} for a full-sib half-sib design on a diploid organism) in the calculation of CV_A (18.4% of studies reviewed reporting CV_As, or 41.2% of the studies with incorrect CV_A), which underestimates the actual CV_A by half (and will also result in a fourfold reduction in the heritability; Fig. 1). A further source of error revealed in 10.5% of the studies (23.5% of studies with incorrect CV_A) was the use of the square root of the ratio of additive genetic variance to the trait mean, instead of the square root of the additive genetic variance. This mistake may lead to significant overestimation of CV_A (see Fig. 1), sometimes by orders of magnitude. Problems with transformation of variables, which undermine the utility of CV_A for comparative purposes, were detected in 13.2% of studies (29.4% of studies with incorrect CV_A), but this is an underestimation of this problem. Undetermined errors were found in 13.2% of studies (29.4% of studies with incorrect CV_A).

Finally, although we have here focused on CV_As, it is important to note that we encountered similar errors (e.g., using V_{sire} instead of V_A) in the reporting of evolvabilities as I_A.

Discussion

THE PROBLEM AND ITS CONSEQUENCES

Comparing evolvabilities within or among populations or species requires standardized measures of additive genetic variation. CV_A

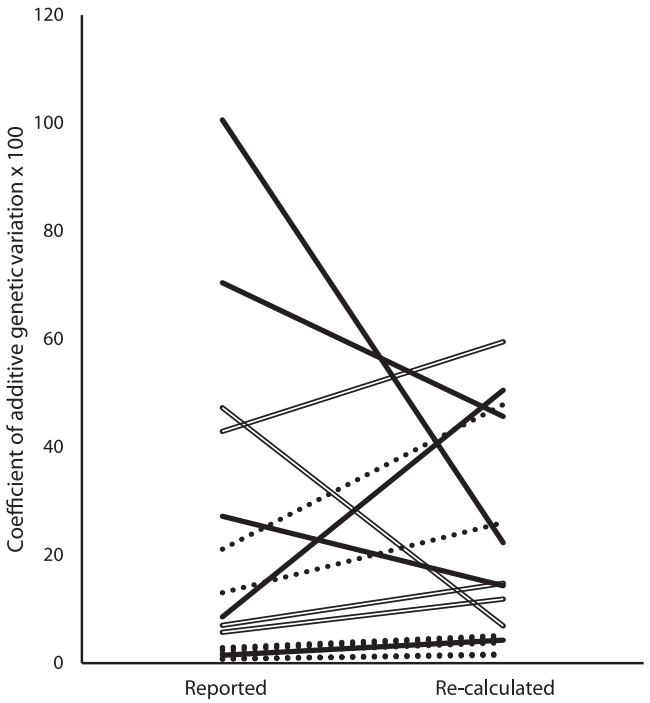


Figure 1. Plot of reported and recalculated CV_A values showing that the extent of over- and underestimation of these coefficients can be substantial in some cases. Cases where the mistake is the incorrect use of V_{sire} instead of V_A in the calculation of CV_A are indicated by dotted lines (note that some lines are overlapping at the bottom of the graph). Cases presenting the "square root" problem (see Results) are indicated with thick lines. The rest of the cases, where the source of the error is unknown, are indicated with open lines. All CV_A values shown are expressed using a 100 multiplier.

and I_A, which standardize V_A by the mean of the trait, are suitable measures of evolvability that do not suffer from the problems arising when using variance-standardized alternatives such as the heritability (Houle 1992; Hansen et al. 2011). Heritability is not a suitable measure of evolvability for comparative purposes because, among other reasons, it standardizes V_A by V_P, which not only contains the former but also comprises other components of variance that can correlate with V_A (Hansen et al. 2011).

Coefficients of additive genetic variation have been reported, or interpreted, widely but our review highlights two areas in need of attention. The first surrounds the calculation of the parameter itself. We found that almost half of studies that reported CV_As had calculated these parameters incorrectly. The most common error involved taking the sire variance component V_{sire} as V_A, thereby underestimating the actual CV_A by half in a full-sib half-sib analysis of diploid organisms. We suspect that this mistake arises because of a misunderstanding of the V_A notation in Houle (1992), or because of misunderstandings regarding quantitative genetic models. Thus, the high frequency of this mistake may reflect a more serious and broader problem affecting the accuracy of other quantitative genetic parameters (including h^2, which depend on V_A) in the literature. We are also aware of the existence of typographic errors in the formula for the calculation of coefficients of genetic variation in some papers (e.g., mutational coefficient of variation in Houle 1998), which may lead to further accumulation of errors. We reiterate that these coefficients should be calculated as in equation (1) where V_A stands for additive genetic variation (but may be replaced by any other desired variance component; e.g., V_R provides CV_R).

