2012

A robust new metric of phenotypic distance to estimate and compare multiple trait differences among populations

Rebecca Safran
University of Colorado, Boulder, rebecca.safran@colorado.edu

Samuel Flaxman
University of Colorado, Boulder, samuel.flaxman@colorado.edu

Michael Kopp
University of Vienna, Austria, michael.kopp@univie.ac.at

Darren E. Irwin
University of British Columbia, irwin@zoology.ubc.ca

Derek Briggs
University of Colorado, Boulder, derek.briggs@colorado.edu

See next page for additional authors

Current Zoology 58 (3): 426–439, 2012

A robust new metric of phenotypic distance to estimate and compare multiple trait differences among populations

Rebecca SAFRAN[1*§], Samuel FLAXMAN[1§], Michael KOPP[2§], Darren E. IRWIN[3], Derek BRIGGS[4], Matthew R. EVANS[5], W. Chris FUNK[6], David A. GRAY[7], Eileen A. HEBETS[8], Nathalie SEDDON[9], Elizabeth SCORDATO[10], Laurel B. SYMES[11], Joseph A. TOBIAS[9], David P. L. TOEWS[3], J. Albert C. UY[12]

[1] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder CO 80309, U.S.A.

[2] Mathematics and Biosciences Group, Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria; Current address: Evolutionary Biology and Modeling Group, Faculty of Sciences, Aix-Marseille University, 13331 Marseille, France.

[3] Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, BC, Canada, V6T 1Z4.

[4] Research & Evaluation Methodology, School of Education, University of Colorado, Boulder, CO 80309, U.S.A.

[5] School of Biological and Chemical Sciences, Queen Mary University of London, London, E1 4NS

[6] Department of Biology, Colorado State University, Fort Collins, CO 80523, U.S.A.

[7] Department of Biology, California State University, Northridge, CA 91330, U.S.A.

[8] School of Biological Sciences, University of Nebraska, Lincoln, NE 68588, U.S.A.

[9] Edward Grey Institute, Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK

[10] Committee on Evolutionary Biology, The University of Chicago, Chicago, IL, 60637,U.S.A.

[11] Department of Biology, Dartmouth College, Hanover, NH 03755, U.S.A.

[12] Department of Biology, University of Miami, Coral Gables, FL 33146, U.S.A.

Abstract Whereas a rich literature exists for estimating population genetic divergence, metrics of phenotypic trait divergence are lacking, particularly for comparing multiple traits among three or more populations. Here, we review and analyze via simulation Hedges' *g*, a widely used parametric estimate of effect size. Our analyses indicate that *g* is sensitive to a combination of unequal trait variances and unequal sample sizes among populations and to changes in the scale of measurement. We then go on to derive and explain a new, non-parametric distance measure, "Δ*p*", which is calculated based upon a joint cumulative distribution function (CDF) from all populations under study. More precisely, *distances are measured in terms of the percentiles in this CDF at which each population's median lies*. Δ*p* combines many desirable features of other distance metrics into a single metric; namely, compared to other metrics, *p* is relatively insensitive to unequal variances and sample sizes among the populations sampled. Furthermore, a key feature of Δ*p*—and our main motivation for developing it—is that it easily accommodates *simultaneous comparisons of any number of traits across any number of populations*. To exemplify its utility, we employ Δ*p* to address a question related to the role of sexual selection in speciation: are sexual signals more divergent than ecological traits in closely related taxa? Using traits of known function in closely related populations, we show that traits predictive of reproductive performance are, indeed, more divergent and more sexually dimorphic than traits related to ecological adaptation [*Current Zoology* 58 (3): 426–439, 2012].

Keywords Effect size, Phenotype divergence, Sexual dimorphism, Sexual selection, Speciation

Inferences about the role of adaptation in population differentiation and speciation are often made by comparing phenotypic divergence and population genetic divergence. An active area of research and debate concerns the role of sexual selection in the process of speciation (e.g., Lande, 1981; West-Eberhard, 1983; Price, 1998, Panhuis et al., 2001; Boul et al., 2007; Ritchie, 2008; van Doorn et al., 2009; Kraaijeveld et al., 2011). Whereas divergence in sexual traits is a common form of phenotypic differentiation among populations and sister taxa (e.g., Endler and Houde, 1995; Seehausen and van Alphen, 1999; Gray and Cade, 2000; Uy and Borgia, 2000; Irwin et al., 2001, 2008; Safran and McGraw; 2004, Rodríguez et al., 2004; Mendelson et al., 2005; Johnsen et al., 2006; Svensson et al., 2006; Boul et al., 2007; Uy et al., 2008; Seddon et al., 2008; Free-

Received Oct. 22, 2011; accepted Feb. 9, 2012.

∗ Corresponding author. E-mail: rebecca.safran@colorado.edu. [§]Equal contribution.

man-Gallant et al., 2009), many questions still remain about how sexual signal divergence is related to speciation. In particular, researchers are interested in estimating differences in the extent of trait divergence that is underlain by ecological adaptation or sexual selection as a way to examine mechanisms that maintain modern population differences. In turn, such analyses can be used to infer a role of either natural or sexual selection in the process of divergence (Mayr, 1947; Maan and Seehausen, 2011).

An issue underlying all research concerning divergence among closely related populations concerns the metrics employed to examine estimates of both phenotypic and genetic distance. Whereas a rich and controversial literature exists for estimates of genetic distance (e.g. Wright, 1943, 1951, 1965, 1973, 1978; Slatkin 1987; Charlesworth, 1998; Excoffier, 2001; Charlesworth et al., 2003; Hedrick, 2005) there are relatively few resources for metrics of phenotypic distance, yet such metrics are fundamental for comparing trait differences among populations. In particular, the literature on metrics of phenotype distance, a sub-set of effect size metrics, (reviewed by Grissom and Kim, 2001; Nagawa and Cuthill, 2007) is focused on comparisons between two populations and on cases where traits follow the assumptions of parametric statistics (e.g., Grissom and Kim, 2001). Yet, limitations in the widespread utility of these metrics exist, particularly when trait distributions deviate from assumptions underlying parametric methods, when traits under study are measured in different units (e.g., size vs. color), when there are unequal sample sizes among groups under comparison, or when simultaneous analysis of more than one trait and/or more than two populations is desired.

Here, we offer a non-parametric and potentially powerful new metric, "Δp", that overcomes the aforementioned limitations. We analyze the performance of Δp by comparing its behavior to that of a very commonly employed effect size metric, Hedges' *g* (Hedges, 1981), which is a variant of Cohen's *d* (Cohen, 1969) and belongs to a class of parametric effect size measures that essentially calculate a difference in means, scaled (divided) by some measure of the standard deviation in one or both groups being compared (see Grissom and Kim, 2001; Nagawa and Cuthill, 2007). (In fact, *g* and *d* are practically identical metrics, with the only difference being that *d* does not utilize the "-2" correction seen below in the denominator of equation (2).) As we illustrate, the behavior of such metrics can be sensitive to unequal variances and sample sizes among groups being compared. We show that Δp does not have this sensitivity, that its behavior is at least as reliable as that of *g* for both normally and non-normally distributed data sets, and that it offers the crucial, additional advantage of being amenable to comparisons involving more than two populations and/or two or more traits simultaneously.

We note here that, like other distance measures, the measures we present below are descriptive. Hence, we suggest the following protocol: standard statistical methods are first used to establish the significance of differences between populations. Then, the method we present below can be used to generate quantitative descriptions of differences between two or more populations for examining questions about (1) the degree of phenotypic divergence of a single trait, (2) the overall degree of phenotypic differentiation across all traits being considered, (3) the degree of sexual dimorphism in a trait within populations (if applicable) relative to the phenotypic differences between populations, and (4) the ranking of traits, regardless of their units of measure, in order of which traits are most phenotypically divergent.

