Alytes, 2001, **18** (3-4): 153-177.

Frog morphometrics: a cautionary tale

Lee-Ann C. HAYEK*, W. Ronald HEYER**[1] & Claude GASCON***

* Mathematics & Statistics, MRC 136, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0136, USA
[hayek.lee-ann@nmnh.si.edu]

** Amphibians & Reptiles, MRC 162, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0162, USA
[heyer.ron@nmnh.si.edu]

*** Field Support Program, Conservation International, 2501 M Street, NW, Suite 200, Washington, DC 20037, USA
[c.gascon@conservation.org]

Scant attention has been paid to measurement error in frog morphometric studies. We study both interobserver effects of measurement on the same specimens of *Vanzolinius discodactylus* (Anura, Leptodactylidae) and intraobserver effect of repeated measurements on a single *V. discodactylus* specimen. Interobserver measurements differ statistically and result in different biological interpretations in some cases. Evidence is provided that log transformation of raw data is often unnecessary. Allometric transformation of measurement variables to remove size effect requires parallel regression slopes of variable against size. This requirement is not met with the *V. discodactylus* data, nor is it likely to be met when several variables are used in a morphometric study. We recommend: assume measurement differences between sexes in frogs and analyze data separately by sex; consider and select the most appropriate statistical model options for data analyses; avoid pseudoprecise measurements; do not rush to logarithmic transformation; remeasure at least one individual frog 20 times to provide an assessment of measurement error in data interpretation; be conservative in drawing biological inferences from morphometric analyses, basing interpretations and conclusions only on very robust effect size estimates and differences.

INTRODUCTION

Frogs are relatively soft-bodied organisms and their preservation requires considerable care. Limbs and body must be correctly positioned to achieve standardized preparation. Unfortunately, different preservatives and different individual techniques result in very different museum preparations for the same species (fig. 1). Therefore, precise, comparable measurements of preserved frogs are difficult. For example, one of the standard measurements taken on frogs, snout-vent length (SVL), is somewhat problematic in larger preserved frogs, because the sacral-urostyle portion of the body usually is fixed at an obtuse angle to the vertebral column. How much one "straightens out" the preserved animal has an effect on the

