Anais da Academia Brasileira de Ciências (2005) 77(1): 45-76
(Annals of the Brazilian Academy of Sciences)
ISSN 0001-3765
www.scielo.br/aabc

Determining sexual dimorphism in frog measurement data: integration of statistical significance, measurement error, effect size and biological significance

LEE-ANN C. HAYEK[1] **and W. RONALD HEYER**[2]

[1]Mathematics and Statistics, MRC 136, National Museum of Natural History, Smithsonian Institution
PO Box 37012, Washington, DC 20013-7012, USA
[2]Amphibians and Reptiles, MRC 162, National Museum of Natural History, Smithsonian Institution
PO Box 37012, Washington, DC 20013-7012, USA

Manuscript received on June 16, 2004; accepted for publication on September 22, 2004;
contributed by WILLIAM RONALD HEYER*

ABSTRACT

Several analytic techniques have been used to determine sexual dimorphism in vertebrate morphological measurement data with no emergent consensus on which technique is superior. A further confounding problem for frog data is the existence of considerable measurement error. To determine dimorphism, we examine a single hypothesis (*Ho* = equal means) for two groups (females and males). We demonstrate that frog measurement data meet assumptions for clearly defined statistical hypothesis testing with statistical linear models rather than those of exploratory multivariate techniques such as principal components, correlation or correspondence analysis. In order to distinguish biological from statistical significance of hypotheses, we propose a new protocol that incorporates measurement error and effect size. Measurement error is evaluated with a novel measurement error index. Effect size, widely used in the behavioral sciences and in meta-analysis studies in biology, proves to be the most useful single metric to evaluate whether statistically significant results are biologically meaningful. Definitions for a range of small, medium, and large effect sizes specifically for frog measurement data are provided. Examples with measurement data for species of the frog genus *Leptodactylus* are presented. The new protocol is recommended not only to evaluate sexual dimorphism for frog data but for any animal measurement data for which the measurement error index and observed or a priori effect sizes can be calculated.

Key words: statistics, sexual dimorphism, measurement error index, effect size, frogs.

INTRODUCTION

Study of animal sexual dimorphism can lead to important biological insights. For example, in a seminal frog paper, Shine (1979) convincingly demonstrated that for species in which male combat occurs, the males are often larger than females. Aside from Shine's paper (1979), the causes of sexual dimorphism in frogs are not known in most cases.

Mouth width is known to correlate with prey size (Duellman and Trueb 1986:238) and hindlimb length with locomotion type (jumping, hopping, burrowing: Duellman and Trueb 1986:356, 365) among species of frogs and both may be of biological significance for sexual differences of these variables within species (Heyer 1978).

*Member Academia Brasileira de Ciências
Correspondence to: W. Ronald Heyer
E-mail: heyer.ron@nmnh.si.edu

There are two outstanding problems when evaluating sexual dimorphism in measurement variables in frogs: (1) large measurement error, and (2) statistical versus biological significance.

Measurement error in frogs is large and impacts both statistical and biological results (Hayek et al. 2001). As part of a recent study, WRH detected an apparent conflict between statistical and biological significance for several morphological variables in a group of large species of the frog genus *Leptodactylus* (Heyer 2005). WRH brought the problem to LCH, who proposed a study on appropriate statistical methodology for evaluating sexual dimorphism for measurement data in frogs. LCH suggested that WRH select a limited number of data sets that would allow for evaluation of problems associated with sample sizes and geographic variation and that would likely exhibit a range of variation in sexual dimorphism. LCH would then use these data to examine appropriate statistical procedures for evaluating sexual dimorphism in the variables measured.

Through review of the literature and analyses of our data we find a new approach to the problem is superior to other methods in use. Our protocol consists of the following sequential steps:

1) Analyze the overall size measurement data (in our case snout-vent lengths [SVL]) with **ANOVA** and the other measurement variables with **ANCOVA** (using SVL as the independent variable) to determine whether the results are statistically significant. If the results are statistically significant, proceed to the next step.

2) Evaluate the statistically significant results from Step 1 with the **measurement error index**, developed herein, to screen out statistically significant results that are compromised by measurement error. For results that are not compromised by measurement error, proceed to the final step.

3) Calculate and use **effect size** (ES) coefficients to evaluate the biological significance of the statistically supported results. Effect size values are standardized scores that can be compared across studies irrespective of sample sizes. We find that small effect size values are not biologically meaningful in our data, but that medium and large effect size values do have biological meaning.

We lay out arguments for the appropriateness of this 3-step protocol for frog measurement data; discuss this protocol in terms of other approaches used in the literature to study sexual dimorphism in measurement data; define small, medium, and large effect sizes for frog measurement data; and show examples of the application of the new protocol with frog data.

We propose that the procedure described in this paper should be adopted in future studies when evaluating sexual dimorphism of measurement data in animals in general.

MATERIALS AND METHODS

Materials

Almost all of the data used in this study come from years of study of the variation in members of the frog genus *Leptodactylus* by WRH. The variables are: Snout-vent length (SVL), a measure of overall size; head length; head width; head area; eye-nostril distance; tympanum diameter; thigh length; shank length; and foot length. Not all of these variables were examined in earlier studies, so, there are no data or smaller sample sizes for eye-nostril distance and tympanum diameter in some cases. Methods for taking the measurements are those found in Heyer (2005). Head area is calculated as one-half an ellipsoidal conic section fit to the triangular area determined from measured head length and head width of each frog in the study.

The data were selected to answer a variety of questions. One problem of concern was whether characterizations of sexual dimorphism based on specimens throughout the geographic range differed from characterizations based on single locality samples. Specifically, should sexual dimorphism always be studied at a local level? Two data sets address this problem: (1) a substantial sample available for *Leptodactylus fuscus* throughout its geographic range (Panama to Argentina) and a single large sample of *L. fuscus* SVL data from Porto Velho, Brasil; and (2) a substantial sample of the widely distributed *Leptodactylus podicipinus* (southern Amazonia, central and eastern Brasil to northern Argentina) and four reasonably-sized samples from single localities (Alejandra, Bolivia; Curuçá, Brasil; Porto Velho, Brasil; Rurrenabaque, Bolivia).

A second problem involved sexual dimorphism of similar species within a single genus. Two sets of data analyzed in previous studies demonstrated different statistically significant results for measurements made between morphologically similar appearing species (Heyer 1978): (1) the species pair *Leptodactylus bufonius* and *L. troglodytes*; and (2) the species pair *Leptodactylus furnarius* and *L. gracilis*.

In addition, questions regarding biological versus statistical significance were raised for the species *Leptodactylus knudseni*, *L. pentadactylus*, and an undescribed species, referred to herein as Middle American pentadactylus (Heyer 2005).

Finally, the importance of determining effect sizes (ES) to define sexual size dimorphism in frogs became clear during the course of our study. A previously assembled data set for *Eleutherodactylus fenestratus* (Heyer and Muñoz 1999) was included because the effect size values of this species would be at the large end of a possible range of values. The difference in male and female size in *E. fenestratus* is obvious by visual inspection.

To evaluate measurement error, individual specimens were measured 20 times. Maximum and minimum values were obtained from these measurements. Three individuals of about the same SVL were selected for measurement at more-or-less regular intervals spanning the adult size ranges of species of *Adenomera* and *Leptodactylus*, with the exception that only one specimen was available for the largest size category. Previous data were available for one individual of *Vanzolinius discodactylus* (Hayek et al. 2001) and a male and female of an undescribed species from Pará, Brasil (Heyer 2005). In addition to the previously available date, the following specimens were re-measured: *Adenomera marmorata* – USNM 209101, 209110, 209112, *Leptodactylus knudseni* – USNM 216785, 531513, *L. labyrinthicus* – USNM 121284, 303175, 370593, 507904, *L. leptodactyloides* – USNM 202519, 202522, 321214, *L. myersi* – USNM 302191, *L. podicipinus* – USNM 148685, 148686, *L. rhodomystax* – USNM 343256, 343257, 531559, *L. vastus* – USNM 109144, 109148. These specimens not only span the size range for leptodactylid frogs in general, but also include examples of well and poorly preserved individuals (Fig. 1). Twenty data forms were produced on which to record the measurement data for these 20 frogs. Only one data sheet was filled out on any given day. The 20 specimens were placed in three containers, one containing the smallest individuals, one the medium-sized, and one the largest. The order of container examination was indicated on the top of each data form so that each container was examined almost equally either first, second, or third in the study. Individuals were haphazardly selected from each container each session. The date and time were also recorded on each form as they were filled out. All measurements were taken by WRH to avoid inter-observer error.

Evaluation of Statistical Methods

Statistical significance for a hypothesis of sexual dimorphism of frog body parts using measurement data indicates whether the study results are due to chance or to sampling variability. Total reliance upon statistical tests for amphibian hypotheses leads to the anomalous results that prompted the present study. Previous studies of size sexual dimorphism in frogs seems to have been reduced to the selection of a fixed level of significance and a desire for a dichotomous reject/ do not reject decision regardless of sample size to test merely whether there is or is not a difference of 0. The alternative hypothesis is, by default, that any unspecified statistically significant sexual size difference at all that is not

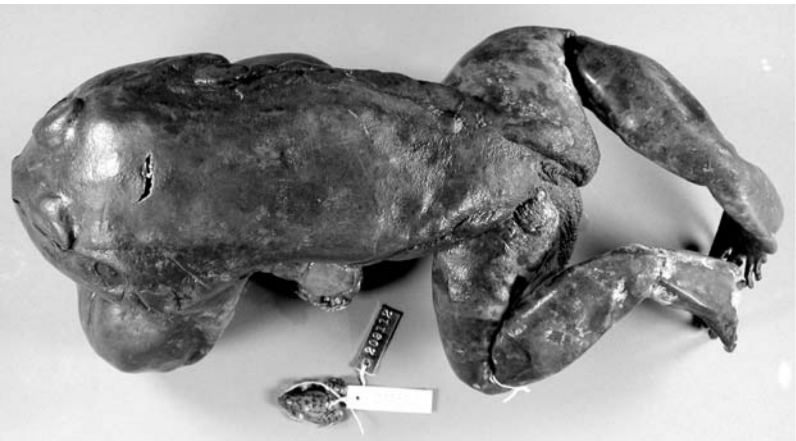

Fig. 1 – Smallest and largest individuals used to determine measurement error, showing both the size differences involved and variation in quality of preservation. Above – *Leptodactylus vastus*, USNM 109144; below – *Adenomera marmorata*, USNM 209112.

0, is equated with biological importance. There is little if any emphasis upon the actual or expected size of that difference in nature or whether such a difference has biological meaning. Therefore, contradictory results occur when the *p*-value is the focal point. Two tests on the same species can lead to results that on the one hand infer sexual dimorphism and on the other hand deny any differences exist. For example, when a total of 35 *L. furnarius* were examined (Heyer 1978), male head length was larger than female. With a sample size of 74 specimens in the present study, the opposite was found. We conclude that emphasis on *p*-value statistical test results alone is not what the researcher should be seeking. Understanding size sexual dimorphism in frogs requires answers to questions of existence, magnitude, and strength of any association or inter-relationship.