After explaining the newly derived phenotype distance metric, we illustrate its utility by applying it to a number of empirical data sets where the function of traits in either a sexual signaling or ecological adaptation context has been previously explored, such that we can compare trait distance between populations for both sexual and ecological traits. We also use this metric to make comparisons of males to females within closely related populations to test a prevailing yet largely untested assumption about using sexual dimorphism as a proxy of sexual selection (e.g., Kraaijeveld et al., 2011). Here, we can address whether known sexual signals - when they are present in both males and females - are more sexually dimorphic than ecological traits. The overall goal of this contribution is to present methodological recommendations, so that appropriate metrics of trait distance are available for making comparisons among closely related populations.

Limitations of Hedges' g In the following, we outline some of the limitations of Hedges' *g* that have motivated the development of our new metric Δp. Many of our points have also been made by Grissom and Kim (2001). We wish to emphasize that our paper is not meant as a general critique of Hedges' *g*, which has many useful properties: one only needs to know means, standard deviations, and sample sizes to calculate it, it is unit-less, and it has achieved widespread usage in a va-

riety of disciplines. Indeed, it is because of the widespread acceptance of g that we choose to use it for comparisons here: the behavior of g sets a standard that a newly proposed metric should meet and exceed.

Hedge's g is computed as

$$g_{y,zj} = \frac{\overline{x}_{yj} - \overline{x}_{zj}}{s^*_{y,zj}} \quad (1)$$

where the subscripts on g in equation (1) indicate that the calculation of g between two populations, denoted y and z, was for the jth trait that was measured in these populations. $\overline{x}_{yj}$ and $\overline{x}_{zj}$ are sample means for the jth trait in the two sampled populations (y and z, respectively) and $s^*_{y,zj}$ is a measure of pooled sample standard deviation. This measure is weighted by sample size and is defined as

$$S^*_{y,zj} \equiv \sqrt{\frac{(n_{yj}-1)S^2_{yj} + (n_{zj}-1)S^2_{zj}}{n_{yj}+n_{zj}-2}}, \quad (2)$$

where n_{yj} and n_{zj} are the sample sizes of observations of the jth trait in the two populations and S^2_{yj} and S^2_{zj} are the sample variances (see Table 1 for definitions of all symbols). Note that Hedges and Olkin (1985) give an additional correction factor for g, which should be applied if the overall sample size is small.

As mentioned above, $g_{y,zj}$ has properties that limit its utility in certain situations.

First, it assumes that the trait has the same "true" variance in both populations. Indeed, the term under the square-root sign in equation (2) is an unbiased estimator for this variance (as it is in the two-sample t-test with equal variance, e.g., Sokal and Rohlf 1995). If, in contrast, the (true) variances in the two populations are different, equation (1) cannot be applied, since the denominator (eq. 2) has no useful interpretation and its expected value will depend on sample sizes. The latter point can be seen by replacing the empirical variances in equation (2) with their "true" counterparts. Then, increasing the sample size for the population with the larger (smaller) true variance will increase (decrease) $s^*_{y,zj}$ and decrease (increase) $g_{y,zj}$. This effect is also illustrated in Fig. 1.

A possible solution to the problem of variance heterogeneity is to define an alternative distance measure in which equation (2) is replaced by the square root of an unweighted average of the sample variances (in analogy to the t-test with unequal variances). However, the interpretation of such a measure poses conceptual difficulties, since the difference in sample means is scaled by a "virtual" standard deviation that does not apply to any real population (Grissom and Kim, 2001).

Table 1 Definitions and explanations of notation employed

Symbol	Meaning	Value(s) or range assigned (if applicable)
N	Number of populations or groups being compared	Integer, ≥ 2
t	Number of traits measured in each population or group	Integer, ≥ 1
i	Index variable for populations	i = 1, 2, …, N
j	Index variable for traits	j = 1, 2, …, t
n_{ij}	Number of observations of jth trait in ith population	Integer, > 0
k	Index variable for observations	k = 1, 2, …, n_{ij}
x_{ijk}	kth observation of jth trait in ith population	Empirically determined
$\overline{x}_{ij}$	Sample mean of jth trait in ith population	Empirically determined
$\hat{x}_{ij}$	Sample median of jth trait in ith population	Empirically determined
s^2_{ij}	Sample variance of jth trait in ith population	Empirically determined
$g_{y,zj}$	Hedges' g statistic computed for the jth trait measured in populations y and z ($y, z \in \{1,2,...,N\}$)	See equation (1)
$s^*_{y,zj}$	Pooled standard deviation used in calculation of Hedges' g statistic	See equation (2)
d	Cohen's d statistic	See text for description
$p_j(u)$	cumulative distribution function for trait j, expressed as a percentage	See equation (3)
$\Delta p_{y,z\bullet}$	Distance between populations y and z, calculated over all traits	See equation (5)
$\Delta p_{y,zj}$	Distance between populations y and z, calculated for trait j	See equation (4)

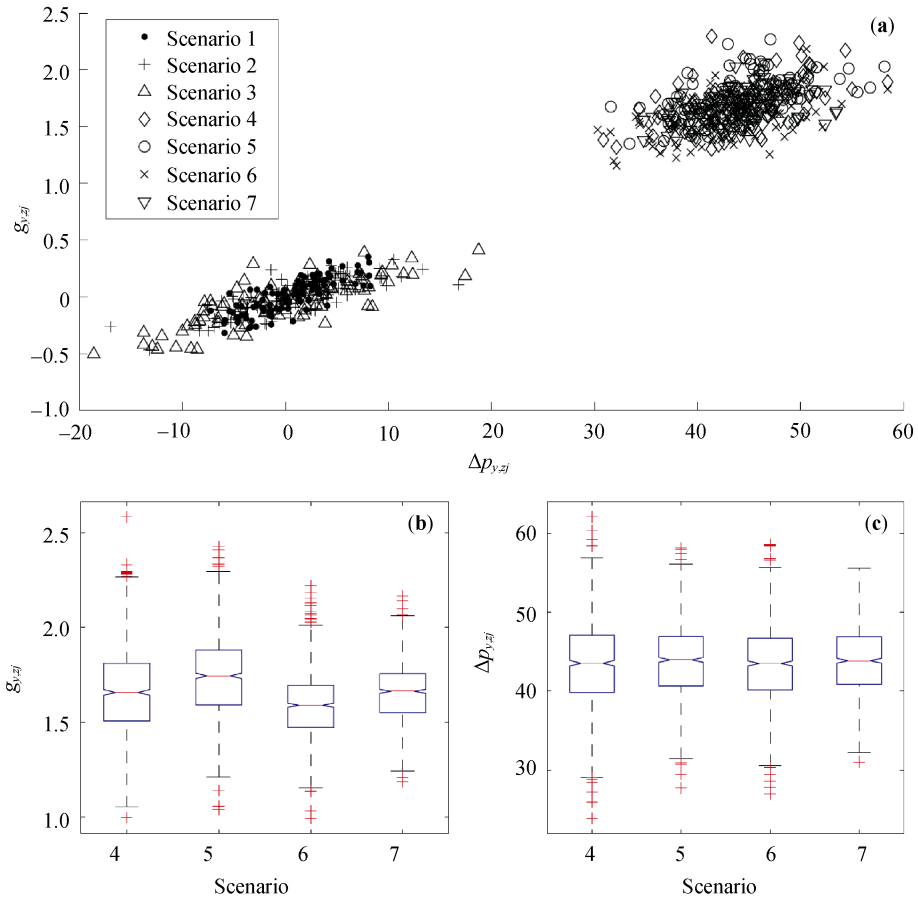

Fig. 1 Comparisons between the behavior of $g_{y,zj}$ and $\Delta p_{y,zj}$ using simulated data following normal distributions

1000 data sets were generated with the parameters listed for each "Scenario" (Table 2). (**a**): $g_{y,zj}$ and $\Delta p_{y,zj}$ are strongly correlated with each other (Pearson's rho = 0.98). In Scenarios 1-3, there was no difference between the populations; in Scenarios 4-7 the means and standard deviations of the populations truly differed. For visual clarity, only 100 randomly selected points from each scenario are plotted here. (**b**) and (**c**): Box-and-whisker-plots of $g_{y,zj}$ and $\Delta p_{y,zj}$ (respectively) in the four scenarios in which the two mock populations truly differed in their underlying means (see Table 2). In these plots, the centerline is the median value of the metric, the box shows the interquartile range (IQR), the whiskers extend to up to 1.5×IQR beyond the box, and the "+" symbols show points outside the latter range.