1. Corresponding author.

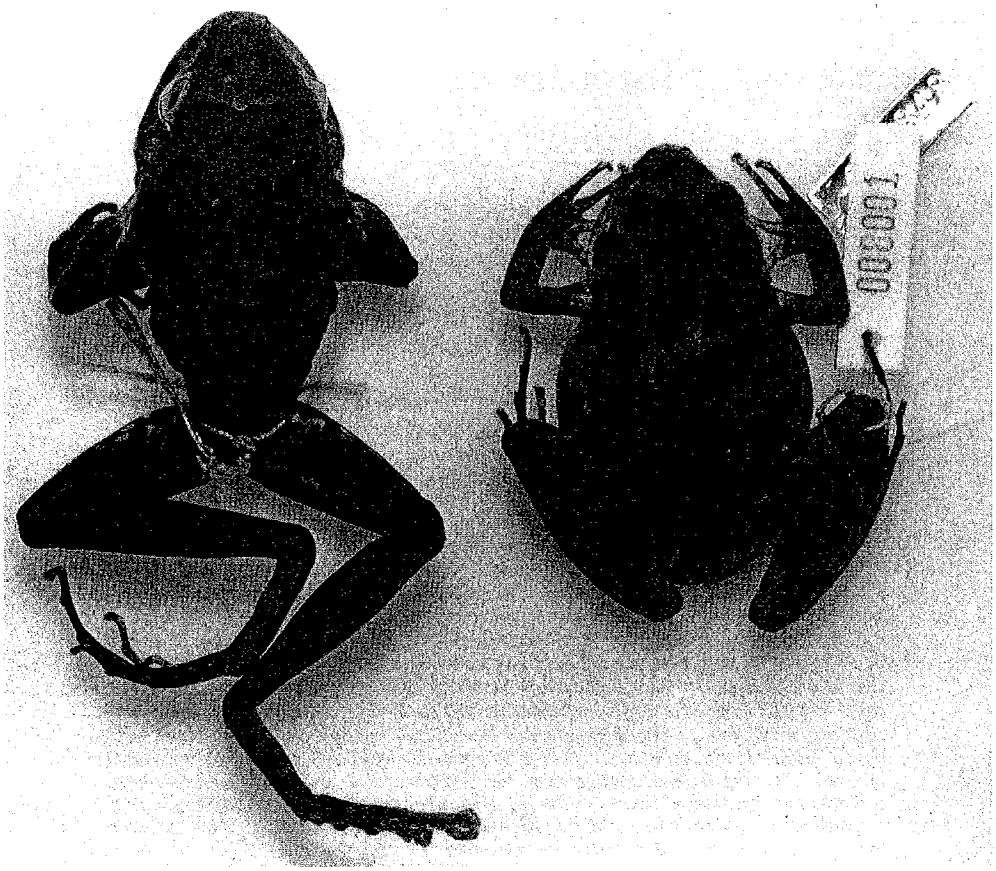

Fig. 1. – *Thoropa miliaris* (USNM 38936 on left, USNM 229848 on right) showing preservation/positioning differences that make accurate, comparable measurements difficult.

resultant measurement. In spite of (or, perhaps oblivious to) these difficulties, researchers have used frog measurement data to address a variety of scientific questions. There has been little attention paid to precision and repeatability of frog measurement data and how this variation might affect the scientific questions being addressed.

We know of only one study (LEE, 1982) that demonstrated important measurement differences between fresh and preserved frogs and differences in measurements taken on the same individuals at the same state of preservation. In that study, Lee took all the measurements himself using the same measuring equipment and methodology throughout. Although LEE (1982) presented extensive literature on the effects of preservation technique on fish morphology and discussed its relevance to frog morphometrics, herpetologists have generally ignored his warnings.

We are not aware of any published studies of the effect of different individual researchers taking the same set of measurements on the same frogs to measure inter-observer variability (although A. Dubois and A. Ohler have unpublished data on this topic, personal communication). Studies on other groups of organisms demonstrate that such differences are not trivial. LEE (1990) found differences in precision between two observers on scale count data taken from the same lizards. YEZERINAC et al. (1992) found that measurement error varied considerably, depending on the variable, for bird skeleton measurement data. In these studies, a constant value was being measured. That is, the number of scales did not change on any individual lizard, nor did the individual bird bones change size or shape. As indicated above, this is not true for whole frog specimens: how the specimen is positioned will determine what the value of the measurement will be for several of the measurements (variables) commonly taken for frog morphometric studies.

PAGANO & JOLY (1999) compared a select group of morphological measures on water frogs with an analysis of allozymic markers. These authors concluded that frog morphology was of limited use for their identification purposes. They determined frog body landmarks for measurement points from digitized photographs of specimens. Data were input and analyzed on a computer. Similar methodology has proved acceptable for characterization of stratigraphic sections (see e.g., BENSON et al., 1995), in which the surfaces are approximately linear and two-dimensional. However, for examination of three-dimensional, soft-bodied organisms, the use of such methods further complicates the measurement process. Despite the stated advantage of magnification of digitized figures for measurement purposes, statistical error minimization has not been proved to be achievable for measurements taken from frog photographs. Based on our experience, we do not recommend using photographs of frogs from which to take morphometric data.

One of us (CG) took a series of measurements on specimens of the frog species *Vanzolinius discodactylus* (Anura, Leptodactylidae) from the Rio Juruá in Brazil to test the riverine hypothesis of speciation (GASCON et. al., 1996). Another of us (WRH) used the same specimens in a study examining differentiation throughout the entire species range of *V. discodactylus* (HEYER, 1997). WRH took the same set of measurements on the same frogs that CG measured. The two data sets were given to LCH to analyze and evaluate. During the course of this study, LCH reevaluated the statistical procedures and assumptions used in the GASCON et al. (1996) study.