Statistical significance, or *p* value, actually provides little insight about frog size dimorphism. The result of the statistical test depends upon sample size, test level, and power at which the test was performed, as well as the difference between the quantities being tested. To reject a null hypothesis of no sexual dimorphism is to reject that the size difference between the sexes is really 0. Since all nature varies, it follows that before any statistical test is even performed, such a strict null hypothesis has to be false (given a large enough sample size). If we reject a 0 difference between the sexes, what is the alternative? Failing to demonstrate an effect is quite distinct from either implicitly or explicitly concluding that no difference exists at all. Is dimorphism then any male-female difference on average? Clearly, hypothesis test results provide no indication of the magnitude of the difference between the sexes, or the actual effect of dimorphism's being observed. For example, based upon hypothesis test results for SVL of $p < 0.05$, both *L. pentadactylus* and *L. troglodytes* exhibit sexual dimorphism. However, for the former species the mean difference between the sexes is 13.7 mm (maximums: 195 mm males; 174 mm females); whereas for the latter, it is 1.3 mm (maximums: 52.8 mm males; 52.7 mm females). Not only are these values highly discrepant, but with a two-sided test ''significance'' indicates ''not equal''. The test result cannot provide inference on significance of males or females being larger. We require that the maximum values exhibit a reasonably large difference in the same direction indicated by the statistical test results. Classical statistical significance tests are not independent of sample: the larger the sample size the more likely is rejection (and in practice power is higher). Thus, there is always a sample size that will allow for the rejection of any non-zero difference; with enough specimens the sexes will be called dimorphic. Because this dependence is so often ignored in amphibian research we propose supplementing significance test results with two factors: a measurement error index; and a standardized, biologically meaningful effect size defined herein specifically for frogs. These two quantities are used in tandem to determine existence of consequential size dimorphism.

Statistical Methods

Descriptive and inferential statistical analyses were computed for each measurement variable for all adult individuals of each species and by sex. Locality analyses were performed on the *L. fuscus* and *L. podicipinus* data. All assumptions, hypothesis tests, and analyses under a general linear model were performed using SPSS (SPSS for Windows, version 11.0, 2001, SPSS Inc., Chicago). Power and effect size calculations were programmed into Mathcad Professional 2000 (Mathsoft Inc., Cambridge, Massachusetts). We used a modification (see Appendix I) of Cohen's **d** (1977) as our effect size measure, $\mathbf{d} = \mathbf{m_f} - \mathbf{m_m}/\sigma$, where **d** = effect size index, $\mathbf{m_f}$, $\mathbf{m_m}$ = population means expressed in original measurement unit, and σ = the standard deviation of either population (assuming they are equal). Cohen's **d** was selected because means are the focus of any study of sexual dimorphism. For our study we defined **d** as the standardized mean difference of female versus male measurements and σ as the pooled standard deviation of the two groups (Appendix I). Under an ANOVA model the numerator is the difference between female and male means. Under an ANCOVA model the numerator uses the difference between covariance-adjusted means. This measure, **d**, can also be computed from regression calculations that give correlation coefficients. That is, **d** is defined as twice a correlation coefficient divided by the square root of one minus the coefficient squared. The two calculation methods yield equivalent values for **d**. Computations of effect size as either average percentile or percentage non-overlap were programmed in Mathcad following Cohen (1977). Reliability calculations for indices and regressions were modeled with SYSTAT (Wilkinson and Coward 2000).

RESULTS

Determination of a New Measurement Error Index

Regression analyses were performed for each measurement variable with mean SVL as the independent variable and the range of each variable as dependent. The data used were the individuals measured 20 times each. The mean SVL is the mean of the 20 measurements of each individual. The range of each variable was the maximum measurement minus the minimum measurement of the 20 measurements of each individual.

For some variables, linear regression was the most appropriate analysis (e.g. head width, Fig. 2), for others, quadratic regression was more appropriate (e.g. SVL, Fig. 3). Table I gives the regression formula suitable for each variable.

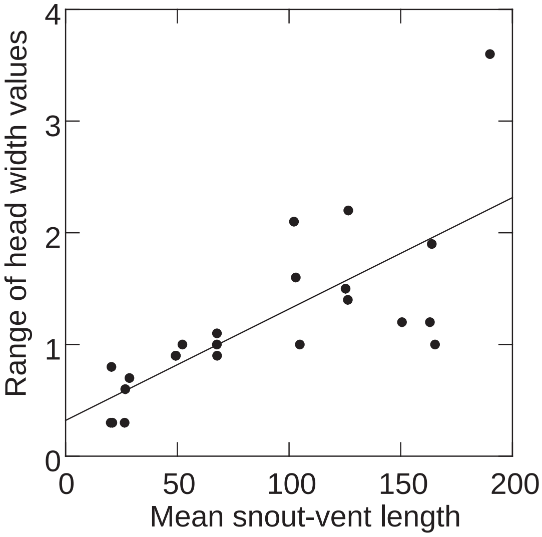

Fig. 2 – Measurement error data for head width with linear smoother. All values in mm.

Measurement errors were greatest for the largest specimens. In general, more manipulation

TABLE I

Regression formulae for measurement error of variables for leptodactylid frogs.

Variable	Regression formula	Significance (p-value)	**Adjusted multiple r^2**
SVL	$y = 2.262 - 0.062x + 0.001x^2$	0.000	**0.876**
Head length	$y = 1.529 - 0.033x + 0.0001x^2$	0.000	**0.772**
Head width	$y = 0.320 + 0.010x$	0.000	**0.534**
Eye-nostril distance	$y = 0.160 + 0.006x$	0.000	**0.755**
Tympanum diameter	$y = 0.027 + 0.007x$	0.000	**0.804**
Thigh length	$y = 0.500 + 0.015x$	0.000	**0.559**
Shank length	For $x < 25$ mm, use 0.5 mm for $x > 25$ mm, $y = 0.159 + 0.007x$	0.000	**0.480**
Foot length	y = 1.031 + 0.008x	0.001	**0.408**

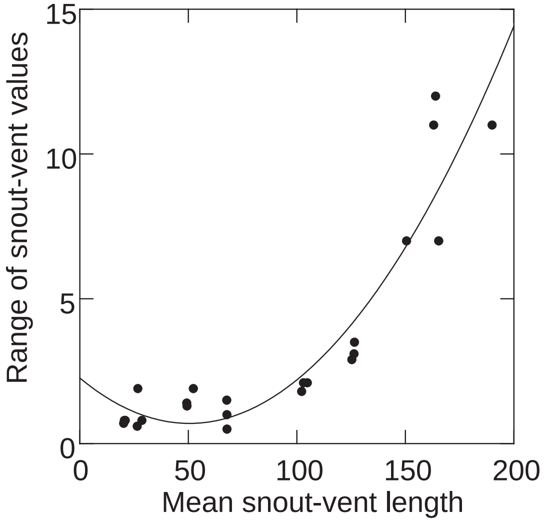

Fig. 3 – Measurement error data for SVL with quadratic smoother. All values in mm.

of specimens is required the larger they are to position them for measurement of each variable. In the case of SVL, such error is particularly true for large, poorly preserved specimens where the specimen must be flattened out to take the measurement. In some cases, measurement error was greater for the smallest size specimens relative to moderate sized specimens (e.g., those variables for which the quadratic regression is most appropriate such as SVL, Fig. 3). For example, to measure head length, the proximal point of the needle nose caliper is ''hooked'' behind the jawbones. For the small specimens, the size of the needle nosed point is large relative to distinguishing the posterior angle of the jawbones from the overlying skin and associated tissues. Some variables were measured more accurately than others. For example, shank length was measured more accurately than either thigh length or foot length (Fig. 4).

The regression formulae determined for mean measurement errors (Table I) are appropriate to evaluate measurement error in this study, since WRH took all of the measurement data. However, we propose that these regression formulae are appropriate to evaluate measurement error in any study involving frogs with similar overall body shapes, such as members of the families Leptodactylidae, Myobatrachidae, and Ranidae. Tree-frogs (Hylidae, Rhacophoridae) and hopping toads (Bufonidae) should be at least spot-checked for some variables to determine whether measurement error results are comparable to those established herein.

We used the above results to define a measurement error index as the mean sexual difference divided by the measurement error regression quantity for the variable of interest (solving for y in the equations of Table I, also see Measurement Error Screening of Statistically Significant Results, below).

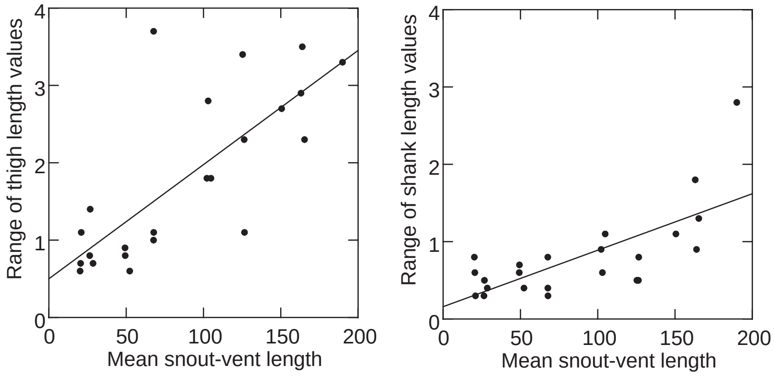

Fig. 4 – Comparison of magnitude of measurement errors for thigh length (left) and shank length (right). All values in mm.

Statistical Tables

Tables II-VI provide comparisons of mean difference values and conventional hypothesis tests of means for males and females across species, by sex and locality. The statistical significance of results is designated by p-value, where $p < 0.05$ indicates the null hypothesis rejection, or the inferred existence of possible sexual dimorphism.

The information in these tables provides the data for the interpretations and further analyses relating to determination of sexual dimorphism in this paper.

Should Raw, Transformed, or Covariate-adjusted Data be Used?

In previous studies of size dimorphism for variables other than SVL, differing types of data have been used: (1) raw versus transformed data; and, (2) raw measures versus ratios of the measures.

In general, morphological measurements on frogs are ratio-scaled (i.e. there is an absolute zero value) and continuous so that results of tests for normality and variance homogeneity in the population show that the raw, untransformed measurement data can be used for general linear modeling. Although tests that reject these assumptions can be found in the literature, they are sample-based. It is actually not appropriate to base tests only on small field samples, especially samples that are unrepresentative of the population. The assumptions concern characteristics of the populations from which the samples are taken.

It is quite usual in studies of sexual dimorphism that the raw measurements are divided by a measure of overall body size before beginning hypothesis testing. For amphibian research the usual denominator of such a ratio is SVL. In turn, such ratios are either used as the variable of interest or are transformed. Across research areas the most commonly applied transformation is the logarithm (Sokal and Rohlf 1969 p. 382). The arcsine transformation has also been used for ratio transformation (e.g. Heyer 1994).

In the present study, with its emphasis on reliability of body part measurements and determination of actual magnitude of sexual differences, we also compared results of covariate – adjusted data, with SVL as the covariate. In statistical application, ANOVA treats sex as a grouping factor, whereas regression models treat sex as the variable being predicted. In this regard, ANCOVA represents a link between the two models. The ANCOVA technique allows the researcher to adjust for body size after the field sampling has been completed and the measurements made. Use of a ratio is for the purported aim of adjusting these same data and therefore ANCOVA can be seen as an alternative.

TABLE II

ANOVA and ANCOVA results for *Leptodactylus fuscus* and *L. podicipinus* measurement data. Meas. = measurement. All mean difference values are positive; female values are greater than male values.