Furthermore, if one wishes to compare more than two populations (e.g. the three possible pairwise comparisons between three populations), each pairwise difference will be scaled by a different pooled standard deviation, rendering it awkward if not impossible to make meaningful quantitative comparisons between them.

A second limitation of $g_{y,zj}$ is that its value depends on the scale of measurement. For example, researchers will often apply non-linear transformations (such as log or arcsin transforms) to make their data meet the assumptions of parametric statistical methods (including normality and variance homogeneity). However, such transformations will also alter the calculated values of $g_{y,zj}$. This may be problematic if one wishes to compare distance measures for different traits and only some of the traits have been transformed or different traits have different natural scales of measurement (e.g., additive vs. multiplicative). Indeed, problems of this kind may often occur in sexual selection research when researchers aim to compare the divergence of naturally- versus sexually-selected traits (e.g., size vs. color).

In some cases, instead of comparing the divergence of different traits, one might want to have a single divergence measure involving multiple traits. Such a measure is given by the Mahalanobis distance (Mahalanobis, 1936; Arnegard et al., 2010), which may be seen as a multivariate generalization of $g_{y,zj}$ that also takes into account correlations between traits. However, the Mahalanobis distance faces the same restrictions as $g_{y,zj}$, that is, it requires a single estimate of the variance-covariance matrix for all populations, and its exact value will depend on scale(s) of measurement.

In summary, the limitations of $g_{y,zj}$ regarding unequal variances and scales of measurement not only affect pairwise comparisons, but also limit its applicability for comparisons involving multiple traits and/or populations.

A more useful distance metric would work for simple pairwise comparisons involving two populations and a single trait, but would also work for considering more than two populations and more than one trait simultaneously, so that (1) all pairwise effect sizes, even for traits measured in different units, would all be on the same scale, and (2) measures of overall distance (involving all traits at once) could be computed.

In the following sections, we introduce the derivation of Δp and via simulation and examples using empirical data, present its utility in a number of contexts in which a flexible measure of trait distance is required.

1 Methods

1.1 A nonparametric distance measure for arbitrary numbers of traits and population

Our new metric Δp does not make any assumptions about trait distributions or variances (i.e., it is nonparametric). Instead, it is based upon the joint (average) cumulative distribution function (cdf) across all populations for a given trait. Suppose there are N populations and t traits being considered. Let x_{ijk} denote the kth observation of the jth trait in the ith population, and n_{ij} the number of samples for trait j taken from population i. The joint empirical cdf for trait j (expressed in percentiles) is defined as

$$p_j(u) = \frac{100}{N}\sum_{i=1}^{N}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\mathbf{1}\{x_{ijk} \le u\} \qquad (3)$$

where u is any given value of trait j, and $\mathbf{1}\{\bullet\}$ is an indicator function that returns 1 if its argument is true and 0 otherwise. For illustration, imagine that the data for trait j from all populations have been pooled and sorted in increasing order. Let $x_{\min,j}$ denote the global minimum and $x_{\max,j}$ the global maximum. $p_j(u)$ is a step function which starts out at zero (for $u < x_{\min,j}$), jumps up by $100/(N\ n_{ij})$ at each x_{ijk}, and reaches 100 at $u = x_{\max,j}$. Importantly, by making the height of the jumps inversely proportional to the size of the sample a given data point stems from, we make sure that each sampled population contributes equally to $p_j(u)$, independent of sample or population size (e.g., for $N = 2$, each population is responsible for 50% of the total increase in $p_j(u)$).

Returning to the individual populations, we then ask: Into what percentile in the overall CDF does the median of each population fall? In other words, we calculate the value of $p_j(\hat{x}_{ij})$, where $\hat{x}_{ij}$ is the median value of trait j in population i, and we repeat this for all N populations. Our measure of phenotypic distance between populations y and z with respect to trait j is then defined as

$$\Delta p_{y,zj} \equiv p_j(\hat{x}_{yj}) - p_j(\hat{x}_{zj}) . \qquad (4)$$

As with $g_{y,zj}$, $\Delta p_{y,zj}$ can be positive or negative, in this case depending on whether population y or population z has the larger median. Importantly—and in contrast to $g_{y,zj}$ and other phenotypic distance measures—if there are more than two populations, all pairwise $\Delta p_{y,zj}$ values will be based on the same overall CDF, and hence, will be directly comparable. A numerical example outlining the above calculations is included in the online supplementary materials (Appendix 1) as a Microsoft Excel spreadsheet.

If data are available for more than one trait, the above analysis can be repeated for each trait separately. Again, the results will be comparable, because each phenotypic distance is measured at the appropriate scale (i.e., with respect to the overall CDF for that trait). In addition, we can also define an overall phenotypic distance for a pair of populations considering all traits simultaneously. The idea is to view the $p_j(\hat{x}_{ij})$ values (with $j = 1,\ldots,t$) of a single population as a set of "coordinates" for that population in a t-dimensional trait-percentile space (illustrated in Fig. 2). The coordinates for different populations naturally lend themselves to a notion of distance: for any number of traits being considered, we can calculate a Euclidean distance between the two populations using their percentile coordinates. We denote this distance between two groups or populations as

$$\Delta p_{y,z\bullet} \equiv \sqrt{\sum_{j=1}^{t}\left(p_j(\hat{x}_{yj}) - p_j(\hat{x}_{zj})\right)^2} , \qquad (5)$$

where $y, z \in \{1, 2, ..., N\}$ refer to two of the populations that were measured, and the subscript "•" denotes that this distance is calculated over all traits. Note that, unlike the single trait distance $\Delta p_{y,zj}$, $\Delta p_{y,z\bullet}$ is always positive; however, in the limiting case of just a single trait (i.e., $j = t = 1$), from equation (5) we obtain $\Delta p_{y,z\bullet} = \left|\Delta p_{y,z1}\right|$. Note also that $\Delta p_{y,z\bullet}$ is expected to increase as more traits are added to the analysis (as is true of any Euclidean distance as more dimensions are added).