The second issue surrounds the quality or quantity of information reported in papers, which in turn are needed to determine whether CV_A and I_A can be calculated, or assessed for accuracy (i.e., whether authors report phenotypic means, and additive genetic variance or phenotypic variances, along with clearly defined methods for how data were handled—e.g., transformations, scales, etc.). We found that a high proportion of studies that report CV_As did not provide sufficient information to assess whether these parameters were calculated accurately; approximately, 37% of the studies reviewed here did not provide V_A or V_{sire} parameters.

What are the consequences of calculating evolvabilities incorrectly? The magnitude of the errors vary to the extent that the actual CV_A value can be grossly over- or underestimated depending on the nature of the error (see Table S1 and Fig. 1). Our analysis suggests that nearly 50% of papers based on half-sib designs calculate CV_As incorrectly. Clearly, this is likely to severely compromise studies that use such estimates for comparative purposes.

Although several other alternatives have been suggested (Roff 1997; Lynch and Walsh 1998; Hereford et al. 2004; Teplitsky et al. 2009), it is generally accepted that CV_A and I_A are in most cases the most appropriate statistics for comparing evolvabilities (e.g., Kruuk et al. 2000; Hansen et al. 2003, 2011). However, researchers need to be aware of the limitations of these coefficients (Lande 1977; Roff 1997; Lynch and Walsh 1998; Teplitsky et al. 2009). In the following section, we comment on important aspects surrounding the interpretation and use of CV_A and I_A. From here onwards, we refer to evolvabilities exclusively as measures based on mean scaling of additive genetic variance, unless stated otherwise.

MEANINGFUL MEASURES OF EVOLVABILITY: A DIMENSIONLESS MEASURE INFLUENCED BY DIMENSIONALITY, AND ISSUES OF SCALE AND TRANSFORMATION

The property making CV_A and I_A suitable measures of evolvability is a consequence of scaling to the mean. The influence of the mean on these two measures is obvious, but often not fully appreciated. Figure 2 illustrates the effects of mean scaling upon

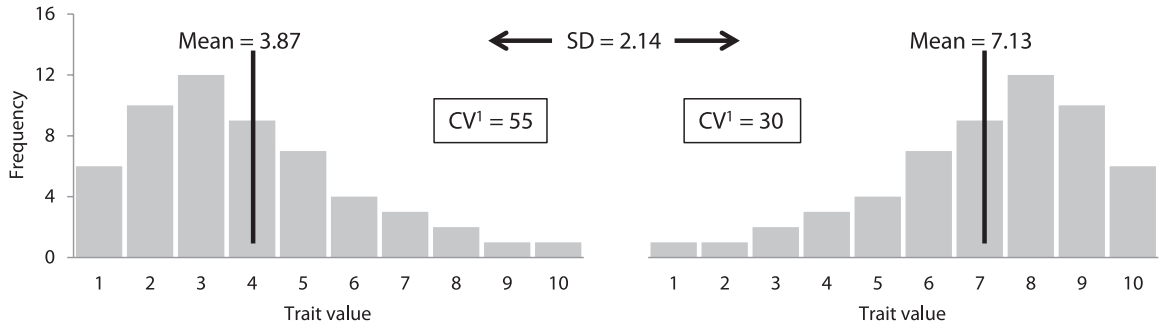


Figure 2. Effects of mean scaling. Two distributions with same variance (each distribution is the mirror image of the other; they have same skew but of different sign) are shown. Note the differences in the coefficient of variation, $CV = \frac{SD}{\bar{X}}$, of both distributions due to mean scaling effects. Note that these effects can seriously influence the numerical values of CV_A and to a higher degree of I_A, because the latter uses the square root of the mean. [1]*CV* shown has been multiplied by 100.