The objectives of this study are: (1) to evaluate inter- and intra-observer statistical differences of measurement sets; (2) to understand the kinds of differences investigators create when measuring frogs; (3) to evaluate the effect of measurement differences on certain statistical procedures that are generally applied in frog morphometric studies; and (4) to judge whether measurement differences yield different biological interpretations.

METHODS AND MATERIALS

Fourteen measurements were made on each frog, following the methodology in GASCON et al. (1996). The fourteen variables are: snout-vent length (SVL), nostril separation, eye

width anterior, eye width posterior, head width, head length, eye to nostril distance, tympanum diameter (tympanum height of GASCON et al., 1996), eye length, thigh length (femur length of GASCON et al., 1996), shank length (tibia length of GASCON et al., 1996), foot length, maximum width of disk on third finger, and maximum width of disk on fourth toe.

Prior to WRH's taking of these data, he confirmed landmarks with CG for a subset of the variables in an attempt to make certain that the measurements would be comparable.

CG and WRH measured each individual one time.

CG used digital calipers linked to an IBM-PC; measurements were made to the closest 0.01 mm and the data were recorded with three decimal places. WRH used Helios dial calipers; measurements were made to the closest 0.1 mm and the data were recorded with one decimal place.

To assess individual measurement error, WRH measured one male, USNM 348976, 20 times over a 12 day period. The eye region on one side of the head is slightly squashed, otherwise this specimen is in reasonable shape. The specimen is about average in overall state of preservation and positioning in terms of ease of measurements. Measurements were taken at various times of the day and measurements were never taken one immediately after the other to eliminate or minimize carry-over effects of learning or memory. For SVL, efforts were made to focus visually on the caliper jaws when measuring the specimen and not to look at the readout dial until after the jaws had been set. All other measurements were taken under a dissecting microscope with the calipers while the measurement readout dial was not visible in the field of observation. Measurements were recorded on dated and timed separate, individual data sheets.

CG and WRH used different criteria to categorize sex of the individuals. CG used three categories: F, M and 0. In cases where CG opened the frog to take tissues, sex and whether the individual was adult or not were determined by the state of its gonads. Individuals recorded as 0 were not opened. These data were recorded under field conditions. For the morphological analyses reported by GASCON et al. (1996), data for adult and non-adult males were combined as were the data for adult and non-adult females. WRH used five categories: M, F, B, G and J. The M (adult male) category was determined by presence of vocal slits in males. The F (adult female) category was determined by presence of developed ova or some curliness of the oviduct in females. The B (juvenile male) category was determined by presence of testes. The G (juvenile female) category was determined by presence of ovaries. The J (juvenile) category was used when sex could not be determined, either because the gonads were indeterminate in very small specimens or the gonads had been removed from the specimens when tissues had been taken. These data were taken in the laboratory with the aid of a Wild stereoscopic dissecting microscope.

Male and female immature gonads of *Vanzolinius discodactylus* are quite similar in appearance and difficult to differentiate without detailed examination under magnification. Both ovaries and testes have a mosaic-like pattern externally. The only consistent difference between immature gonads is that the testes have a smooth external surface, whereas ovaries have an irregular external surface. Not surprisingly, the difficulty of differentiating gonads using the unaided eye resulted in several different interpretations of sex by CG and WRH. The differences are (CG determination, followed by WRH determination): INPA 2410 (F, B); INPA 2371, 2433, 3397, 5605, 5671, 5728, 5735, 5736, 5799, 5801 (M, G); INPA 3572, 5571 (F,

J, gonads now removed in both); INPA 3177, 3573, 5524, 5592, 5670, 5697, 5730 (M, J, gonads now removed in all).