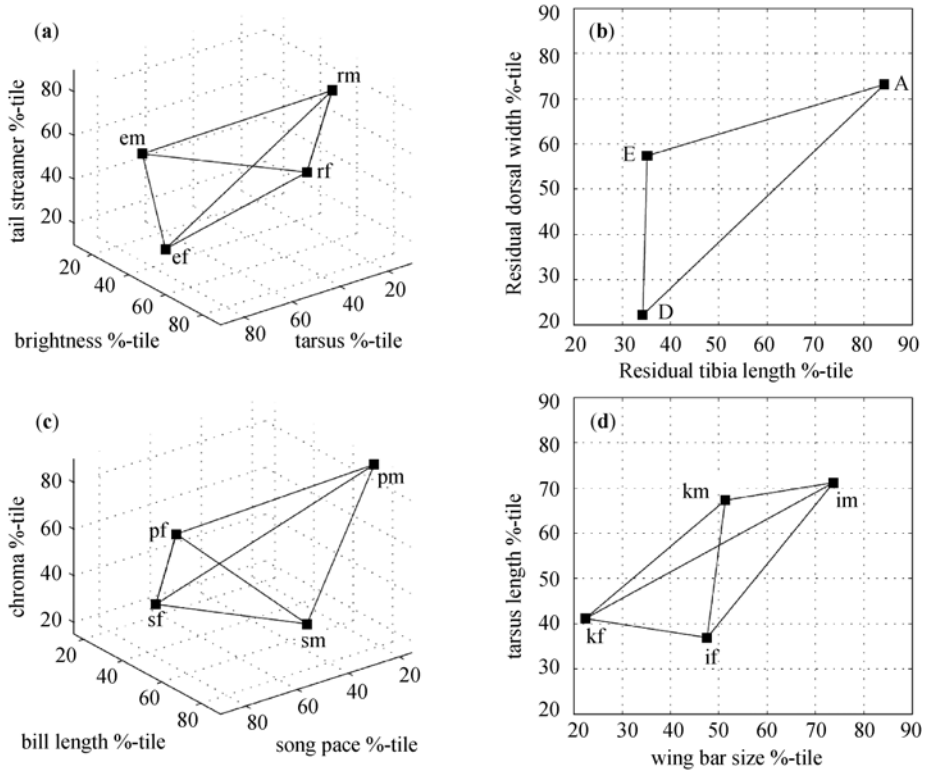

Fig. 2 Representations of population locations in trait-percentile spaces

(**a**) is based upon data on three traits from males and females (with sexes considered as separate "groups") for each of two subspecies of barn swallows *Hirundo rustica*. "em" and "ef" represent *H. r. erythrogaster* males and females (respectively; sample sizes = 71-98 for the three traits); "rm" and "rf" represent *H. r. rustica* males and females (sample sizes = 32-76 for the three traits). (**b**) Data on two traits from females in three species of painted forest toadlets ("A", "D", and "E" refer, respectively, to *Engystomops petersi* sp. A and sp. D and *E. freibergi*). Sample sizes for both traits for A, D, and E are 9, 9, and 36, respectively. (**c**) Data on *Hypocnemis peruviana males* ("pm") and females ("pf") and *Hypocnemis subflava* males ("sm") and females ("sf"). Sample sizes vary from 20-27 for both sexes for bill length and chroma; sample sizes for song pace are 5 (sf), 9 (pf), 17 (pm), and 21 (sm). (**d**) Data on populations of Hume's Warblers from Kyrgyzstan (males = "km"; females = "kf"; sample sizes = 39 for both sexes for both traits) and India (males = "im", sample size = 56; females = "if", sample size = 45). Information about the data contained in these figures can be found in online Appendix 3.

We note that for $t > 1$, the interpretation of $\Delta p_{y,z\bullet}$ as a Euclidean distance neglects correlations between traits within samples (e.g., correlations between traits j and j+1 within population y). Our justification is that, while the amount and direction of divergence relative to within-population correlations poses some extremely important questions (e.g. Schluter 1996), it does not seem possible to derive a single metric that captures all aspects of this problem, especially if the orientation of principal components differs between populations. (If the orientation of principle components is similar across populations, then if significant correlations between traits exist, $\Delta p_{y,z\bullet}$ can be calculated using results from principle components analysis.)

We have developed a MATLAB script to import data, perform all of the above calculations, and automatically generate a .csv file of results on Hedges' g values, $\Delta p_{y,z\bullet}$, $\Delta p_{y,zj}$, and other useful descriptive statistics. The commented source code (.m files), along with plain text explanations, metadata, and example input and output files are all freely available from the second author (SMF) upon request and have also been archived at SourceForge.net (http://sourceforge.net/projects/deltap/files/). Details of the derivation of confidence intervals are located in online Appendix 4.

1.2 Evaluating the performance of Δp

In order to explore and illustrate the behavior of $\Delta p_{y,z\bullet}$, we used numerical simulations to generate pseudo-random data sets on hypothetical traits from mock populations. We then applied the above methods to these data sets to explore realistic scenarios involving equal and unequal means for traits among populations, unequal sample sizes among populations, and unequal variances in a trait among populations (Table 2). We compared the behavior of $\Delta p_{y,z\bullet}$ in these "Scenarios" (Table 2) to the behavior of $g_{y,zj}$. By necessity, for the

purposes of directly comparing $g_{y,zj}$ and $\Delta p_{y,z\bullet}$, we considered only two mock populations and a single hypothetical trait (since that is all that can be used in a single calculation of $g_{y,zj}$); that is, we compare $g_{y,zj}$ and $\Delta p_{y,zj}$.

For these comparisons, we generated pseudo-random data following normal and exponential distributions. We show results below for several scenarios involving normally distributed data; additional scenarios with different sample sizes, means, standard deviations, and non-normal data are given in the supplemental materials (see online Appendix 2). For each "Scenario" in Table 2, we generated 1000 pairs of random samples from the two populations, each with the specified sample size and following the specified distribution. For each Scenario, we could thus calculate $g_{y,zj}$ and $\Delta p_{y,zj}$ 1000 times, independently. Analysis of variance (ANOVA) was performed to compare whether $g_{y,zj}$ systematically differed in different Scenarios, and likewise for $\Delta p_{y,zj}$.

1.3 Applying Δp to empirical data sets involving traits of known function

To demonstrate the application of Δp we solicited data from researchers working on systems where traits related to sexual signaling and ecological adaptation are well characterized. Criteria for inclusion of data were as follows: 1) the underlying mechanisms generating trait variation are fairly well-understood, such that one trait can be assumed to be predominantly underlain by natural selection via ecological adaptation and another to be predominantly underlain by sexual selection via variation in reproductive performance. 2) Data are from closely related taxa, ranging from sister taxa to sub-species to geographically isolated populations. See online Appendix 3 for details about the individual study systems, and the field and lab methods used to generate the unpublished data given in Tables 3 and 4; references for published data are given when available in Tables 3 and 4.

2 Results

2.1 Evaluating the performance of Δp

In simulations comparing the performance of $\Delta p_{y,zj}$ with $g_{y,zj}$, two important categories of results emerged (Fig. 1). First, if $\Delta p_{y,zj}$ is a valid distance metric, it should reproduce some aspects of the behavior of the well-established metric, $g_{y,zj}$. This was indeed the case: (1) $\Delta p_{y,zj}$ and $g_{y,zj}$ were very tightly correlated, (2) they both were centered on zero for cases when populations did not truly differ (Scenarios 1–3 in Figure 1a), and (3) they were both larger than zero for cases when the means of the populations truly differed (Scenarios 4–7 in all panels of Fig. 1). Second, however, we also found that $g_{y,zj}$ is much more sensitive to the combination of unequal variances and sample sizes (e.g., Grissom and Kim 2001; Fig. 1b) than $\Delta p_{y,zj}$ (Fig. 1c). In Scenarios 3–7, the true difference between populations was constant, but the simulated values of $g_{y,zj}$ varied systematically depending upon which population had the larger sample size and which had the larger variance (ANOVA on data in Fig. 1b: $F^3_{3996} = 104.75$, $P < 10^{-64}$. When the population with the smaller variance is sampled the most, then $g_{y,zj}$ will tend to overestimate the distance between populations; when the population with the larger variance is sampled the

Table 2 Parameters[a] used in simulated "Scenarios" used to compare $g_{y,zj}$ and $\Delta p_{y,zj}$

Scenario	Population y			Population z		
	n_{yj}	μ_{yj}[b]	σ_{yj}[c]	n_{yj}	μ_{yj}[b]	σ_{yj}[c]
1: unequal sample sizes only	56	103.36	8.89	120	103.36	8.89
2: unequal sample sizes only; means are equivalent but differ from scenario 1	56	90.27	6.77	120	90.27	6.77
3: unequal variances only	56	90.27	6.77	56	90.27	8.89
4: unequal means and variances only	56	103.36	8.89	56	90.27	6.77
5: unequal means, variances, and sample sizes	56	103.36	8.89	120	90.27	6.77
6: as in 5, but with reversed sample sizes	120	103.36	8.89	56	90.27	6.77
7: as in 4, but with larger sample sizes	120	103.36	8.89	120	90.27	6.77

[a] Sample sizes, means, and standard deviations were all inspired by a real data set on tail streamer lengths (in mm) for two subspecies of barn swallows (Safran and Evans, unpublished data).