***Leptodactylus fuscus* – ANOVA for entire species sample**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	240	202	ns	0.007	0.407	0.663	1.44	0.5	55.3	56.3
Head length	239	202	0.013	0.014	0.703	0.358	0.30	1.2	20.3	20.3
Head width	239	202	ns	0.002	0.132	0.098	0.75	0.1	18.3	18.8
Head area	239	202	0.028	0.073	0.599	35.211				
Eye-nostril distance	21	20	ns	0.036	0.220	0.074	0.42	0.2	4.4	4.6
Tympanum diameter	21	20	ns	0.017	0.127	0.054	0.33	0.2	3.7	3.6
Thigh	239	201	0.000	0.034	0.974	0.748	1.14	0.6	25.6	26.2
Shank	239	201	0.001	0.029	0.898	0.806	0.46	1.8	32.1	32.0
Foot	239	201	0.009	0.015	0.739	0.620	1.38	0.4	30.9	30.8

***L. fuscus* – ANCOVA for entire species sample**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	239	202	0.035	0.010	0.560
Head width	239	202	ns	0.005	0.304
Head area	239	202			
Eye-nostril distance	21	20	ns	0.053	0.295
Tympanum diameter	21	20	ns	0.037	0.220
Thigh	239	201	0.000	0.046	0.995
Shank	239	201	0.000	0.028	0.941
Foot	239	201	0.039	0.010	0.542

The statistical assumption of ANCOVA that the regression data be linear is not violated by frog data, because the adults we measure do not exhibit allometry in size as static individuals. In fact, there is no indication of allometry for juveniles and adults in *L. knudseni*, a species for which allometry in head width was anticipated by WRH (Fig. 5). Rather than simple division to form a ratio (a quantity with known properties that disallow the use of parametric linear models in general) ANCOVA provides statistical control whereby the influence of the covariate is removed from the comparison on the measurement of interest.

Table VI, which provides results for an example species *L. podicipinus*, illustrates that regardless of transformation, or of body measurement considered, when we compare the ratio results we find that the effect sizes and power of the tests are virtually identical. From a standpoint of detectable male-female difference these results are equivalent as well. When results on the raw data are compared with ratio results it is clear that in general the division by body size changes and often greatly reduces the observed effect size from that seen with the raw measure. We therefore present our results for both ANOVA and ANCOVA on the raw measures only.

TABLE II (continuation)

***L. fuscus* – ANOVA for Porto Velho, Brazil sample**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	104	102	0.009	0.047	.886	0.906	1.37	0.7	43.7	44.2

***Leptodactylus podicipinus* – ANOVA for entire species sample**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	419	528	0.000	0.324	1.00	4.855	1.33	3.6	43.3	54.3
Head length	419	528	0.000	0.309	1.00	1.390	0.45	3.1	14.8	17.4
Head width	419	528	0.000	0.277	1.00	1.343	0.69	2.0	13.9	19.0
Head area	419	528	0.000	0.303	1.00	191.568				
Eye-nostril distance	21	21	0.000	0.565	1.00	0.638	0.38	1.7	3.8	4.6
Tympanum diameter	419	528	0.000	0.174	1.00	0.255	0.28	0.9	3.3	3.7
Thigh	419	528	0.000	0.196	1.00	1.428	1.05	1.4	17.1	19.7
Shank	419	528	0.000	0.253	1.00	1.501	0.42	3.6	17.7	20.2
Foot	419	528	0.000	0.285	1.00	1.790	1.32	1.4	21.1	23.9

***L. podicipinus* – ANCOVA for entire species sample**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	419	528	0.001	0.012	0.93
Head width	419	528	ns	0.000	0.06
Head area	419	528	0.044	0.004	0.52
Eye-nostril distance	21	21	0.023	0.126	0.64
Tympanum diameter	419	528	0.021	0.006	0.63
Thigh	419	528	ns	0.001	0.13
Shank	419	528	ns	0.001	0.20
Foot	419	528	0.015	0.006	0.68

Measurement Error Screening of Statistically Significant Results

For SVL, there are three additional aspects of the data that address the stability and reliability of statistically significant results for a test of sexual dimorphism: (1) measurement error, (2) maximum specimen sizes, and (3) corresponding size differences in the other variables. To evaluate the influence of measurement error on dimorphism test results, we use a simple index of measurement error calculated as the mean difference determined between males and females for the variable involved, divided by the mean measurement error as determined by the regression formulae in Table I. A value of 1 indicates that the degree of the measurement error is of the same magnitude as the observed mean differences. Values in the range of 0.7 or less indicate that the measurement error is much larger than the observed measurement differences between males

TABLE II (continuation)

***L. podicipinus* – ANOVA for Alejandra, Bolivia sample**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	38	41	0.000	0.697	1.00	1.492	1.43	1.0	43.3	47.9
Head length	38	41	0.000	0.734	1.00	1.704	0.31	5.5	14.4	16.1
Head width	38	41	0.000	0.646	1.00	1.529	0.74	2.1	13.9	15.5
Head area	38	41	0.000	0.714	1.00					
Eye-nostril distance	0	0								
Tympanum diameter	38	41	0.000	0.522	1.00	0.310	0.31	1.0	3.2	3.5
Thigh	38	41	0.000	0.514	1.00	1.522	1.52	1.3	16.9	18.1
Shank	38	41	0.000	0.595	1.00	1.492	1.49	3.3	17.1	18.7
Foot	38	41	0.000	0.601	1.00	1.801	1.80	1.3	20.8	22.8

***L. podicipinus* – ANCOVA for Alejandra, Bolivia sample**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	38	41	0.000	0.155	0.958
Head width	38	41	ns	0.032	0.345
Head area	38	41	0.003	0.107	0.846
Eye-nostril distance	0	0			
Tympanum diameter	38	41	ns	0.004	0.085
Thigh	38	41	ns	0.000	0.054
Shank	38	41	ns	0.017	0.202
Foot	38	41	0.013	0.079	0.714

and females and that any statistically significant results are probably spurious. Values in the range of 2 or greater indicate that measurement error is not influencing variability of the effect size. Maximum size data have been addressed in preceding examples (Materials and Methods: Evaluation of Statistical Methods). When SVLs differ between sexes, one would expect that overall size difference would be evidenced in ANOVA results with most or all of the other variables. When these three criteria are used to supplement and assess the robustness of the statistically significant results for SVL differences between males and females, the following species are considered to not demonstrate meaningful differences in SVL with the available data: *L. troglodytes* (measurement error index is equivocal; the maximum sizes of males and females are virtually identical; and 6 of 8 other variables do not differ in the ANOVA analyses, Table III) and Middle American pentadactylus (the measurement index is very small; the maximum female size is larger than the maximum male size, but not impressively so; and 6 out of 8 of the other variables do not differ in the ANOVA analyses, Table IV).

The following two species results are equivocal concerning whether the statistically significant differences in SVL are meaningful: *L. bufonius* (the measurement error is moderate; maximum size differences between males and females are small relative to the mean differences; and the ANOVA results for the other variables support the statistical results, Table III) and *L. fuscus* from Porto Velho (the mea-

TABLE II (continuation)

***Leptodactylus podicipinus* – ANOVA for Curuçá, Brazil sample**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	32	19	0.000	0.480	1.00	6.204	1.31	4.7	36.4	43.7
Head length	32	19	0.000	0.456	1.00	1.857	0.51	3.6	13.4	15.1
Head width	32	19	0.000	0.480	1.00	2.070	0.66	3.1	12.3	14.9
Head area	32	19	0.000	0.488	1.00					
Eye-nostril distance	17	6	0.001	0.408	0.95	0.562	0.37	1.5	3.8	4.1
Tympanum diameter	32	19	0.000	0.340	1.00	0.340	0.27	1.3	2.9	3.2
Thigh	32	19	0.000	0.363	1.00	1.993	1.02	2.0	16.4	18.2
Shank	32	19	0.000	0.463	1.00	2.127	0.40	5.3	16.0	18.0
Foot	32	19	0.000	0.469	1.00	2.454	1.31	1.9	19.5	21.6

***L. podicipinus* – ANCOVA for Curuçá, Brazil sample**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	32	19	ns	0.000	0.051
Head width	32	19	ns	0.012	0.119
Head area	32	19	ns	0.019	0.159
Eye-nostril distance	17	6	ns	0.097	0.288
Tympanum diameter	32	19	ns	0.004	0.073
Thigh	32	19	ns	0.045	0.313
Shank	32	19	ns	0.011	0.110
Foot	32	19	ns	0.049	0.337

surement error index is borderline; the maximum size differences between males and females is small relative to the mean differences; [no data available for other variables], Table II).

To assess the biological implications of the ANCOVA results that are statistically significant, only one of the three criteria described above can be applied, namely the measurement error index. Using the measurement error index criterion, the following statistically significant results are not considered to be meaningful with the available data: *L. bufonius* foot length (Table III), *L. fuscus* thigh length (Table II), foot length (Table II), Middle American pentadactylus head length (Table IV), head width (Table IV), *L. troglodytes* head length (Table III), shank length (Table III), foot length (Table III).

Effect Size as a Conveyor of Biological Meaning

Testing for statistical rejection of the null hypothesis is a necessary first step for scientific investigation, even though it provides little practical biological information about the parameters that demonstrate statistical significance. There are two additional problems to consider when testing for sexual dimorphism: (1) the magnitude of the difference that we are trying to detect (or define), and (2) the size of the sample. Clearly the column of raw mean differences contains values that are both sample and sample size dependent as well as being variable and noncomparable across species, localities or subgroups. Therefore, we require a method for comparing sexual differences that is ''dimensionless'' in the sense

TABLE II (continuation)

L. podicipinus **– ANOVA for Porto Velho, Brazil sample**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	81	113	0.000	0.403	1.00	4.776	1.34	3.6	38.2	47.6
Head length	81	113	0.000	0.328	1.00	1.271	0.44	2.9	14.1	16.1
Head width	81	113	0.000	0.369	1.00	1.350	0.69	2.0	12.9	16.1
Head area	81	113	0.000	0.354	1.00					
Eye-nostril distance	0	0								
Tympanum diameter	81	113	0.000	0.172	1.00	0.228	0.29	0.8	2.9	3.6
Thigh	81	113	0.000	0.242	1.00	1.654	1.06	1.6	16.4	18.8
Shank	81	113	0.000	0.347	1.00	1.562	0.42	3.7	15.9	19.1
Foot	81	113	0.000	0.385	1.00	1.792	1.33	1.4	19.1	22.3

L. podicipinus **– ANCOVA for Porto Velho, Brazil sample**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	81	113	ns	0.000	0.050
Head width	81	113	ns	0.008	0.246
Head area	81	113	ns	0.001	0.064
Eye-nostril distance	0	0			
Tympanum diameter	81	113	0.016	0.030	0.678
Thigh	81	113	ns	0.000	0.053
Shank	81	113	ns	0.001	0.065
Foot	81	113	0.030	0.024	0.586

of a correlation coefficient or normal deviate.

The columns labeled ‘‘effect size’’ in Tables II-V contain values that are standardized and tell the researcher how much sexual difference actually exists. This measure quantifies the magnitude of the difference between the sexes. The division of the mean difference by the standard deviation standardizes the difference between the male and female means and puts the difference on a scale that is adjusted for the standard deviation of the measure. This produces the same result as when raw scores are converted to standard-normal or z-scores. Therefore, effect size can be used to compare results from studies on different species or genera, even when unequal sample sizes are involved.

Comparing columns for p-value and effect size in Tables II-V clarifies that a statistically significant test result can obtain either (a) when sample size is excessive and effect size small, or (b) when there is small sample size and large effect size. Thus, tests of sexual dimorphism with very large sample sizes can demonstrate statistical significance, yet the raw differences involved may be biologically trivial or meaningless. The ANCOVA results for head length (with sex as the covariate) in the total sample of *L. podicipinus* provides a good example. The test result is statistically significant at the observed 0.001 level of probability (power of 0.93), yet the effect size for this variable is only 0.012 (Table II, Fig. 7A), a value so small that there is likely to be negligible biologically meaningful information in the population differences of head length between males and females.