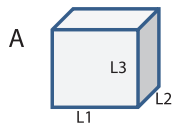
coefficients of variance with an example of two equal-variance different-skew distributions.

Houle (1992) pointed out the necessity to correct for scaling effects in comparative analysis of variability. Scaling effects deal with the relationship between means and variances, and although they may have biological relevance they can also be statistical artifacts. Thus, the interpretation surrounding mean-scaled additive genetic variances (CV_A and I_A) can be complicated unless the scaling effects are properly accounted for or eliminated. Where higher measurement errors are associated with small means (as one might expect), traits with smaller means will generally have comparatively higher coefficients of variance (or I_A; Houle 1992 and references therein). The same problem of a negative relationship between means and variances applies, regardless of measurement errors, when analyzing meristic traits (Lande 1977; Houle 1992; Lynch and Walsh 1998 pp. 302–305).

Houle (1992) also emphasized Lande's (1977) point that the relationship between means and variances is influenced by the covariance between the different components of a unit, and the extent to which these parts or components combine multiplicatively (e.g., length, area, and volume), or additively, to make up the whole. A well-known consequence of these purely mathematical influences on *CV*s is that, for instance, *CV*s of body volume would be larger than those of linear body measures (e.g., length, width, and height) as long as these linear traits are positively correlated (Lande 1977). To compare CV_As among traits with different dimensionalities, some authors scale downward the CV_As of areas and volumes by a factor of 2 and 3, respectively, based on the assumption that the relative magnitude of the *CV*s of linear, area, and volume measurements approximates 1, 2, 3, respectively (see Lande 1977). However, in most cases dividing *CV*s by their dimensionalities is not an adequate correction (but see Houle 1992), as it would only apply to more or less perfectly geometrically proportioned objects (i.e., where the correlation between the linear measurements is close to 1), and even in such cases it is only an approximation when dealing with objects that have low coefficients of variation for the linear dimensions (<10%; Lande 1977). In most cases of biological variation, where variation in shape is likely, the ratio 1:2:3 for linear, area, and volume dimensions needs to be considered an upper limit for the comparison of *CV*s of these different dimensions (Lande 1977), and not a rule. Thus, correcting for the effects of dimensionality, beyond acknowledging that due to mathematical constraints some *CV*s are expected to be higher than others, is not straightforward (see Fig. 3 for an illustration of this point; and see also Milner et al. 2000).

These considerations raise interesting questions in regards to the causes of variation in evolvabilities. For example, if fitness depends strongly on body size, fitness can be expected to have higher evolvability than linear morphological traits (e.g., leg length) for several reasons. First, fitness would be affected by a higher number of genetic events, which would capture higher levels of genetic variance than less polygenic traits (Houle 1992; Houle et al. 1996; Rowe and Houle 1996). Second, fitness would be determined by a higher number of dimensions, and of relationships among traits. The extent to which the evolvability of fitness is affected by these factors would depend on the sign and strength of the relationships between the components that make up fitness (Lande 1977; Price and Schluter 1991; and see Kirkpatrick 2009), something that in most cases is unknown.

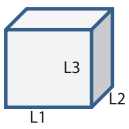
A crucial issue deserving attention relates to measurement scale and scale transformation. Only data on ratio and log-interval scales produce meaningful CV_A and I_A (Hansen et al. 2011; Houle et al. 2011). The key point here is to confirm that mean-standardized measures of additive genetic variation are based on data that have scales with meaningful zero values and that ratios of the data values are relevant. For example, length, mass, duration, or age measured in any unit of time meet these two requirements (a mass gain of 0 g means that there is no mass gain; a 20-cm long wing is twice as long as a 10-cm wing, etc.). Temperature in



Correlation

L1vL2	0.93
L2vL3	0.97
L1vL3	0.96

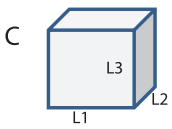
	Mean	SD	CV[1]	Scaled CV[1]
L1	10.0	1.0	10.3	
L2	10.0	1.1	11.2	
L3	10.0	1.1	10.8	
Area (L1xL2)	100.4	20.8	20.7	10.4
Volume	1022.6	314.4	30.7	10.2