[b] Assumed true population mean used for generating pseudorandom data.

[c] Assumed true population standard deviation used for generating pseudorandom data.

Table 3 Side by side comparisons of $g_{y,zj}$ and $\Delta p_{y,zj}$ for sexual vs. ecological traits in two closely related populations

Species	Comparison[1]	Sex Trait	$g_{y,zj}$	$\Delta p_{y,zj}$	Ecological Trait	$g_{y,zj}$	$\Delta p_{y,zj}$	source
Barn swallows *Hirundo rustica*	*H. r. erythrogaster* vs. *H. r. rustica* $n_{1s} = 71$, $n_{2s} = 53$, $n_{1e} = 82$, $n_{2e} = 76$	Tail length	−1.80	−46.36 (−49.18, −40.49)	Tarsus length	1.53	39.25 (29.67, 48.,44)	Safran and Evans *unpubl*
Hume's warbler *Phylloscopus humei*	Kyrgyztan vs India $n_{1s} = n_{1e} = 39$; $n_{2s} = n_{2e} = 56$	Wing bar size	−0.71	−28.50 (−38.10, −13.28)	Tarsus length	−0.25	−1.75 (−29.95,18.57)	Scordato *unpubl*
Greenish warblers *Phylloscopus trochiloides*	*P. viridanus* vs. *P. plumbeitarsus* $n_{1s} = 5$, $n_{2s} = 9$, $n_{1e} = 12$, $n_{2e} = 15$	Song units	−3.96	−47.78 (−57.89, −27.78)	Tarsus length	0.24	11.67 (−29.17, 40.83)	Irwin et al., 2001; Irwin et al., 2009
Pacific / Winter wrens *Troglodytes pacificus / hiemalis*	*T. hiemalis* vs. *T. pacificus* $n_{1s} = 13$, $n_{2s} = 19$, $n_{1e} = 9$, $n_{2e} = 34$	Song freq	−3.62	−49.39 (−54.66, −37.85)	Tarsus length	0.36	12.91 (−24.35, 44.28)	Toews and Irwin, 2008
Tree crickets *Oecanthus forbesi*	Wisconsin vs. Ohio $n_{1s} = 8$, $n_{2s} = 14$, $n_{1e} = 8$, $n_{2e} = 13$	File tooth #	−1.03	−40.18 (−64.29, −16.96)	Underwing length	−0.24	−3.87 (−40.38, 35.10)	Symes *unpubl*
Field Crickets *Gryllus*	*G. texensis vs G. rubens* [2] $n_{1s} = 164$, $n_{2s} = 122$, $n_{1e} = 119$, $n_{2e} = 102$	Pulse rate	4.79	50.91 (49.38, 54.27)	Ovipositor length	−2.40	−47.90 (−49.37, −44.75)	Izzo and Gray, 2004; Gray et al., 2001; Gray *unpubl*
Painted forest toadlet *Engystomops petersi*	*E. petersi* sp. A vs. sp. D. $n_{1s} = 4$, $n_{2s} = 12$, $n_{1e} = 14$, $n_{2e} = 19$	Call dom freq	−2.43	−50.00 (−66.67, −29.17)	Tibia length	1.85	41.54 (10.53, 48.68)	Boul et al., 2007; Funk et al., 2008; Funk *unpubl.*
Warbling antbirds *Hypocnemis* spp.	*H. peruviana* vs. *H. subflava* $n_{1s} = 17$, $n_{2s} = 21$, $n_{1e} = 20$, $n_{2e} = 27$	Song pace	−1.16	−33.75 (−45.67, −4.76)	Bill length	0.34	16.11 (−13.61, 41.02)	Tobias and Seddon, 2009 (song); Seddon and Tobias unpubl (morphology)
Manakins *Manacus*	*M. vitellinus* vs. *M. candei* $n_{1s} = n_{1e} = 17$, $n_{2s} = n_{2e} = 15$	Plumage brightness	−3.85	−50.20 (−56.86, −44.81)	Bill length	−0.24	0 (−25.10, 40.59)	Stein and Uy, 2006; Uy *unpubl*
Wolf spiders *Schizocosa*	*S. bilineata* vs *S. crassipalpata* $n_{1s} = 87$, $n_{2s} = 51$, $n_{1e} = 86$, $n_{2e} = 52$	Leg length	−1.25	−41.68 (−47.50, −33.40)	Cephalathorax width	−1.38	−43.43 (−47.29, −34.35)	Hebets *unpubl*

[1] In the "Comparison" column, n_{1s} denotes the sample size for the sexual trait in the first named population in the comparison, n_{1e} the sample size for the ecological trait in that population, n_{2s} the sample size for the sexual trait in the second named population, and n_{2e} the sample size for the ecological trait in the second population.

[2] For this comparison only, the sex trait occurs in males only and the ecological trait in females only.

Comparisons are males vs. males with the exception of *Gryllus* crickets in which case only males possess the sexual signal and females the ecologically relevant trait. A negative value of $g_{y,zj}$ and $\Delta p_{y,zj}$ indicates that the first population named has a smaller mean trait value compared to the second. For example, in the case of barn swallows, the length of tail streamers in the subspecies *erythrogaster* is shorter compared to *rustica*. Numbers in parentheses following $\Delta p_{y,zj}$ are 95% confidence intervals (see methods).

Table 4 Using $\Delta p_{y,zj}$ to compare sexual and ecological trait dimorphism

Species	Comparison	Dimorphism Sex Trait	$\Delta p_{y,zj}$		Dimorphism Ecological Trait	$\Delta p_{y,zj}$		source
			Pop 1	Pop 2		Pop 1	Pop 2	
Barn swallows *Hirundo rustica*	*H. r. rustica* vs. *H. r. erythrogaster*	Tail length	35.40 (29.22, 42.92)	38.24 (34.01, 44.70)	Tarsus length	−8.79 (−26.10, 0.00)	0.56 (−8.06, 8.25)	Safran and Evans, *unpubl*.
Hume's warbler *Phylloscopus humei*	Kyrgyztan vs. India	Wing bar size	28.96 (15.74, 41.83)	26.13 (10.42, 41.28)	Tarsus length	26.19 (8.86, 41.13)	34.24 (13.67, 48.06)	Scordato, *unpubl*.
Warbling antbirds *Hypocnemis* spp.	*H. peruviana* vs. *H. subflava*	Song pace	−62.08 (−77.24, −39.94)	−42.43 (−60.92, −30.24)	Bill length	20.23 (−13.70, 42.69)	21.81 (−1.04, 44.77)	Tobias and Seddon, 2009 (song); Seddon and Tobias, *unpubl* (morphology).
Mankins *Manacus*	*M. vitellinus* vs. *M. candei*	Plumage brightness	25.74 (20.74, 28.68)	74.70 (70.15, 78.03)	Bill length	0 (−16.08, 47.04)	−31.41 (−56.92, 0.00)	Stein and Uy, 2006; Uy, *unpubl*.
Wolf spiders *Schizocosa*	*S. bilineata* vs *S. crassipalpata*	Leg length	7.79 (−16.38, 25.41)	33.11 (21.49, 41.42)	Cephalathorax width	−47.69 (−54.13,-37.04)	−35.95 (−47.31, −24.01)	Hebets, *unpubl*.