TABLE II (continuation)

***L. podicipinus* – ANOVA for Rurrenabaque, Bolivia sample**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	28	25	0.000	0.626	1.00	5.791	1.31	4.4	36.4	42.5
Head length	28	25	0.000	0.556	1.00	1.619	0.50	3.2	13.1	14.8
Head width	28	25	0.000	0.609	1.00	1.738	0.67	2.6	12.3	14.0
Head area	28	25	0.000	0.606						
Eye-nostril distance	0	0								
Tympanum diameter	28	25	0.000	0.496	1.00	0.365	0.27	1.4	2.6	3.0
Thigh	28	25	0.000	0.445	1.00	1.549	1.02	1.5	15.5	16.2
Shank	28	25	0.000	0.597	1.00	1.762	0.40	4.4	15.0	16.2
Foot	28	25	0.000	0.630	1.00	2.359	1.31	1.8	18.3	21.0

***L. podicipinus* – ANCOVA for Rurrenabaque, Bolivia sample**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	28	25	ns	0.009	0.102
Head width	28	25	ns	0.046	0.333
Head area	28	25	ns	0.038	0.284
Eye-nostril distance	0	0			
Tympanum diameter	28	25	ns	0.018	0.157
Thigh	28	25	ns	0.009	0.103
Shank	28	25	ns	0.030	0.233
Foot	28	25	0.025	0.097	0.622

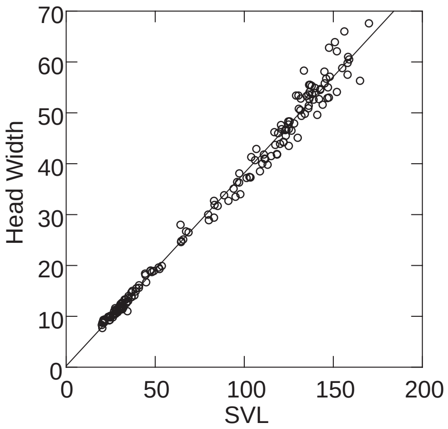

Fig. 5 – Regression of head width on SVL for juvenile and adult male specimens of *Leptodactylus knudseni*. Adjusted multiple $r^2 = 0.989$, $p = 0.000$. Values in mm.

The second sample size problem is at the other end of the spectrum. When available sample sizes are small or not representative of the population as a whole there can be interpretation problems. An example from our data is the difference in SVL length of male and female *L. pentadactylus.* The ANOVA results for SVL are significant, with the mean size of females being 13.7 mm larger than the mean size for males. However, there are two other features of the data that militate against this result being considered biologically meaningful. First, the mean measurement error is large for *L. pentadactylus*, 14.9 mm, just exceeding the mean size differences between the sexes. Second, the largest male in the sample is 195.0 mm SVL, whereas the largest female is only 174.2 mm. A plausible explanation to ac-

TABLE III

ANOVA and ANCOVA results for *Leptodactylus bufonius*, *L. troglodytes*, *L. furnarius*, and *L. gracilis* data. Meas. = measurement. Negative mean difference values (bold) indicate the male values are greater than female values.

***Leptodactylus bufonius* – ANOVA**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	106	76	0.000	0.172	1.000	2.502	1.79	1.4	59.4	61.8
Head length	106	76	0.000	0.083	0.980	0.514	0.06	0.9	21.9	21.7
Head width	106	76	0.000	0.107	0.996	0.708	0.85	0.8	20.5	21.1
Head area	106	76	0.000	0.104	0.995	120.671				
Eye-nostril distance	21	21	0.000	0.352	0.995	0.419	0.48	0.9	5.8	6.4
Tympanum diameter	21	21	0.004	0.188	0.844	0.262	0.40	0.7	5.2	5.2
Thigh	106	76	0.000	0.134	1.000	1.098	1.30	0.8	22.8	24.0
Shank	106	76	0.000	0.137	1.000	0.896	0.53	1.7	23.8	24.5
Foot	106	76	0.000	0.157	1.000	0.919	1.46	0.6	22.2	23.2

***L. bufonius* – ANCOVA**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	106	76	ns	0.000	0.058
Head width	106	76	ns	0.000	0.055
Head area	106	76	ns	0.000	0.060
Eye-nostril distance	21	21	0.003	0.202	0.866
Tympanum diameter	21	21	ns	0.057	0.321
Thigh	106	76	ns	0.020	0.476
Shank	106	76	ns	0.016	0.392
Foot	106	76	0.002	0.052	0.873

count for this apparent conflict may be found in life history information. Reproductively active male *L. pentadactylus* are territorial and apparently reside in burrows in the forest floor where a foam nest is laid and in which all larval development takes place. These males associated with burrows are extremely difficult to capture. The burrows seem to be limited resources. It is unlikely that male *L. pentadactylus* are able to excavate burrows from scratch but rather modify existing burrows made by other organisms. It is reasonable to assume that younger males that are unable to oust resident males, are more likely to be collected because they spend all their time on the forest floor. Thus, it is important to examine each statistically significant result to evaluate whether the results are biologically meaningful.

An effect size has other interpretations that make it superior to a *p*-value as an aid in evaluating sexual dimorphism.

First, effect size is the extent to which the populations of the two sexes do not overlap (Table VII). That is, if there were no overlap at all (or 100% non-overlap), then every single female would be larger than every single male, or vice versa. Surely we would agree to a conclusion of dimorphism at this level. The largest non-overlap value we ob-

TABLE III (continuation)

***Leptodactylus troglodytes* – ANOVA**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	40	26	0.012	0.094	0.717	1.386	1.62	0.9	52.8	52.7
Head length	40	26	ns	0.008	0.112	**–0.014**	0.15	0.1	19.8	20.0
Head width	40	26	ns	0.025	0.244	0.261	0.81	0.3	18.2	18.3
Head area	40	26	ns	0.000	0.051	3.351				
Eye-nostril distance	22	20	0.000	.402	0.999	0.328	0.45	0.7	5.2	5.6
Tympanum diameter	22	20	0.026	0.118	0.615	0.209	0.37	0.6	4.9	5.2
Thigh	40	26	ns	0.005	0.088	0.165	1.23	0.1	21.8	21.1
Shank	40	26	ns	0	0.050	0.014	0.50	0.0	22.8	21.0
Foot	40	26	ns	0.003	0.073	**–0.102**	1.42	0.1	21.4	20.5

***L. troglodytes* – ANCOVA**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	40	26	0.000	0.207	0.979
Head width	40	26	ns	0.012	0.138
Head area	40	26	0.001	0.175	0.949
Eye-nostril distance	22	20	0.001	0.260	0.950
Tympanum diameter	22	20	ns	0.028	0.178
Thigh	40	26	ns	0.032	0.293
Shank	40	26	0.006	0.113	0.798
Foot	40	26	0.003	0.136	0.873

served was for the *E. fenestratus* SVL data, Table V, for which ES = 0.834, or approximately 48% non-overlap. Alternatively, if the spread of SVL values were large and the overlap wider than the difference between average SVL values, then the effect observed would not seem to be biologically important. An ES = 0 means that the male and female distributions completely overlap, indicating there is 0% non-overlap. With zero observed non-overlap (or 100% overlap), clearly the sexes could not be dimorphic. In Tables II and VII we find that with an observed ES = 0.017 there is less than 1% non-overlap of the populations of tympanum diameters of *L. fuscus* males and females. This result obviously speaks more directly to our question of sexual dimorphism than the $p = 0.000$ that resulted, and was based to a great extent upon the large sample sizes for male and female *L. fuscus*. Note here that despite wide-held belief among many practitioners, it is clearly not true that the smaller the observed *p*-value the more dimorphism exists. Consideration of effect size illuminates this issue.

A second interpretation is that of an average percentile. For example, when ES = 0.2 (Table VII), this indicates that the mean of the males (females) is at the 15th percentile of the distribution of the females (males).

Finally, we can use an observed effect size value from our study as a comparative value with any range of effect size values defined specifically for frog species. That is, we can use frog-specific effect size values or ranges as a starting point for

TABLE III (continuation)

***Leptodactylus furnarius* – ANOVA**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	54	20	0.000	0.385	1.00	4.171	1.34	3.1	39.4	44.8
Head length	54	20	0.000	0.338	1.00	1.176	0.43	2.7	14.9	16.8
Head width	54	20	0.000	0.256	0.998	0.944	0.70	1.4	13.6	13.9
Head area	54	20	0.000	0.346	1.00	162.327				
Eye-nostril distance	34	11	0.001	0.241	0.950	0.276	0.39	0.7	4.3	4.6
Tympanum diameter	34	11	0.042	0.092	0.534	0.171	0.29	0.6	3.0	3.2
Thigh	54	20	0.000	0.267	0.999	1.932	1.06	1.8	20.5	23.1
Shank	54	20	0.000	0.286	1.00	2.393	0.42	5.7	24.8	29.2
Foot	54	20	0.000	0.196	0.985	2.059	1.33	1.5	28.3	28.6

***L. furnarius* – ANCOVA**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	54	20	ns	0.010	0.135
Head width	54	20	ns	0.005	0.088
Head area	54	20	ns	0.006	0.097
Eye-nostril distance	34	11	ns	0.045	0.280
Tympanum diameter	34	11	ns	0.075	0.439
Thigh	54	20	ns	0.000	0.050
Shank	54	20	ns	0.001	0.056
Foot	54	20	ns	0.008	0.119

our study or for computation of power of the test. Guidelines are provided and discussed below.

DISCUSSION

Small, Medium, and Large Effect Sizes for Sexual Dimorphism in Frogs

Cohen (1977) is the authority for the rationale underlying effect size usage in the behavioral sciences. In his seminal work, Cohen (1977:12) proposed: ‘‘...*as a convention*, ES [effect size] values to serve as operational definitions of the qualitative adjectives ‘small’, ‘medium’, and ‘large’.’’ He went on to clarify (p. 13): ‘‘Although arbitrary, the proposed conventions will be found to be reasonable by reasonable people. An effort was made in selecting these operational criteria to use levels of ES which (sic) accord with a subjective average of effect sizes such as are encountered in behavioral science. ‘Small’ effect sizes must not be so small that seeking them amidst the inevitable operation of measurement and experimental bias and lack of fidelity is a bootless task, yet not so large as to make them fairly perceptible to the naked observational eye. Many effects... are likely to be small effects as here defined, both because of the attenuation in validity of the measures employed and the subtlety of the issues frequently involved. In contrast, large effects must not be defined as so large that their quest by statistical methods is wholly a labor of supererogation, or to use Tukey’s delightful term, ‘statistical sanctification’. That is, the difference in size between apples and

TABLE III (continuation)

***Leptodactylus gracilis* – ANOVA**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	44	33	ns	0.020	0.228	1.245	1.44	0.9	52.7	50.7
Head length	44	33	ns	0.020	0.229	0.389	0.30	1.3	19.2	18.5
Head width	44	33	ns	0.009	0.127	0.270	0.75	0.4	17.1	15.8
Head area	44	33	ns	0.012	0.152	48.192				
Eye-nostril distance	29	21	0.012	0.125	0.727	0.289	0.42	0.7	4.9	5.2
Tympanum diameter	28	21	0.019	0.111	0.660	0.227	0.32	0.7	3.6	3.6
Thigh	44	33	ns	0.006	0.104	0.414	1.14	0.4	26.1	25.0
Shank	44	33	ns	0.006	0.102	0.509	0.46	1.1	30.2	30.8
Foot	44	33	ns	0.028	0.308	0.880	1.38	0.6	30.4	29.9

***Leptodactylus gracilis* – ANCOVA**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	44	33	ns	0.001	0.056
Head width	44	33	ns	0.009	0.127
Head area	44	33	ns	0.007	0.114
Eye-nostril distance	29	21	ns	0.011	0.110
Tympanum diameter	28	21	ns	0.002	0.058
Thigh	44	33	ns	0.004	0.080
Shank	44	33	ns	0.006	0.102
Foot	44	33	ns	0.009	0.132

pineapples is of an order that hardly requires an approach via statistical analysis. On the other side, it cannot be defined so as to encroach on a reasonable range of values called medium.'' Cohen's (1977) characterizations of small, medium, and large effect sizes have become the standards used subsequently in the behavioral and most other sciences. However, as early as 1982 Cohen and colleagues (Welkowitz et al. 1982:220) explicitly stated that their values defining small, medium, and large not be used as conventions ''if you can specify [effect size] values that are appropriate to the specific problem or field of research.'' To our knowledge, conventions for small, medium, and large effect sizes have not been established for measurement data used to evaluate sexual dimorphism in frogs.