Correlation

L1vL2	0.16
L2vL3	0.16
L1vL3	0.18

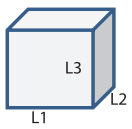
	Mean	SD	CV[1]	Scaled CV[1]
L1	10.0	1.0	9.9	
L2	10.0	1.0	9.5	
L3	10.0	1.0	9.6	
Area (L1xL2)	100.1	14.7	14.7	7.4
Volume	1002.1	193.4	19.3	6.4



Correlation

L1vL2	0.03
L2vL3	0.04
L1vL3	0.05

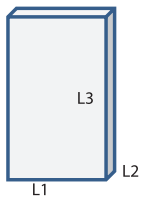
	Mean	SD	CV[1]	Scaled CV[1]
L1	10.0	1.0	10.2	
L2	10.0	1.0	9.9	
L3	10.0	1.0	10.2	
Area (L1xL2)	99.5	14.5	14.5	7.3
Volume	997.8	183.3	18.4	6.1



Correlation

L1vL2	0.00
L2vL3	0.01
L1vL3	0.00

	Mean	SD	CV[1]	Scaled CV[1]
L1	10.0	1.0	10.0	
L2	10.0	0.1	1.0	
L3	9.9	1.9	19.5	
Area (L1xL2)	100.0	10.0	10.1	5.0
Volume	984.9	217.7	22.1	7.4



Correlation

L1vL2	0.00
L2vL3	0.01
L1vL3	0.02

	Mean	SD	CV[1]	Scaled CV[1]
L1	10.1	1.0	10.1	
L2	5.0	1.0	19.6	
L3	20.1	1.0	4.8	
Area (L1xL2)	50.3	11.1	22.1	11.0
Volume	1008.2	229.7	22.8	7.6

Figure 3. **Illustration of the difficulties encountered when comparing coefficients of variation, $CV = \frac{SD}{\bar{X}}$, for different dimensions both at the within- and between-population level. Shown are the mean shape and associated means and variances for five populations (A–E) of objects, each with a sample size of 1000. Scaled *CV*s have been calculated by a factor of 2 (*CV* area) or 3 (*CV* volume). (A). A geometrically proportioned population of objects where the correlations among the linear measures of length (L1), width (L2), and height (L3) are near 1. This is the only case where the 1:2:3 scaling for length, surface, and volume's *CV*s is appropriate; scaled *CV*s for area and volume are a good approximation of the *CV*s for the linear measures that make up these dimensions. (B). Population with similar means and variances for L1–L3 as population A but where the correlations among L1-L2-L3 are low. Note that the scaling correction is not appropriate. (C–D). Populations where correlations among linear measurements are nearly zero. (C). Population with same means and variances for L1–L3 as A and B. Note that the scaling correction is not appropriate. (D). Population with same means for L1–L3 but with different variances for each of these linear measurements. (E). Population with dissimilar means but similar variances for the linear measurements. Note the difficulties for making comparisons both within- and between populations. [1]*CV* and scaled *CV* have been multiplied by 100.**

degrees Celsius or dates are counter examples: they do not have an absolute, nonarbitrary, zero value (a temperature of 0°C does not mean "no temperature"; Houle et al. 2011). Moreover, the ratio between their measurements has no meaning (5°C is not five times hotter than 1°C). Critically, transformations of variables that are in ratio or log-interval scales lead to different scales that no longer meet the criteria above (Hansen et al. 2011; Houle et al. 2011). Thus, CV_A and I_A have no meaning if they are calculated on transformed scales.

It is easy to see then that the problem of scale also applies to standardization of variables, to the use of scales yielding negative means, and to the use of residuals. CV_A and I_A cannot be calculated on variables with zero means (e.g., *z*-scores, principal components, and relative warps from geomorphic morphometric analyses) or residuals (see Kotiaho 1999; Birkhead et al. 2006 for discussion on residuals).

In summary, the utility of mean-scaled additive genetic variances as measures of evolvability in a broad range of situations is unquestionable. However, these parameters, as many others, are subject to errors of calculation, assumptions, and limitations, and all of these issues need to be accounted for when reporting and comparing evolvabilities. We encourage researchers to examine in detail the existing literature discussing the advantages, use, and limitations of evolvability measures, and in particular Houle's (1992), Lande's (1977), Roff's (1997), Lynch and Walsh's (1998), Hansen et al.'s (2011), and Houle et al.'s (2011) collective advice before undertaking a comparative study of evolvabilities.