In each analysis shown below, four distinct groups were included simultaneously (2 sexes × 2 populations). A positive value indicates that males have the larger mean value for a trait; a negative value indicates that females have the larger mean value.

most, $g_{y,zj}$ will tend to underestimate distance between populations. With large numbers of repeated simulations, slight differences could also be detected for $\Delta p_{y,zj}$ (ANOVA on data in Fig. 1c: $F^3_{3996} = 3.36,\ P < 0.02$). However, the degree of this sensitivity was an order of magnitude less for $\Delta p_{y,zj}$ than for $g_{y,zj}$: mean values of $g_{y,zj}$ ranged from 1.589 in Scenario 6 to 1.744 in Scenario 5, a difference of 10%. By contrast mean values of $\Delta p_{y,zj}$ ranged from 43.34 in Scenario5 to 43.83 in Scenario 5, a difference of only 1%.

2.2 Applying Δp to empirical data sets involving traits of known function

The results summarized in Fig. 2 and Table 3 exemplify the utility of $\Delta p_{y,zj}$ and $\Delta p_{y,z\bullet}$. Namely, multiple traits measured in different units, from across multiple populations can be compared simultaneously. In Figures 2a, 2c, and 2d, simultaneous comparisons of males from different populations, females from different populations, and dimorphism within populations can all be made. For example, Fig. 2a shows the extent to which barn swallow tail streamers are (i) sexually dimorphic in both populations (compare "ef" to "em" and "rf" to "rm" on the z-axis), (ii) divergent across populations (compare "em" to "rm" and "ef" to "rf"), and (iii) similar between *H. r. rustica* females and *H. r. erythrogaster* males (compare "rf" and "em" on the z-axis). Fig. 2b, data from two traits in females from three closely related populations of painted forest toadlets indicates the different axes of phenotype distance among these three closely related populations.

Although no formal conclusions about the relative significance of sexual selection and ecological adaption in the process of population divergence can be drawn from Table 3 (as these require phylogenetic correction and time-since-divergence analyses), our comparisons strongly indicate greater distances between sexual traits compared to ecological traits, leading to the inference that sexual traits are more divergent in closely related taxa compared to those traits related to ecological adaptation. This conclusion is supported by two aspects of the results shown in Table 3. First, the point estimates of $\Delta p_{y,zj}$ are greater (in magnitude) for the sexual trait than the ecological trait in 9 of 10 cases. Secondly, the 95% confidence intervals around $\Delta p_{y,zj}$ do not include zero for any of the sexual traits, yet they do include zero for six of the 10 ecological traits. Moreover, Table 4 indicates that sexual trait dimorphism may generally be greater than ecological trait dimorphism, where the function of each phenotypic trait has been addressed through empirical field study. In 7 of 10 comparisons (using results within each population in Table 4), traits with known sexual signaling function are more dimorphic compared to traits related to ecological adaptation.

3 Discussion

Testing predictions of hypotheses about the role of sexual selection in speciation - and many other investigations related to trait divergence - requires researchers to compare the relative degree of inter-population divergence for very different types of traits (e.g. size and color). Here, we have emphasized that commonly used parametric distance metrics, such as Hedge's g ($g_{y,zj}$), have several drawbacks, which limit their usefulness in such studies. First, the definitions of many of these metrics assume that the trait distributions in divergent populations have equal variances (reviewed above). If the variances are unequal (which will not always be known or apparent with empirical data), the expected value obtained from equation (1) depends on differences in sample size (Fig. 1). Second, the numerical value of $g_{y,zj}$ depends on the scale of measurement, and this metric will be affected if the data are subjected to a nonlinear transformation. This makes it difficult to compare the degree of divergence of different traits that may have been measured in very different ways (i.e., the problem of comparing "apples with oranges").

With these problems in mind, we developed a novel, non-parametric distance measure, Δp, which does not depend on equality of variances, is independent of the scale of measurement (because it is non-parametric), and facilitates comparisons of several traits across several populations. Δp is based on comparing the location of population medians in the joint (trait-wise) cumulative distribution function (CDF) across all populations. Viewed differently, Δp compares population medians after transforming the data into percentiles of the joint cdf (this view of percentiles as an alternative scale of measurement is illustrated in all panels of Fig. 2). The percentile scale serves as a common frame of reference for all comparisons involving a given trait. In addition, percentiles provide a natural normalization (since they always range from 0 to 100), and they are independent of the original scale of measurement (because they only depend on the ranking of the raw data). These properties, in turn, allow for meaningful comparisons of divergence

measures for different traits. In sum, measuring divergence at the percentile scale makes it possible to really compare "apples to apples".

Δp may also be interpreted as a measure of overlap between two distributions (see Huberty and Lowman, 2000). This is most clearly seen in the case of two populations and a single trait. If we assume, for simplicity, that both sample distributions are symmetric, then the maximal possible value of $\Delta p_{y,zj}$ is 50 (because the median of the smaller distribution is at least at the 25 percentile of the joint CDF, and the median of the larger distribution is at most at the 75 percentile). The difference between the actual value of $\Delta p_{y,zj}$ and the maximal value (50) is determined by how much the lower tail of the larger distribution overlaps with the median of the smaller distribution, and vice versa.

3.1 Empirical comparisons

Recent hypotheses about speciation propose that sexual signal divergence is accompanied by ecological trait divergence, predicting that sexual selection plays a role in speciation – in cases with and without gene flow – when ecological contexts differ (e.g., van Doorn et al., 2009). According to this model, sexual trait divergence in closely related populations should coincide with ecological trait divergence, but this is not the case in the various systems explored to demonstrate the utility of Δp (Table 3). Table 3 presents data on the divergence shaped predominantly by sexual or natural selection. Although not a formal quantitative comparison in which phylogenetic relationships or a metric of time since divergence would need to be accounted for, a striking pattern *when comparing closely related species only* is that sexual signals are more strongly divergent than ecological traits among disparate taxonomic groups. Moreover, the values of $\Delta p_{y,zj}$ are estimated on the same scale although these various acoustic signals, color variation, and morphological traits are measured in fundamentally different units. Thus, although in most cases $g_{y,zj}$ and $\Delta p_{y,zj}$ provide similar information about which traits are more divergent, $\Delta p_{y,zj}$ provides the advantage that ecological and sexual trait differentiation are *directly* comparable. An interesting exception is the wolf spider case, which suggests that the ecological trait is slightly more divergent compared to the sexual trait.

For those taxa in which sexual signals are present in both males and females, we derived dimorphism estimates using $\Delta p_{y,zj}$ to compute the differences in ecological and sexual traits between males and females. Similar to the case in Table 3, Table 4 is not a formal analysis of whether sexual traits are more dimorphic than ecological traits, though among the five taxa examined, support for greater dimorphism in sexual signals is evident. An interesting exception, again, is the wolf spiders which suggest that leg length (a putative sexual trait in these species) is either hardly dimorphic (*S. bilineata*) or very dimorphic (*S. crassipalpata*) and that in both taxa, the ecological trait (cephalothorax width) is equally dimorphic but the direction of dimorphism differs (in *S. crassipalpata* females are larger than males). It is important to note that *S. bilineata* males develop brushes upon their tibial forelegs upon maturation – a secondary sexual trait that makes them distinctly dimorphic (Stratton 2005), potentially relieving foreleg length from sexual selection in this species. Additionally, due to the potential for sexual cannibalism in spiders, selection from ecological selection versus sexual selection is often intertwined, making predictions less apparent. Whereas the data from Table 4 are not conclusive evidence to support the use of sexual dimorphism as a proxy of sexual selection on phenotypic traits (e.g., Kraaijeveld et al., 2011), they do indicate that – in traits of known function – sexual traits may tend to be more dimorphic compared to those underlain predominantly by natural selection in the study systems described in Table 4.