For ANOVA and ANCOVA, Cohen (1977:285-287) defined a small effect size as 0.10, a medium effect size as 0.25, and a large effect size as 0.40. The nature of our data indicates that it is inappropriate to use a single definition of effect size for all frog body measurement variables.

Sexual dimorphism of overall size, as reflected by SVL in our data, can fit into the category of ''statistical sanctification'' cited above. That is, in some species of frogs, the males are very much smaller than the females – no statistical analyses are necessary to demonstrate what is obvious from visual inspection. In order to know what such large effect size values would be, we included the data on *Eleutherodactylus fenestratus*, in which there is a gap, or no evidence of overlap, in the SVL mea-

TABLE IV

Sexual dimorphism statistics for *Leptodactylus knudseni*, Middle American pentadactylus, and *L. pentadactylus* measurement data. Meas. = measurement. Negative mean difference values (bold) indicate that male mean measurements are greater than female values.

***Leptodactylus knudseni* – ANOVA**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	78	37	ns	0	0.055	0.679	11.42	0.1	170.0	154.0
Head length	78	37	ns	0	0.054	**–0.230**	1.08	0.2	58.8	55.6
Head width	78	37	ns	0.005	0.121	**–1.075**	1.64	0.7	67.6	58.8
Head area	78	37	ns	0.004	0.098	**–410.36**				
Eye-nostril distance	78	37	ns	0	0.054	**–0.062**	0.95	0.1	15.7	14.6
Tympanum diameter	77	32	ns	0.002	0.069	0.118	0.95	0.1	12.5	11.4
Thigh	78	37	ns	0.003	0.093	**–0.901**	2.47	0.4	70.6	62.9
Shank	78	37	ns	0	0.055	**–0.303**	1.08	0.3	69.8	66.4
Foot	76	37	ns	0.001	0.065	0.523	2.08	0.2	72.2	67.9

***L. knudseni* – ANCOVA**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	78	37	ns	0.008	0.159
Head width	78	37	0.006	0.065	0.792
Head area	78	37	0.026	0.043	0.608
Eye-nostril distance	78	37	ns	0.008	0.158
Tympanum diameter	77	32	ns	0.007	0.142
Thigh	78	37	ns	0.031	0.466
Shank	78	37	ns	0.012	0.215
Foot	76	37	ns	0.000	0.053

surements between the males and females. To be useful, an effect size should represent the smallest effect that would be of substantive (biological) significance to the researcher. That is, not every possible non-zero difference is important. For example, the mean raw or unstandardized difference for *L. knudseni* between the sexes' SVLs is only about 0.7 mm and the test result (at negligible .055 power, ES = 0.000) was not significant (Table IV). If one were to decide that a difference of about this magnitude could be important, then with the same means (132.05 and 131.37) and standard deviations (11.00 and 17.57), it would take about 18,800 specimens of *L. knudseni* to attain statistical significance. This sample size would provide about 80% power to detect such an 'important' difference. The selection of critical differences must have some better and more realistic basis than merely selecting any non-zero value that arises. Based on the range of standardized effect size values in our data (Fig. 6A), we propose that appropriate effect size conventions for evaluating hypotheses with SVL data are small = 0.20, medium = 0.45, and large = 0.70.

Effect size values for the ANCOVA results for

TABLE IV (continuation)

Middle American pentadactylus – ANOVA

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	75	74	0.032	0.031	0.574	3.882	12.14	0.3	156.3	164.1
Head length	75	74	ns	0.005	0.134	0.556	1.10	0.5	59.0	59.6
Head width	75	74	ns	0.001	0.061	0.241	1.67	0.1	62.7	65.1
Head area	75	74	ns	0.002	0.089	218.356				
Eye-nostril distance	75	74	0.043	0.028	0.527	0.355	0.97	0.3	14.5	15.3
Tympanum diameter	30	44	ns	0.110	0.137	0.181	0.97	0.2	10.8	11.2
Thigh	75	74	ns	0.007	0.178	0.877	2.53	0.4	67.7	68.9
Shank	75	74	ns	0.024	0.466	1.309	1.10	1.2	67.8	69.7
Foot	74	71	0.037	0.030	0.547	1.546	2.11	0.7	70.7	73.0

Middle American pentadactylus – ANCOVA

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	75	74	0.009	0.046	0.746
Head width	75	74	0.000	0.103	0.983
Head area	75	74	0.000	0.086	0.957
Eye-nostril distance	75	74	ns	0.002	0.075
Tympanum diameter	30	74	ns	0.000	0.050
Thigh	75	74	ns	0.006	0.152
Shank	75	74	ns	0.000	0.050
Foot	74	71	ns	0.004	0.125

the variables other than SVL extend over a much smaller range and would be expected to do so, since means are adjusted. In no case is a statistically significant ANCOVA result for effect size obvious to the eye for the specimens themselves. To interpret effect size values for ANCOVA analyses, it is useful visually to examine the data over the range of values obtained in this study (Fig. 6B). Figure 7A shows an example for which large sample size induces a statistically significant result for a very small effect size, which can readily be interpreted as not having biological significance. The graphs for the largest ANCOVA effect sizes for our data (Fig. 7E, F) demonstrate differences that probably do have biological meaning. Given the small number of ANCOVA significant effect size results we have in our study, as a first approximation, we propose adopting Cohen's conventions, namely small = 0.10, medium = 0.25, large = 0.40 for the ANCOVA-based effect sizes.

We emphasize that the effect size characterizations we propose are just that – proposals. The actual characterizations should come from testing our proposals against multiple frog measurement data sets before being adopted as conventions.

Comparison with Previously Published Results

Previous analyses of sexual dimorphism in the morphologically similar species *L. bufonius* and *L. troglodytes* indicated differences in sexual dimorphism in the majority of the measurement variables analyzed. Both species are stocky and short legged.

TABLE IV (continuation)

***Leptodactylus pentadactylus* – ANOVA**

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	26	26	0.033	0.160	0.856	13.731	14.92	0.9	195.0	174.2
Head length	26	26	0.013	0.118	0.718	4.204	1.16	3.6	67.9	67.1
Head width	26	26	0.014	0.114	0.701	4.419	1.80	2.5	75.1	71.2
Head area	26	26	0.018	0.107	0.671					
Eye-nostril distance	26	26	0.039	0.083	0.548	0.812	1.05	0.8	16.0	16.6
Tympanum diameter	26	26	0.043	0.080	0.532	0.588	1.06	0.6	13.7	11.6
Thigh	26	26	0.002	0.178	0.897	6.015	2.72	2.2	80.8	78.9
Shank	26	26	0.000	0.224	0.961	5.715	1.19	4.8	76.6	77.5
Foot	26	26	0.001	0.215	0.952	5.885	2.21	2.7	82.1	80.6

***L. pentadactylus* – ANCOVA**

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	26	26	ns	0.001	0.057
Head width	26	26	ns	0.002	0.062
Head area	26	26	ns	0.005	0.077
Eye-nostril distance	26	26	ns	0.014	0.127
Tympanum diameter	26	26	ns	0.002	0.062
Thigh	26	26	ns	0.025	0.197
Shank	26	26	0.039	0.084	0.549
Foot	26	26	ns	0.068	0.458

In the previous study (Heyer 1978), *L. bufonius* demonstrated sexual dimorphism in SVL (females larger), head length (male heads longer), head width (male heads wider), whereas *L. troglodytes* demonstrated sexual dimorphism in head length (male heads longer), shank length (male shanks longer), and foot length (male feet longer). In this study, both *L. bufonius* and *L. troglodytes* demonstrate statistically significant differences in female-male SVL, but SVL differences are considered not meaningful and can not be demonstrated to be dimorphic for *L. troglodytes* with the available data. The effect size for SVL in *L. bufonius* is 0.172, a small effect size as defined herein. Head length is not sexually dimorphic for *L. bufonius* as analyzed herein. Although head length is statistically significant for *L. troglodytes* in our results, it is considered to be not meaningful due to the large measurement error relative to the actual measurement differences between the sexes (measurement error index = 0.1). Head width is not statistically different between males and females in our results for *L. bufonius*. For both shank and foot lengths, the ANCOVA results are statistically significant for *L. troglodytes* but are considered not meaningful due to the large measurement errors relative to actual measurement differences in the available data (measurement error index = 0.0, 0.1 respectively). Foot length dimorphism in *L. bufonius* is statistically significant but is considered not meaningful, also due to large measurement errors relative to actual measurement differences (measurement error index = 0.6). Our

TABLE V

Sexual dimorphism statistics for *Eleutherodactylus fenestratus* measurement data. Meas. = measurement. All mean difference values are positive; female values are greater than male values.

ANOVA

Variable	N		*p*	Effect size	Observed power	Mean difference	Mean meas. error	Meas. error index	Maximum	
	♂	♀							♂	♀
SVL	11	14	0.000	.834	1.00	12.729	1.36	9.3	34.5	52.3
Head length	11	14	0.000	.817	1.00	5.071	0.39	12.9	13.9	20.9
Head width	11	14	0.000	.809	1.00	4.650	0.71	6.6	12.8	18.4
Head area	11	14	0.000	.796	1.00					
Eye-nostril distance	11	14	0.000	.811	1.00	1.659	0.39	4.2	4.8	6.7
Tympanum diameter	11	14	0.000	.723	1.00	0.848	0.30	2.8	2.6	3.6
Thigh	11	14	0.000	.852	1.00	6.875	1.08	6.3	17.1	26.5
Shank	11	14	0.000	.872	1.00	7.702	0.43	17.8	19.1	29.2
Foot	11	14	0.000	.786	1.00	5.966	1.34	4.4	17.9	26.2

ANCOVA

Variable	N		*p*	Effect size	Observed power
	♂	♀			
Head length	11	14	ns	.003	0.057
Head width	11	14	ns	.003	0.057
Head area	11	14	ns	.047	0.169
Eye-nostril distance	11	14	ns	.042	0.155
Tympanum diameter	11	14	ns	.002	0.005
Thigh	11	14	ns	.131	0.413
Shank	11	14	0.018	.228	0.684
Foot	11	14	ns	.000	0.050

results indicate that there is only one sexual difference that passes the first two steps of our protocol, namely SVL in *L. bufonius*, but the magnitude of the observed effect is small, hence not biologically meaningful.