Conclusions

While highlighting several common errors surrounding the calculation and use of CV_As, we also advocate some very simple guidelines for avoiding the miscalculation of evolvability estimates. The broad solution is that empiricists should strive for clarity and accuracy when reporting evolvabilities while supplementing these estimates with fundamental summary statistics on the raw scale,

and by explicitly noting the scales of measurement. Our findings also highlight the utility of depositing data files in public archives, such as the Dryad data repository (http://datadryad.org/). As for researchers using CV_A or I_A for review studies, our results suggest that the probability that the evolvability value in any given paper is incorrect may be considerable, and so they underscore the need to confirm the accuracy and validity of these statistics before they are included in any formal analysis.

We advocate that researchers adopt the following practices when reporting quantitative genetic data:

(1) Consistency in the calculation of CV_A, as in equation (1), and of I_A as in equation (2). Note that using a 100 multiplier when reporting CV_A is optional, although we recommend reporting CV_A as in equation (1), as there is no justified reason to use the 100 multiplier. Moreover, it is easier to relate CV_A without the 100 multiplier to I_A, which in turn translates directly into the standardized Lande (1979) equation (Hansen et al. 2003; Hereford et al. 2004; Hansen et al. 2011). In any case, we emphasize that studies need to state clearly whether CV_A is expressed with or without the 100 multiplier.

(2) CV_A and I_A must be calculated using the raw (untransformed) scale, and data need to be on ratio or log-interval scale (see above).

(3) Transparency in the reporting of methods, which should cover the inclusion of the formula used for the calculation of evolvabilities (important given some of the mistakes in the literature). Also critical, complete clarity on the use of transformations and clear descriptions of scale details and measurement units (including whether traits are measured in absolute units, proportions, or percentages), or otherwise, of any methods that may affect the estimation of variability.

(4) Reporting of all summary statistics necessary for the calculation of CV_A and I_A (Houle 1992). As a minimum, we suggest that the phenotypic mean and SD (or variance, or SE together with the sample size) is always provided (see also Wilson 2008) together with observational components of variation, causal components of additive genetic variance (V_A) and residual variance (V_R), narrow-sense heritability estimates (h^2), coefficients of additive genetic, residual, and phenotypic variation (CV_A, CV_R, CV_P) and I_A.

(5) In addition, we recommend, where possible, reporting the SEs of CV_A and I_A. SEs of these derived statistics would allow researchers to carry out unbiased meta-analyses of data on evolvabilities. An approximation to the sampling variance of a *CV* is given in Appendix 1 of Lynch and Walsh (1998; pp. 819–821); the SE of *CV* is simply the square root of such (large-sample) sampling variance (Lynch and Walsh 1998; p. 812; note that the SE of an estimate, the square root of the error variance of such estimate, is different from the SE of a sample, the sample's SD divided by the square root of the sample size). Based on the same procedure used by Lynch and Walsh (1998) to approximate the sampling variance of *CV* (The Delta Method, consisting of a Taylor series expansion), a normal approximation to $\sigma[CV_A]$, the SE of CV_A (measured as in eq. 1), assuming multivariate normality of the sampling error of V_A and the mean (μ) is

$$\sigma[CV_A] \approx \sqrt{\left(\frac{\sigma[V_A]}{2\mu\sqrt{V_A}}\right)^2 + \left(\frac{-\sigma[\mu]\sqrt{V_A}}{\mu^2}\right)^2 + 2r[V_A, \mu]\frac{-\sigma[\mu]\sigma[V_A]\sqrt{V_A}}{2\mu^3\sqrt{V_A}}}, \quad (4)$$

where $\sigma[V_A]$ denotes the SE of V_A, $\sigma[\mu]$ denotes the SE of the mean, and $r[V_A, \mu]$ is the error correlation of V_A and μ.