Finally, as illustrated in Fig. 2, whereas an overall metric of distance can be obtained across multiple traits from multiple populations, the advantage of $\Delta p_{y,z\bullet}$ is that the effect of one trait on overall distance among taxa can be quantified. For example, in Hume's Warblers (Fig. 2d), it is clear that the sexual signal wing bar size rather than tarsus length is a major contributor to overall phenotype distance between these closely related taxa (compare females, "if" and "kf", on the two axes; compare males, "im" and "km" on the two axes). Fig. 2d also shows that sexual dimorphism within populations is at least as pronounced as phenotypic divergence (within a sex) among populations.

3.2 Caveats and cautions in using $\Delta p_{y,zj}$ and $\Delta p_{y,z\bullet}$

One important consideration to keep in mind with $\Delta p_{y,zj}$ is that while the ability to use more than two populations simultaneously is a strength of this distance measure, the magnitude of $\Delta p_{y,zj}$ will change if a new population is added in the construction of $p_j(u)$. For example, suppose that $p_j(u)$ is constructed for two

populations (y and z) and $\Delta p_{y,zj}$ is calculated. Now suppose that observations from trait j in a third population (w) are added, and $p_j(u)$ is recalculated to reflect the observations on all three populations. $\Delta p_{y,zj}$ may now be reduced in magnitude if population w had more extreme trait values than the other populations; alternatively, if w was intermediate between y and z, then $\Delta p_{y,zj}$ would be increased in magnitude. This property of $\Delta p_{y,zj}$ is a direct consequence of the fact that percentiles given by $p_j(u)$ are always bounded on the interval [0,100], regardless of how many populations are being considered. The important consideration here is that if one wishes to compare the magnitudes of different $\Delta p_{y,zj}$ values that were calculated *independently* from one another—for example, as might be done in a meta-analysis—then it is important that the calculations (1) involve the same set of populations (or at least, comparable sets of populations, e.g. 2 sympatric and one allopatric population) and (2) do not involve "saturation" of the metric (see below). In order to facilitate ease of conducting meta-analyses, we suggest it might be useful for any researcher reporting $\Delta p_{y,zj}$ to report the pairwise distances (calculated from just two populations) along with the distances calculated for >2 populations. However, whenever possible, this issue should be avoided by using one of the strengths that $\Delta p_{y,zj}$ offers: all the populations should be put into the same analysis (rather than calculating $\Delta p_{y,zj}$ values independently in different analyses). When all populations are compared in a single analysis, all comparisons of $\Delta p_{y,zj}$ values will truly be "apples to apples". The more general point here is that one of our main motivations for developing $\Delta p_{y,z\bullet}$ was the need for a way to fairly compare distances among arbitrary numbers of populations and traits *simultaneously* and all on the same scale. When analyses are performed that way (i.e., one analysis using all appropriate data simultaneously), comparisons of magnitudes of $\Delta p_{y,z\bullet}$ values (and $\Delta p_{y,zj}$ values calculated as part of $\Delta p_{y,z\bullet}$) will be valid.

A second consideration is that as differences between groups being compared become large, $\Delta p_{y,zj}$ will eventually "saturate." For example, in a pairwise comparison of body mass of hummingbirds and cheetahs, $\Delta p_{y,zj}$ will be at its expected maximum magnitude (approximately 50 for a two-population analysis: see note below). The same would be true of an independent pairwise comparison of body mass between hummingbirds and elephants. The solution here is once again to use the features that $\Delta p_{y,zj}$ offers: the data on hummingbirds, cheetahs, and elephants should all be included in a single analysis, in which case $\Delta p_{y,zj}$ will resolve distances between populations appropriately. Another point to consider here is that we expect that most applications of $\Delta p_{y,zj}$ and $\Delta p_{y,z\bullet}$ will involve closely related groups, in which case "saturation" of the metric is unlikely to diminish its utility. For example, across the wide range of taxa and types of traits shown in Tables 3 and 4, it would have been problematic if a comparison involved two values of $\Delta p_{y,zj}$ that were both near saturation values. There was only one case in which this occurred: in Table 3, in the row for field crickets in the genus *Gryllus*, $\left|\Delta p_{y,zj}\right|$ was near 50 for *both* the sexual trait and the ecological trait. However, in this case, the narrow, non-overlapping confidence intervals around each estimate of $\Delta p_{y,zj}$ still permit a meaningful comparison showing that it is highly likely that the sexual trait is more phenotypically divergent than the ecological trait. In other cases where comparisons did not produce unequivocal differences, it is sample size (and associated wide confidence intervals) rather than saturation that is the limiting factor. We note that when only two populations are considered, the theoretical expected maximum value of $\Delta p_{y,zj}$ with infinite sample sizes is 50. However, values slightly larger than this can be realized for real data sets and for the confidence intervals around $\Delta p_{y,zj}$, especially when sample sizes are small, as is seen occasionally in Table 3. This is because—with finite sample sizes—there is no reason that the medians of two non-overlapping trait distributions must fall exactly at the 25th and 75th percentiles in the joint CDF, $p_j(u)$. In particular, deviations can occur when the medians coincide exactly with one or more trait values. We also note that Table 4 has values larger than 50 for a different reason: there are 4 populations included simultaneously in the calculations of $\Delta p_{y,zj}$.

A third and practical consideration is that calculating $\Delta p_{y,zj}$, and thus $\Delta p_{y,z\bullet}$, requires raw data. Calculation

of $g_{y,zj}$ requires only having means, standard deviations, and sample sizes, which are often easy to obtain from published works; by contrast, $\Delta p_{y,zj}$ utilizes a distribution of data. While the latter contributes to its desirable properties, it also means that one cannot calculate $\Delta p_{y,zj}$ without access to original data sets (or at least, a random subsample of data from an original data set). In the current academic climate of free, electronic access to original data sets—which indeed, is now required upon publication by a number of journals in ecology and evolutionary biology (Fairburn, 2011)—we expect that the need for original data will be much less of an impediment than it might have been even just a decade ago. In light of this transition and because of issues related to the number of populations in a study and saturation, we recommend publishing both $\Delta p_{y,zj}$ and $g_{y,zj}$ side by side in studies related to phenotype distances, noting the advantages and disadvantages associated with each effect size metric.

Acknowlegements We thank Matthew Arnegard, Carlos Botero, Tamra Mendelson, Rafael Rodriquéz and Sander van Doorn for excellent discussions about the need for a new phenotypic distance metric and Maria Servedio for the invitation and encouragement to formalize our ideas. This research was supported as part of the Sexual Selection and Speciation working group by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606. RJS and SMF were supported by the University of Colorado and National Science Foundation grant IOS-0717421to RJS. MK was supported by a grant from the Vienna Science and Technology Fund (WWTF) to the Mathematics and Biosciences Group at the University of Vienna. EAH thanks Mitch Bern for use of his Master's thesis data and was supported by the National Science Foundation grant IOS - 0643179. DEI and DPLT were supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grants 311931-2005 and 311931-2010 to DEI, CGS-D to DPLT). NS and JAT were supported by the Royal Society, British Ecological Society and John Fell Fund (Oxford University). ES supported by NSF-DDIG, the American Ornithologists Union, the University of Chicago, and the American Philosophical Society Lewis and Clark award. JACU was funded by National Science Foundation grant IOS 0306175.