Leptodactylus furnarius and *L. gracilis* are both gracile, long-legged species, but are readily morphologically distinguishable from each other whereas *L. bufonius* and *L. troglodytes* are difficult at best to tell apart morphologically. In a previous study (Heyer 1978), *L. furnarius* (as *L. laurae*) demonstrated sexual dimorphism in SVL (females larger) and no dimorphism in thigh, shank, or foot length; *L. gracilis* demonstrated dimorphism only for head width (male heads longer) for the variables analyzed. Our results demonstrate statistically significant results solely for SVL in *L. furnarius* (females larger), with an effect size of 0.385, a medium effect size as defined herein.

There are two differences between the previous and current studies involving *L. bufonius*, *L furnarius*, *L. gracilis*, and *L. troglodytes*. First, the earlier study employed t-tests and the analysis of ratio data for all variables other than SVL, while ANOVA (equivalent to *t*-test) for SVL and ANCOVA for untransformed variable data were used in this study.

TABLE VI

Comparison of ANOVA and ANCOVA analyses on raw and transformed data for *Leptodactylus podicipinus*, N = 528 females, 419 males for all variables except EN, N = 21 females and males.

ANOVA – Raw data

Variable	*p*	Effect size	Power	Levene's test df	Levene's test *p*	♀ mean	♀ sd	♂ mean	♂ sd
SVL	0.000	0.324	1.00	1,945	0.000	37.69	3.129	31.43	3.215
HL	0.000	0.309	1.00	1,945	0.000	13.53	0.959	11.49	1.108
HW	0.000	0.277	1.00	1,945	0.000	12.59	1.009	10.57	1.057
H area	0.000	0.303	1.00	1,945	0.000	951.92	139.566	682.77	130.170
EN	0.000	0.565	1.00	1,40	0.000	3.97	0.256	3.33	0.315
TD	0.000	0.174	1.00	1,945	0.000	2.78	0.251	2.40	0.257
Thigh	0.000	0.196	1.00	1,945	0.000	15.43	1.145	13.83	1.616
Shank	0.000	0.253	1.00	1,945	0.000	15.87	1.062	13.76	1.259
Foot	0.000	0.285	1.00	1,945	0.000	19.30	1.209	16.68	1.596

ANOVA – Untransformed ratios

Variable	*p*	Effect size	Power	Levene's test df	Levene's test *p*	♀ mean	♀ sd	♂ mean	♂ sd
HL/SVL	0.000	0.073	1.00	1,945	ns	0.352	0.016	0.362	0.017
HW/SVL	0.000	0.069	1.00	1,945	ns	0.329	0.013	0.337	0.014
H area/SVL	0.000	0.193	1.00	1,945	0.011	24.998	2.153	22.992	1.892
EN/SVL	ns	0.003	0.06	1,40	ns	0.106	0.006	0.106	0.005
TD/SVL	0.000	0.069	1.00	1,945	ns	0.073	0.005	0.076	0.005
Thigh/SVL	0.000	0.062	1.00	1,945	ns	0.394	0.027	0.408	0.028
Shank/SVL	0.000	0.118	1.00	1,945	ns	0.409	0.019	0.423	0.019
Foot/SVL	0.000	0.104	1.00	1,945	ns	0.497	0.027	0.515	0.025

ANOVA – Arcsine transformed ratios

Variable	*p*	Effect size	Power	Levene's test df	Levene's test *p*	♀ mean	♀ sd	♂ mean	♂ sd
HL/SVL	0.000	0.073	1.00	1,945	ns	0.361	0.017	0.371	0.018
HW/SVL	0.000	0.069	1.00	1,945	ns	0.336	0.014	0.344	0.015
H area/SVL	0.000	0.193	1.00	1,945	ns				
EN/SVL	ns	0.003	0.06	1,40	ns	0.106	0.006	0.106	0.005
TD/SVL	0.000	0.069	1.00	1,945	ns	0.073	0.005	0.076	0.006
Thigh/SVL	0.000	0.062	1.00	1,945	ns	0.405	0.030	0.421	0.030
Shank/SVL	0.000	0.118	1.00	1,945	ns	0.422	0.021	0.432	0.021
Foot/SVL	0.000	0.104	1.00	1,945	ns	0.520	0.031	0.541	0.029

TABLE VI (continuation)

ANOVA – Log transformed ratios

Variable	*p*	Effect size	Power	Levene's test df	Levene's test *p*	♀ mean	♀ sd	♂ mean	♂ sd
HL/SVL	0.000	0.073	1.00	1,945	ns	–1.043	0.046	–1.016	0.047
HW/SVL	0.000	0.070	1.00	1,945	ns	–1.111	0.040	–1.089	0.041
H area/SVL	0.000	0.196	1.00	1,945	ns	3.215	0.085	3.132	0.082
EN/SVL	ns	0.003	0.06	1,40	ns	–2.250	0.055	–2.245	0.046
TD/SVL	0.000	0.068	1.00	1,945	ns	–2.624	0.070	–2.586	0.072
Thigh/SVL	0.000	0.061	1.00	1,945	ns	–0.934	0.070	–0.898	0.069
Shank/SVL	0.000	0.118	1.00	1,945	ns	–0.895	0.047	–0.861	0.045
Foot/SVL	0.000	0.104	1.00	1,945	0.024	–0.701	0.054	–0.665	0.048

ANCOVA – Raw data, by sex, SVL as covariate

Variable	Model *p*	Model effect size	Model power	Sex *p*	Sex effect size
HL	0.000	0.831	1.00	0.001	0.012
HW	0.000	0.866	1.00	ns	0.000
H area	0.000	0.872	1.00	0.044	0.004
EN	0.000	0.867	1.00	0.023	0.126
TD	0.000	0.638	1.00	0.021	0.006
Thigh	0.000	0.643	1.00	ns	0.001
Shank	0.000	0.820	1.00	ns	0.001
Foot	0.000	0.784	1.00	0.015	0.024

df = degrees of freedom; EN = eye-nostril distance; H area = head area; HL = head length; HW = head width; sd = standard deviation; SVL = snout-vent length; TD = tympanum diameter.

Second, the data sets analyzed herein are larger because measurement data were added for each species over the years between the studies. Given these differences, one would not expect the results to be the same between the studies. Overall, the statistically significant results between the studies are quite similar. The major differences between the studies lie in the variables considered to be biologically meaningful based on effect sizes and measurement error relative to the magnitude of the mean differences in the variables between females and males – in these terms, the results of the two studies are quite different.

Data for SVL, head length, head width, eye-nostril distance, tympanum diameter, thigh length, shank length, and foot length were analyzed previously for *L. knudseni*, *L. pentadactylus*, and Middle American pentadactylus (Heyer 2005). As for the Heyer (1994) study, the data were analyzed using t-tests and for all variables other than SVL, arcsine transformed ratio data were used. Although the arcsine is not the most appropriate transformation, almost identical effect size results obtain if the more appropriate log transformations or the untransformed ratios are used (Table VI) as we have mentioned. Sample sizes are identical for the previous and current analyses for these three species. In the previous study (Heyer 2005), *L. knudseni* demonstrated statistically significant differences only in head width (male heads wider); *L. pentadacty-*

TABLE VII

Interpretations of Effect Size values. Percent non-overlap is the amount of overlap between two groups: An Effect Size = 0.0 indicates that the distribution of the female measurement data totally overlaps that for the males, i.e., 100% overlap or 0% non-overlap. Percentile standing: The percentage of the female population data that the upper half of the male population data exceeds, i.e., Effect Size = 0.0 indicates that the mean of the female data is at the 50th percentile of the male data and an Effect Size = 0.8 indicates that the mean of the females is at the 79th percentile of the male distribution.

Cohen's convention = proposed ANCOVA convention	Proposed SVL convention	Effect size	% non-overlap	Percentile standing
		2.0	81.1	97.7
		1.9	79.4	97.1
		1.8	77.4	96.4
		1.7	75.4	95.5
		1.6	73.1	94.5
		1.5	70.7	93.3
		1.4	68.1	91.9
		1.3	65.3	90.0
		1.2	62.2	88.0
		1.1	58.9	86.0
		1.0	55.4	84.0
		0.9	51.6	82.0
		0.8	47.4	79.0
	large	0.7	43.0	76.0
		0.6	38.2	73.0
		0.5	33.0	69.0
	medium	0.45	30.0	68.0
large		0.4	27.4	66.0
		0.3	21.3	62.0
medium		0.25	18.1	60.2
	small	0.2	14.7	58.0
small		0.1	7.7	54.0
		0.0	0.0	50.0

lus for SVL (females larger) and eye-nostril distance (male distances longer); and Middle American pentadactylus for SVL (females larger), head length (male heads longer), and head width (male heads wider). The results from this study are exactly the same for *L. knudseni*. The statistical results are the same for Middle American pentadactylus for all variables except head length and head width for which the opposite sex demonstrated the larger variable values (females with longer and wider heads in the results in this study); effect size values for both head length (0.005) and width (0.001) are negligi-

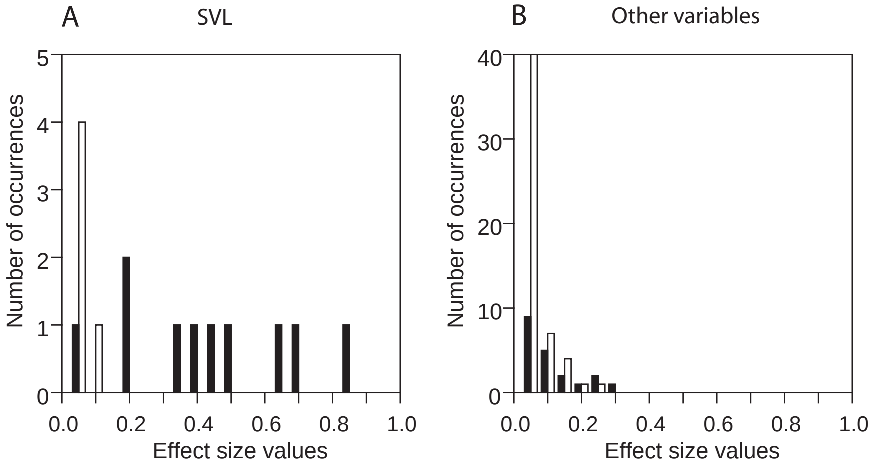

Fig. 6 – Distribution of effect size values for SVL (A) and other variables combined (head length, head width, etc.) (B). Solid bars are biologically significant values. Open bars are biologically insignificant values. Open bar on left of B has been truncated to 40 occurrences for purposes of display; the actual value is 76.

ble. Both sets of statistical results are the same for SVL in *L. pentadactylus*, but in this study there is no dimorphism for eye-nostril distance (ES = 0.002), while there is statistical support for shank length (ES = 0.224; female shank longer). The effect size for SVL dimorphism in *L. pentadactylus* is 0.160, a small effect size as defined herein, hence not biologically meaningful (also see discussion in Effect Size as a Conveyor of Biological Meaning). The statistically significant results for Middle American pentadactylus SVL (ES = 0.031), eye-nostril (ES = 0.028), and foot (ES = 0.030) in this study are considered to be biologically insignificant. As for the above previous study comparisons, the overall statistically significant results are again more similar between the studies than are the biologically significant results.