If $r[V_A, \mu]$ is lacking or not easily obtained, an approximation would be

$$\sigma[CV_A] \approx \sqrt{\left(\frac{\sigma[V_A]}{2\mu\sqrt{V_A}}\right)^2 + \left(\frac{-\sigma[\mu]\sqrt{V_A}}{\mu^2}\right)^2}. \quad (5)$$

Also, the sampling error of the mean will typically be modest. In fact, lacking $\sigma[\mu]$, an approximation is

$$\sigma[CV_A] \approx \frac{\sigma[V_A]}{2\mu\sqrt{V_A}}. \quad (6)$$

For I_A, the equations equivalent to equations (4)–(6) are

$$\sigma[I_A] \approx \sqrt{\left(\frac{\sigma[V_A]}{\mu^2}\right)^2 + \left(\frac{-2\sigma[\mu]}{\mu^3}\right)^2 + 2r[V_A, \mu]\frac{-2\sigma[\mu]\sigma[V_A]}{\mu^5}}, \quad (7)$$

$$\sigma[I_A] \approx \sqrt{\left(\frac{\sigma[V_A]}{\mu^2}\right)^2 + \left(\frac{-2\sigma[\mu]}{\mu^3}\right)^2}, \quad (8)$$

$$\sigma[I_A] \approx \frac{\sigma[V_A]}{\mu^2}. \quad (9)$$

The SE of V_A, $\sigma[V_A]$ (the square root of the large sample variance of V_A), can be obtained in most REML software for animal models. Also, Lynch and Walsh (1998) provide equations to calculate large sample variances of variance components for sib designs (see eq. 18.20b in page 561 and see also page 577). For nested full-sib half-sib mode, $\sigma[V_A]$ would be four times (twice in the case of full-sib models) the SE or the large sample variance of V_{sire}. The Delta Method assumes that the sampling errors of V_A and μ are multivariate normal, and additionally, that errors of CV_A and I_A are normal.

Markov chain Monte Carlo (MCMC)-based Bayesian analysis of quantitative genetic breeding experiments will also allow the calculation of the SEs of evolvabilities. Bootstrapping and jackknifing can also be considered as alternatives for some breeding designs. All these methods lead to appropriate characterization of uncertainty of CV_A and I_A estimates, and they would be of particular value in cases where the sampling error of the additive genetic variance does not distribute normally or symmetrically.

Adopting the five measures above will enable researchers to have a complete understanding of the meaning and utility of the genetic variance estimates reported in individual studies. They will allow the independent calculation of evolvabilities measured as CV_A or I_A, and will enable performing unbiased meta-analyses of these data. We anticipate that the adoption of these practices will broaden the scope and value of future investigations on variability in evolvabilities.

ACKNOWLEDGMENTS

We are very grateful to D. Houle, M. Morrissey, and an anonymous reviewer for useful suggestions that greatly improved the final version of the manuscript. We are most grateful to M. Morrissey for pointing out the approaches relating to the calculation of sampling errors and for discussion on their applications. We thank the Australian Research Council for financial support to FG-G, LWS, JLT, and JPE, and to the Academy of Finland's Centre of Excellence in Evolutionary Research for support to JSK.

LITERATURE CITED

Barton, N. H., and M. Turelli. 1989. Evolutionary quantitative genetics: how little do we know. Ann. Rev. Genet. 23:337–370.

Becker, W. A. 1984. Manual of quantitative genetics. Pullman, Washington, DC.

Birkhead, T. R., E. J. Pellatt, I. M. Matthews, N. J. Roddis, F. M. Hunter, F. McPhie, and H. Castillo-Juarez. 2006. Genic capture and the genetic basis of sexually selected traits in the zebra finch. Evolution 60:2389–2398.

Charlesworth, B. 1987. The heritability of fitness. Pp. 21–40 *in* J. W. Bradbury and M. B. Andersson, eds. Sexual selection: testing the alternatives. John Wiley & Sons Limited, New York.

Coltman, D. W., P. O'Donoghue, J. T. Hogg, and M. Festa-Bianchet. 2005. Selection and genetic (CO)variance in bighorn sheep. Evolution 59:1372–1382.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. Longman, Essex.

Fisher, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

Gustafsson, L. 1986. Lifetime reproductive success and heritability: empirical support for Fisher's fundamental theorem. Am. Nat. 128:761–764.

Hansen, T., C. Pélabon, and D. Houle. 2011. Heritability is not evolvability. Evol. Biol. 38:258–277.

Hansen, T. F. 2006. The evolution of genetic architecture. Annu. Rev. Ecol. Evol. Syst. 37:123–157.