References

Arnegard ME, McIntyre PB, Harmon LB, Zelditch ML, Crampton WGR et al., 2010. Sexual signal evolution outpaces ecological divergence during electric fish species radiation. Am. Nat. 176: 335–356.

Boul KE, Funk WC, Darst CR, Cannatella DC, Ryan MJ, 2007. Sexual Selection Drives Speciation in an Amazonian Frog. Proc. Roy. Soc. B 274: 399–406.

Charlesworth B, 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15: 538–543.

Charlesworth B, Charlesworth D, Barton NH, 2003. The effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. Evol. Syst. 34: 99–125.

Cohen J, 1969. Statistical power analysis for the behavioral sciences. 1st edn. New York: Academic Press.

Endler J, Houde AE, 1995. Geographic variation in female preferences for male traits in *Poecilia reticulate.* Evolution 49: 456–468.

Excoffier L, 2001. Analysis of population subdivision. In: Balding DJ, Bishop M, Cannings C ed. Handbook of Statistical Genetics. New York: John Wiley, 271– 307.

Fairburn DJ, 2011. The advent of mandatory data archiving. Evolution 65: 1–2.

Freeman-Gallant CR, Taff CC, Morin DF, Dunn PO, Whittingham LA et al., 2009. Sexual selection, multiple ornaments, and age- and condition- dependent signaling in the common yellowthroat. Evolution 64: 1007–1017.

Funk WC, Angulo A, Caldwell JP, Ryan MJ, Cannatella DC, 2008. Comparison of morphology and calls of two cryptic species of *Physalaemus* (Anura: Leiuperidae). Herpetologica 64: 290–304.

Gray DA, Cade WH, 2000. Sexual Selection and Speciation. Proc. Natl. Acad. Sci. USA 97: 14449–14454.

Gray DA, Walker TJ, Conley BE, Cade WH, 2001. A morphological means of distinguishing females of the cryptic field cricket species *Gryllus rubens* and *G. texensis* (Orthoptera: Gryllidae). Florida Entomologist 84: 314-315.

Grissom RJ, Kim JJ, 2001. Review of assumptions and problems in the appropriate conceptualization of effect size. Psych. Methods: 6, 135–146.

Hedges LV, 1981. Distribution theory for Glass's estimator of effect size and related estimators. J. Educ. Stat. 6: 107–128.

Hedrick PW, 2005. A standardized genetic differentiation measure. Evolution 59: 1633–1638.

Huberty CJ, Lowman LL, 2000, Group overlap as a basis for effect size. Educ. Psych. Measurement 60: 543–563.

Irwin DE, Bensch S, Price TD, 2001. Speciation in a ring. Nature 409: 333–337.

Irwin DE, Thimgan MP, Irwin JH, 2008. Call divergence is correlated with geographic and genetic distance in greenish warblers *Phylloscopus trochiloides*: A strong role for stochasticity in signal evolution? Journal of Evolutionary Biology 21: 435–448.

Izzo AS, Gray DA, 2004. Cricket song in sympatry: Examining reproductive character displacement and species specificity of song in *Gryllus rubens*. Annals of the Entomological Society of America 97: 831–837.

Johnsen A, Andersson S, Fernandez JG, Kempenaers B, Pavel V et

al., 2006. Molecular and phenotypic divergence in the bluethroat subspecies complex. Molec. Ecol. 15: 4033–4047.

Kraaijeveld K, Femmie Kraaijeveld-Smit JL, Maan M, 2011. Sexual selection and speciation: The comparative evidence revisited. Biol. Rev. 86: 367–377.

Lande R, 1981. Models of speciation by sexual selection on polygenic traits. Proc. Natl. Acad. Sci USA 78: 3721–3725.

Maan ME, Seehausen O, 2011. Ecology, sexual selection and speciation. Ecol. Letters. 14: 591–602.

Mahalanobis PC, 1936. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2: 49–55.

Mayr E, 1947. Ecological factors in speciation. Evolution 1: 263–288.

Mendelson TC, Shaw KL, 2005. Sexual behaviour: Rapid speciation in an arthropod. Nature 433: 375–376.

Nagawa S, Cuthill IC, 2005. Effect size, confidence interval and statistical significance: A practical guide for biologists. Biol. Rev. 82: 591–605.

Panhuis TM, Butlin R, Zuk M, Tregenza T, 2001. Sexual selection and speciation. Trends Ecol. Evol. 6: 364–371.

Price TD, 1998. Sexual selection and natural selection in bird speciation. Philosophical Transactions of the Royal Society of London B 353: 251–260.

Price TD, 2008. Speciation in Birds. Greenwood Village: Roberts and Company.

Ritchie MG, 2007. Sexual selection and speciation. Annual Review of Ecology, Evolution and Systematics 38: 79–102.

Rodríguez RL, Sullivan LE, Cocroft RB, 2004. Vibrational Communication and reproductive isolation in the *Enchenopa binotata* species complex of treehoppers (Hemiptera: Membracidae). Evolution 58, 571–578.

Safran RJ, McGraw KJ, 2004. Plumage coloration, not length or symmetry of tail-streamers, is a sexually selected trait in North American barn swallows. Behavioral Ecology 15: 455–461.

Seehausen O, Van Alphen JM, 1999. Can sympatric speciation by disruptive sexual selection explain rapid evolution of cichlid diversity in Lake Victoria? Ecology Letters 2: 262–271.

Seddon N, Merrill RM, Tobias JA, 2008. Sexually selected traits predict patterns of species richness in a diverse clade of suboscine birds. American Naturalist 171: 620–631.

Schluter D, 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50: 1766–1774.

Slatkin M, 1987. Gene flow and the geographic structure of natural populations. Science 236: 787–792

Sokal RR, Rohlf FJ, 1995. Biometry: The Principles and Practice of Statistics in Biological Research. 3rd edn. Freeman: San Francisco.

Stein AC, Uy JAC, 2006. Plumage brightness predicts male mating success in the lekking golden-collared manakin. Behavioral Ecology 17: 41–47.

Stratton GE, 2005. Evolution of ornamentation and courtship behavior in Schizocosa: Insights from a phylogeny based on morphology (Araneae, Lycosidea). Journal of Arachnology 33: 347–376

Svensson EI, Eroukhmanoff F, Friberg M, 2006. Effects of natural and sexual selection on adaptive population divergence and premating isolation in a damselfly. Evolution 60: 1242–1253.

Tobias JA, Seddon N, 2009. Signal design and perception in *Hypocnemis antbirds*: Evidence for convergent evolution via social selection. Evolution 63: 3169–3189.

Toews DPL, Irwin DE, 2008. Cryptic speciation in a Holarctic passerine revealed by genetic and bioacoustic analyses. Molecular Ecology 17: 2691–2705.

Uy JAC, Borgia G, 2000. Sexual selection drives rapid divergence in bowerbird display traits. Evolution 54: 273–278.

Uy JAC, Moyle RG, Filardi CE, 2008. Plumage color and song differences mediate species recognition between incipient flycatcher species of the solomon islands. Evolution 63: 153–164

van Doorn S, Edelaar P, Weissing FJ, 2009. On the origin of species by natural and sexual selection. Science 326: 1704–1707.

West- Eberhard MJ, 1983. Sexual selection, social competition, and speciation. Quarterly Review of Biology 58: 155–183

Wright S, 1943. Isolation by distance. Genetics 28: 114–128.

Wright S, 1951. The genetical structure of populations. Ann. Eugen. 15: 323–354.

Wright S, 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19: 395–420.

Wright S, 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. 70: 3321–3323

Wright S, 1978. Evolution and the Genetics of Populations: Variability within and among Natural Populations. Chicago: Chicago Univ. Press.