Biological Implications

As indicated in the introduction, one of the main interests in analyzing sexual dimorphism of measurement data in frogs is to gain insights to their biology. From the results discussed above, *L. furnarius* demonstrates sexual dimorphism in size (females larger) whereas *L. gracilis* does not demonstrate size dimorphism. Based upon our tests and supplemental methodology, these results are robust and most likely have an as yet undetermined biological explanation.

The lack of sexual dimorphism for SVL in *L. knudseni* is robust, whereas the results for dimorphism in SVL for Middle American pentadactylus and *L. pentadactylus* require further investigation. The lack of sexual dimorphism in size may relate to territorial defense and fighting as indicated by Shine (1979), since males are typically smaller than females in most species of frogs.

Some samples were included in this study to assess whether geographic variation may have a confounding effect when trying to understand sexual dimorphism for the species. Only SVL data were available for this aspect in *L. fuscus* (Table II). The sample size for Porto Velho is large enough that it almost certainly characterizes the range of SVL values for the species at that locality. The range of SVL values at Porto Velho is less than half the range for the species as a whole (Porto Velho male SVL range 34.2-43.7 mm, female range 34.3-44.2 mm; for en-

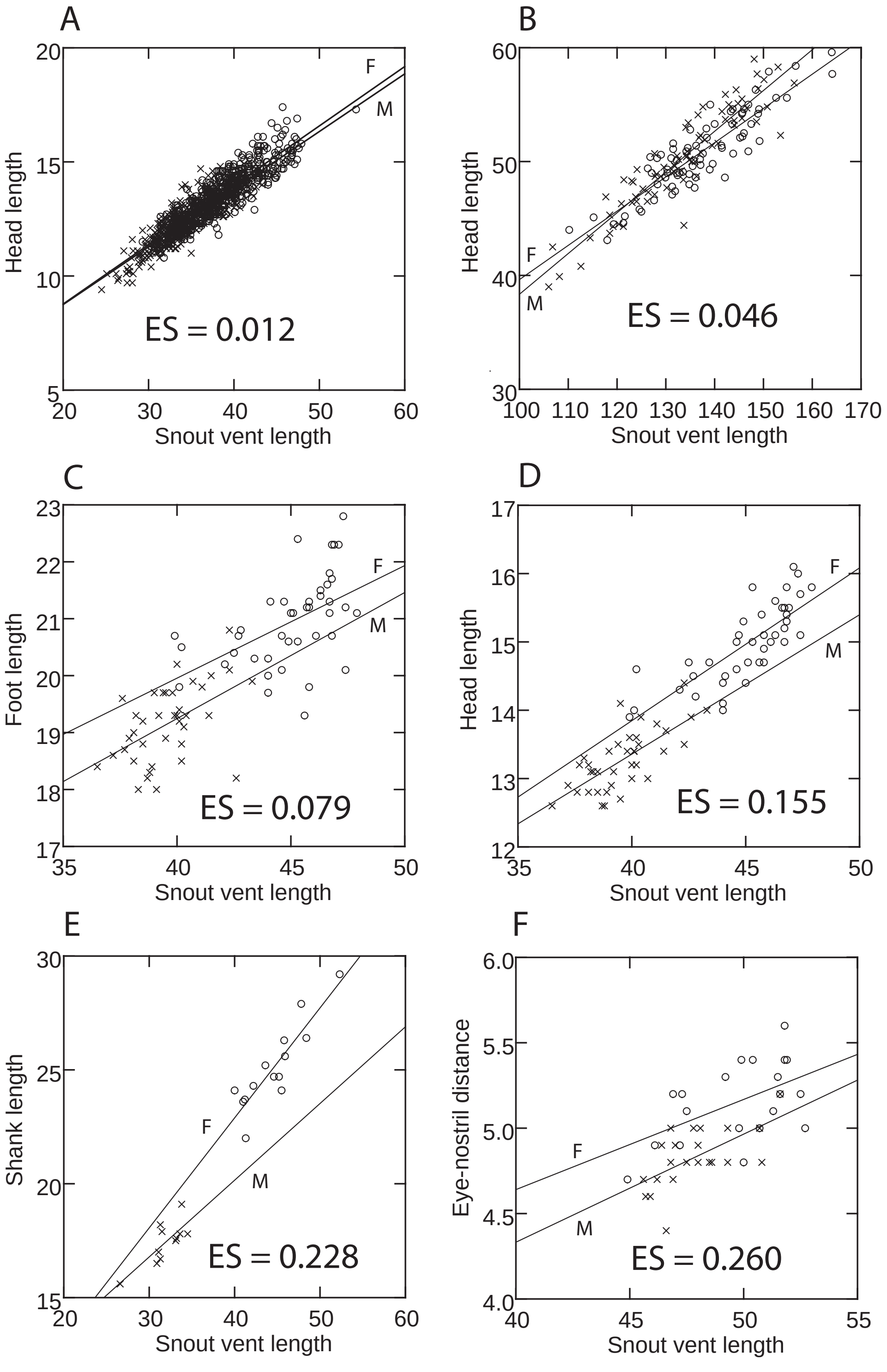

Fig. 7 – Scatterplots of variables with different effect sizes (ES). A = *Leptodactylus podicipinus* (all data), B = Middle American pentadactylus, C and D = *L. podicipinus* – Alejandra, Bolivia sample; E = *Eleutherodactylus fenestratus*, F = *L. troglodytes*. F (within figures) = female data regression line, M = male data regression line. All values in mm.

tire species sample male SVL range 32.4-55.3 mm, female range 32.2-56.3 mm). Thus, there is meaningful geographic variation in SVL that exceeds the range of intra-population variation. For the entire species sample, SVL sexual dimorphism is not statistically significant. For the Porto Velho sample, sexual dimorphism in SVL is statistically significant but the effect size is so small (ES = 0.047) that it is biologically meaningless. More data are available to address the problem of species versus population variation in sexual dimorphism for *L. podicipinus*. The sample from Porto Velho is large enough to characterize the range of measurement variables for the frogs at that site. In all cases, the sample for the entire species exceeds the ranges of values for the Porto Velho sample, but there is variation in the magnitude of the differences (Table VIII). The variation in SVL is meaningfully greater than the intra-population variation at Porto Velho, whereas the variation in tympanum diameter at Porto Velho is approximately equivalent to that observed in the data for the entire sample. Thus, the magnitude of geographic variation varies depending on the variable involved. Sexual dimorphism in SVL demonstrates similar results for the entire species sample and for each of the four individual locality samples: for the entire species sample the females are significantly larger with an effect size between small and medium (ES = 0.324); the four locality samples demonstrate that the females are also significantly larger with medium to large effect sizes (ES values range 0.403-0.697). Head length and width are both statistically greater in females in the entire species sample, with medium effect size (ES = 0.309, 0.277) and in each of the localities with medium to large effects (ES range 0.328-0.734). Despite the small sample size, eye-nostril distance is significantly different for the entire species sample with a medium-large effect (ES = 0.565). Data for eye-nostril distance are only available for the Curuçá specimens and achieve statistical significance with an ES = 0.408. Dimorphism in tympanum diameter is statistically significant for the entire species sample, but the effect size is small (ES = 0.174) and the individual localities mirror this result. The very limited data analyzed herein suggest that although there is geographic variation in the variables, the variables that show biologically meaningful sexual dimorphism show it for not only the individual locality samples, but also the entire species sample as well.

For variables other than SVL, those having at least medium effect sizes appear to be biologically meaningful (Fig. 7). The differences between male and female eye-nostril distances in *L. troglodytes* (Table III, Fig. 7F) are of a magnitude (ES = 0.260) that indicates some as yet unknown biological significance. The effect size differences between male and female shank lengths in *Eleutherodactylus fenestratus* (ES = 0.228) also appear to be biologically meaningful (Table V, Fig. 7E). In this case, a plausible biological explanation is that because the mass of females is much greater than the mass in males, females would require longer legs to jump similar distances as males.

Null Hypothesis Testing Versus Scientific Inference

The use of null hypothesis testing in the ecological literature is well established but the limitations of this approach are less well recognized. Emphases are placed on rejecting the null hypothesis and the size of the *p*-value rather than on the data and whether or not it supports the scientific contention. Null hypothesis testing is not solely a dichotomous decision on whether to reject or not. It is also a procedure that gives the researcher a method for determining whether present sample sizes are adequate or need to be increased to demonstrate meaningful statistical results. Thus, this approach should only be used in circumstances where additional data can be obtained. Indeed, there are popular and wide-spread misinterpretations of these distinctions concerning statistical and scientific results. The two errors most commonly seen in the ecological and herpetological literature are: (1) believing that the *p*-value is the probability that the null hypothesis is actually

TABLE VIII

Ranges (maximum-minimum values) of measurement variables for male and female *Leptodactylus podicipinus*. Values in mm.

	Entire sample		Porto Velho sample	
	♂	♀	♂	♀
SVL	18.8	25.0	10.1	16.6
Head length	5.4	6.6	4.4	5.3
Head width	5.4	6.9	3.9	5.5
Eye-nostril distance	1.0	1.1		
Tympanum diameter	1.6	1.7	0.9	1.6
Thigh length	6.6	9.4	5.5	7.0
Shank length	6.8	8.6	4.0	6.5
Foot length	8.0	9.7	4.1	6.9

true; and, (2) interpreting the p-value for hypothesis rejection as the probability that a substantive effect exists in the population (i.e., the smaller the p-value, the larger the biological effect).

An error of major import in herpetological studies is the equating of very small p-values with the existence of meaningful differences between the groups or species being compared (e.g., $p = 0.0001$ is a much more meaningful result than $p = 0.0499$). Although it is a necessary part of the quantitative evaluation of field results, the p-value alone cannot provide the researcher with this information. Even though the testing of the null hypotheses of no effect and the estimation the size of an effect are closely related, there has been total reliance upon the former and lack of interest in the latter in the ecological literature (e.g., Mapstone 1995).

All basic statistical texts contain a section on the inter-relationships involved in a statistical test of hypothesis. We learn that type 1 error, type 2 error, power, and sample size are all related. This information, if considered at all for analysis of field research data, urges researchers to consider the power of the test (Hayes and Steidl 1997, Peterman 1990, Reed and Blaustein 1995, Taylor and Gerrodette 1993, Toft and Shea 1983, Yezerinac et al. 1992, Zielinski and Stauffer 1996). However, in the published literature dealing with sexual dimorphism, a null hypothesis of no difference between the sexes has been set up based upon the data obtained from those specimens that were observed or captured. The refrain is commonly heard that a predetermined sample size is useless in field research because we ''obtain what we can'' given time, funding and behavioral characteristics of the species of interest. There are two facets of power: (1) prospective power, which can be used to determine what samples should be used or if sample sizes should be increased, and (2) retrospective observed power, determined after data collection, which actually can confuse without providing insight. Because retrospective or observed power is clearly a decreasing function of the observed p-value, we need only one of these quantities. The p-value is clearly the best known and more easily calculated.