Hansen, T. F., and D. Houle. 2008. Measuring and comparing evolvability and constraint in multivariate characters. J. Evol. Biol. 21:1201–1219.

Hansen, T. F., C. Pelabon, W. S. Armbruster, and M. L. Carlson. 2003. Evolvability and genetic constraint in Dalechampia blossoms: components of variance and measures of evolvability. J. Evol. Biol. 16:754–766.

Hereford, J., T. F. Hansen, and D. Houle. 2004. Comparing strengths of directional selection: how strong is strong? Evolution 58:2133–2143.

Houle, D. 1992. Comparing evolvability and variability of quantitative traits. Genetics 130:195–204.

———. 1998. How should we explain variation in the genetic variance of traits? Genetica 102/103:241–253.

Houle, D., B. Morikawa, and M. Lynch. 1996. Comparing mutational variabilities. Genetics 143:1467–1483.

Houle, D., C. Pelabon, G. P. Wagner, and T. F. Hansen. 2011. Measurement and meaning in biology. Quat. Rev. Biol. 86:3–34.

Journal Citation Reports®. 2009. Science Edition (Thomson Reuters, 2010).

Kirkpatrick, M. 2009. Patterns of quantitative genetic variation in multiple dimensions. Genetica 136:271–284.

Kotiaho, J. S. 1999. Estimating fitness: comparison of body condition indices revisited. Oikos 87:399–400.

Kotiaho, J. S., N. R. Lebas, M. Puurtinen, and J. L. Tomkins. 2008. On the resolution of the lek paradox. Trends Ecol. Evol. 23:1–3.

Kruuk, L. E. B., T. H. Clutton-Brock, J. Slate, J. M. Pemberton, S. Brotherstone, and F. E. Guinness. 2000. Heritability of fitness in a wild mammal population. Proc. Natl. Acad. Sci. USA 97:698–703.

Lande, R. 1977. On comparing coefficients of variation. Syst. Zool. 26:214–217.

———. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, MA.

McCleery, R. H., R. A. Pettifor, P. Armbruster, K. Meyer, B. C. Sheldon, and C. M. Perrins. 2004. Components of variance underlying fitness in a natural population of the great tit *Parus major*. Am. Nat. 164:E62–E72.

Merila, J., and B. C. Sheldon. 1999. Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. Heredity 83:103–109.

———. 2000. Lifetime reproductive success and heritability in nature. Am. Nat. 155:301–310.

Milner, J. M., J. M. Pemberton, S. Brotherstone, and S. D. Albon. 2000. Estimating variance components and heritabilities in the wild: a case study using the 'animal model' approach. J. Evol. Biol. 13:804–813.

Mousseau, T. A., and D. A. Roff. 1987. Natural-selection and the heritability of fitness components. Heredity 59:181–197.

Pigliucci, M. 2008. Is evolvability evolvable? Nat. Rev. Genet. 9:75–82.

Price, T., and D. Schluter. 1991. On the low heritability of life-history traits. Evolution 45:853–861.

Roff, D. A. 1997. Evolutionary quantitative genetics. Chapman and Hall, New York.

Roff, D. A., and T. A. Mousseau. 1987. Quantitative genetics and fitness: lessons from *Drosophila*. Heredity 58:103–118.

Rowe, L., and D. Houle. 1996. The lek paradox and the capture of genetic variance by condition dependent traits. Proc. R. Soc. Lond. B 263:1415–1421.

Sniegowski, P. D., and H. A. Murphy. 2006. Evolvability. Curr. Biol. 16:R831–R834.

Teplitsky, C., J. A. Mills, J. W. Yarrall, and J. Merila. 2009. Heritability of fitness components in a wild bird population. Evolution 63: 716–726.

Tomkins, J. L., J. Radwan, J. S. Kotiaho, and T. Tregenza. 2004. Genic capture and resolving the lek paradox. Trends Ecol. Evol. 19: 323–328.

Wilson, A. J. 2008. Why h(2) does not always equal V-A/V-P? J. Evol. Biol. 21:647–650.

Associate Editor: L. Kruuk

Supporting Information

The following supplementary material is available for this article:

Table S1. List of studies using nested full-sib half-sib designs providing CV_A data included in our search (see Methods), and information relative to these studies (see main article for description of errors).

Supporting Information may be found in the online version of this article.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.