Null Hypothesis Testing Versus Data Exploration

Investigation into sexual dimorphism and its correlates is a common theme in amphibian and other literatures. Such research usually focuses on size differences, but distinctions of shape and patterns of adaptation, evolutionary, or ecological influences can be of interest as well. The statistical approach is determined by the specific question being framed within the research regardless of the number of foci being considered. In the present study our interest lies in the single unambiguous question ''Is there a difference in the size of the chosen measure between males and females of a given species?'' We therefore rely upon a test of the null hypothesis of no difference within a univariate model framework. Our investigations examine a single alternative hypothesis and its relationship to a standardized measure of the definable difference between the sexes for the given variable. Alternatively, many researchers desire to incorporate related influences into their study and thus advocate multivariate methods (e.g. Butler and Losos 2002). Such an approach requires consideration of measurement data that may need to be adjusted for morphological, phylogenetic or other concerns. That is, these adjustments are beyond the usual statistical methods. Also, many researchers employ exploratory data analytic techniques; that is, techniques that explore the data rather than test a hypothesis about the data (e.g. Butler and Losos 2002, multivariate general linear models). The present study is not intended as a generalized treatment of all of the possible questions involving dimorphism. Rather, it is a unified treatment of the most effective methods that amphibian workers may use to obtain a substantively or operationally significant answer to the single question of the existence of sexual dimorphism in size variables.

Conclusions and Recommendations

We consider the concept of ''effect size'' to be an instrument for the incorporation of biological meaning into the testing methodology as well as an interrelated factor in statistical hypothesis testing. A statistical significance test merges information on size of an effect observable in the data with information on the sample size. For this reason the p-value is not the correct device for evaluating the magnitude of frog population differences. Effect size is a scale-free and standardized measure of the relative magnitude of the effect of interest, in our case sexual dimorphism. Effect size and the ability to detect it are directly related. The larger the effect size, the easier it is to detect, as demonstrated by the *Eleutherodactylus fenestratus* data (Table V). Conversely, the smaller the dimorphism effect, the more difficult it is to demonstrate, as shown with the *L. knudseni* example. A larger sample size generally leads to parameter estimates with smaller variances resulting in a statistically significant difference for small effect sizes.

Any frog study must be of an adequate sample size relative to the study's goals. Thus, the age-old problem arises of ''what should be the sample size?'' A statistically significant result can occur if either the effect size is very ''big'' (despite having a small sample), or, if the sample size is very ''big'' (despite a very small effect size). A review of Tables II-V shows that the sample should be big enough so that an effect size that is biologically interesting or important will be recognized as statistically significant. Consequently, we developed a range of ES values that have biological meaning for the species included in this study, and for other frog species as well. Use of these conventions will allow the researcher to evaluate power in past or future studies as well as to determine when the sample size for the study should be enlarged in a future project, rather than merely ignoring the results as ''non-significant'', or worse, accepting statistically significant results that are biologically trivial as biologically meaningful, and not pursuing the research any further. Sample size is important. An undersized study wastes valuable resources because it is not capable of producing useful results. A study that is too large requires greater resources and the cost benefit ratio is excessive.

In our quest for biological meaning or importance, we incorporate the concept of measurement error. It is well recognized that for many frog morphometric variables, measurement error is high (e.g. Hayek et al. 2001). Our measurement error index provides insight to the relationship of the impact of measurement error for each variable on the ability to detect meaningful sexual dimorphism in the data.

Recommendations.

1. Studies of sexual dimorphism based on measurement data should rely on more than the dichotomous decision to either accept or reject the null hypothesis of no dimorphism made on an arbitrary number of specimens for which sample sizes cannot be increased (most museum-based specimen data).

2. The results of any hypothesis test for sexual or geographic dimorphism should be supplemented with information on measurement error for the morphological variable of interest. Interpretation of effect size and results for the entire variable may be problematic if measurement error is high. Results can be evaluated by use of our measurement error index.

3. We recommend the use of effect size as the primary statistic to evaluate sexual dimorphism in measurement data. Power has been suggested as the primary statistic to evaluate biological magnitude of statistical analyses. However, others say it is not worthwhile, or it is too complicated a factor to consider and report on without a pilot study. We avoid the argument by noting that observed power increases with probability level. Thus, p-values can be used as a proxy for power for researchers who wish to compare power among studies.

4. Adequate sample size, relative to study goals, can be determined by use of effect size. The range of effect size values provided in this study will enable the researcher to determine a sample size large enough to garner statistically significant and biologically meaningful results. Alternatively, the effect size values will help the researcher identify sample sizes so large that a statistically detectable result is of no scientific importance.

5. Effect size information can be used for planning as well as synthesizing studies and their results. Use of an effect size with its confidence interval conveys the same information as the usual hypothesis test of significance, but the emphasis is on the significance of the effect or actual difference between the sexes rather than on the arbitrary sample size. Reporting and interpretation of effect sizes in addition to statistical test results is simple and more effective than other statistical approaches currently in use, particularly for field-based research that can not be controlled experimentally.

APPENDIX I

CALCULATION OF EFFECT SIZE VALUES

Cohen's original work was in the areas of psychology and education, which quite commonly deal with relationships between independent and dependent variables. In such cases an effect size is a standardized measure of the change in the dependent variable as explained by or as a result of change in the independent variable. Thus, standardization was first accomplished by dividing by σ = the standard deviation of the control or independent group. This allowed for the measurement of the effectiveness of the treatment with reference to the group not affected by that treatment.

We present results based upon general linear modeling with ANOVA and ANCOVA. Field and museum specimens used for sexual or geographic dimorphism study do not involve ''control'' and ''treatment'' or ''experimental'' groups. Therefore, we adjust the data used for the standard deviation in order to standardize our effect size. We desire to examine the difference between male and female specimens and relate this difference to, or standardize by, the within group dispersion. Selecting one of the two standard deviations would make an appreciable difference in our value of **d**, so we established a pooled estimate of the standard deviation. The formulae used were:

$$d = \frac{(m_f - m_m)}{\hat{\sigma}_{pooled}}$$

where

$$\hat{\sigma}_{pooled} = \sqrt{\frac{(n_f - 1) \cdot \hat{\sigma}_f + (n_m - 1) \cdot \hat{\sigma}_m}{(n_f + n_m - 2)}}\,.$$

The use of this pooled estimate of the standard deviation depends on the assumption that the two calculated standard deviations are estimates of the same population value, or differ only with sampling variability. This of course is the null hypothesis.

It is advantageous to use the pooled standard deviation because there is an alternative method for calculation of **d**, easily computed from computerized printouts. For ANOVA and ANCOVA: $SS_{effect}/(SS_{effect} + SS_{error})$. Also called Partial Eta Squared, this is the proportion of the effect plus error variance that is attributable to the effect of dimorphism. This is the quantity that we report in this paper.

ACKNOWLEDGMENTS

Drs. M. H. Heyer, P. E. Vanzolini, and G. R. Zug provided comments that were useful in clarifying our presentation. P.E. Vanzolini kindly prepared the Resumo. The research for this paper received support from the Smithsonian Institution, National Museum of Natural History, USA (Lee-Ann C. Hayek and W. Ronald Heyer) and the National Science Foundation, USA (award 9815787 to Rafael O. de Sá and W. Ronald Heyer).

RESUMO

Técnicas analíticas variadas têm sido usadas para avaliar o dimorfismo sexual em medidas de vertebrados, mas não há consenso sobre o melhor procedimento. Um problema adicional, no caso dos anfíbios, é a presença de ponderável erro de medida. Para analisar dimorfismo sexual examinamos uma única hipótese (*Ho* = médias iguais) para dois grupos (fêmeas e machos). Demonstramos que dados de anfíbios preenchem as premissas para hipóteses estatísticas claramente definidas, usando modelos lineares em vez de técnicas exploratórias multivaraiadas, tais como components principais, correlação ou análise de correspondências. Para distinguir significância biológica de significância estatística nas hipóteses, propomos um protocolo incorporando erro de medida e ''effect size''. O erro de medida é avaliado por meio de um novo índice específico. Demonstramos que ''effect size'', amplamente usado nas ciências do comportamento e em meta-análises biológicas, é a medida mais útil na discriminação entre significância biológica e significância estatística. São dadas definições de uma ampla gama de ''effect sizes'' para dados anfibiológicos. São apresentados exemplos com medidas do gênero *Leptodactylus*. O novo protocolo é recomenadado não apenas no caso de anfíbios, mas em todos os casos de vertebrados em que possam ser calculados erros de medida e ''effect sizes'' observados ou determinados a priori.

Palavras-chave: estatística, dimorfismo sexual, índice de erro de medida, ''effect size'', rãs.

REFERENCES

Butler MA and Losos JB. 2002. Multivariate sexual dimorphism, sexual selection, and adaptation in greater Antillean *Anolis* lizards. Ecol Monogr 72: 541–559.

Cohen J. 1977. Statistical power analysis for the behavioral sciences. Revised edition. New York, London, Toronto, Sydney, San Francisco: Academic Press, Inc., 474 p.

Duellman WE and Trueb L. 1986. Biology of amphibians. New York: McGraw-Hill Book Co., 670 p.

Hayek LC, Heyer WR and Gascon C. 2001. Frog morphometrics: A cautionary tale. Alytes 18: 153–177.

Hayes JP and Steidl RJ. 1997. Statistical power analysis and amphibian population trends. Conserv Biol 11: 273–275.

Heyer WR. 1978. Systematics of the *fuscus* group of the frog genus *Leptodactylus* (Amphibia, Leptodactylidae). Nat Hist Mus Los Angeles County Sci Bull 29: 1–85.

Heyer WR. 1994. Variation within the *Leptodactylus podicipinus-wagneri* complex of frogs (Amphibia: Leptodactylidae). Smithsonian Contrib Zool 546: 1–124.

Heyer WR. 2005. Variation and taxonomic clarification of the large species of the *Leptodactylus pentadactylus* species group (Amphibia: Leptodactylidae) from Middle America, northern South America, and Amazonia. Arq Zool. (In press).

HEYER WR AND MUÑOZ AM. 1999. Validation of *Eleutherodactylus crepitans* Bokermann, 1965, notes on the types and type locality of *Telatrema* [sic] *heterodactylum* Miranda-Ribeiro, 1937, and description of a new species of *Eleutherodactylus* from Mato Grosso, Brazil (Amphibia: Anura: Leptodactylidae). Proc Biol Soc Washington 112: 1–18.

MAPSTONE BD. 1995. Scalable decision rules for environmental impact studies: Effect size, type I and type II errors. Ecol Applic 5: 401–410.

PETERMAN RM. 1990. The importance of reporting statistical power: The forest decline and acidic deposition example. Ecology 71: 2024–2027.

REED JM AND BLAUSTEIN AR. 1995. Assessment of ''nondeclining'' amphibian populations using power analysis. Conserv Biol 9: 1299–1300.

SHINE R. 1979. Sexual selection and sexual dimorphism in the Amphibia. Copeia 1979: 297–306.

SOKAL RR AND ROHLF FJ. 1969. Biometry. The principles and practice of statistics in biological research. San Francisco: W.H. Freeman & Company, 776 p.

TAYLOR BL AND GERRODETTE T. 1993. The uses of statistical power in conservation biology: The vaquita and northern spotted owl. Conserv Biol 7: 489–500.

TOFT CA AND SHEA PJ. 1983. Detecting community-wide patterns: Estimating power strengthens statistical inference. Am Natur 122: 618–625.

WELKOWITZ J, EWEN RB AND COHEN J. 1982. Introductory statistics for behavioral sciences. Third edition. New York & other cities: Academic Press, 369 p.

WILKINSON L AND COWARD M. 2000. Linear models I: Linear regression. In: SYSTAT 10, Statistics I. Chicago: SPSS Inc, p. I-399–I-430.

YEZERINAC SM, LOUGHEED SC AND HANDFORD P. 1992. Measurement error and morphometric studies: Statistical power and observer experience. Syst Biol 41: 471–482.

ZIELINSKI WJ AND STAUFFER HB. 1996. Monitoring *Martes* populations in California: Survey design and power analysis. Ecol Applic 6: 1254–1267.