WRH's categories of adult male (M) and adult female (F) are used in the analysis section for both the CG and WRH measurement data sets unless otherwise noted. Using this categorization, 88 adult individuals are available for analysis. Each variable was examined and summarized separately for male and female adults. Graphs and descriptive statistics were calculated and assumptions tested prior to means tests or predictive analyses. Logarithmic transformations were performed and descriptive statistics calculated on the transformed values as well. Tests of normality were performed and discussed below.

In this study, we cannot calculate residual measurement error because we do not have the "true" value of the variable for any individual specimen. Similarly, we are unable to assess a statistical variability estimate for the factors involved in the overall measuring error. That is, we cannot remove intra-observer variability from inter-observer measurement error. We therefore evaluate the two factors separately.

We distinguish "precision" from "accuracy". Accuracy is the closeness of an observer's measurement to the quantity intended to be measured. In our case, this is unknown for the true value of the frog's morphological measurement but can be evaluated by considering the closeness of the results of the two observer's values. Precision refers to the entire class of measurements and how well repeated measurements self-conform. In this case, the mean value does not have to be the "true" value of the variable. To examine these characteristics we calculated both inter- and intra-observer variability estimates and also descriptive measures for qualitative evaluation of the frog data.

Data were analysed either using direct mathematical formulae or using the software package SPSS 8.0 (ANONYMOUS, 1998). Although the discriminant function analyses were done using SPSS 8.0 (ANONYMOUS, 1998), the figures were produced using either SYSTAT versions 7 (ANONYMOUS, 1997, for fig. 7) or 9 (ANONYMOUS, 1999, for fig. 5-6).

THE APPROPRIATENESS OF RAW DATA TRANSFORMATION PROCEDURES IN FROG MORPHOMETRIC STUDIES

GASCON et al. (1996) used an allometric transformation procedure described by THORPE (1976) in an effort to remove size effects from the data. The Thorpe procedure (presented in detail in THORPE, 1975) involves two steps: (1) log-transforming the original measurement data; and (2) transforming the log values using a common slope based on the entire data set. The topic of transforming raw data is discussed first, followed by demonstration that the statistical assumptions of the Thorpe procedure are not met by the *Vanzolinius* data as used by GASCON et al. (1996).

Although not specifically mentioned by GASCON et al. (1996), the raw measurement data were log-transformed as part of THORPE's (1976) transformation procedure. Raw data are transformed as a matter of course in many multivariate analyses of frog morphometric data (for a recent example see GREEN et al., 1997). SOKAL & ROHLF (1969) state that log transformation is the most common transformation for biological data and they provide a cogent

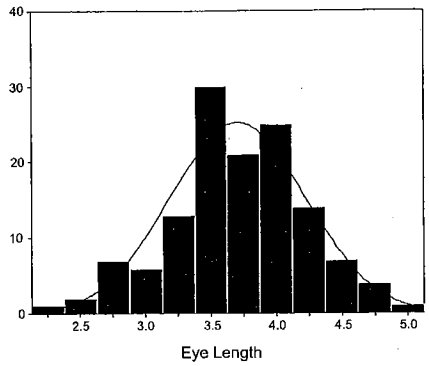

Fig. 2. – Histogram of eye length values measured by CG on total sample of 131 frogs with normal distribution best fit.

discussion on the topic of log-transforming variables as a way to meet some statistical test assumptions that are not met by raw variable data. However, this transformation is often applied routinely, when, in fact, it may be either unnecessary or incorrect to do so.

Replacing each measurement by its logarithm may result in more approximate variance equality. Also, for many biological applications the data can be normalized by this change. The assumption of concern for our purposes is whether the variables are normally distributed. Using BESTFIT (ANONYMOUS, 1995) on the data as analyzed by GASCON et al. (1996), untransformed variables for the entire sample size of 131 individuals were fit with a normal distribution (see fig. 2 for an example). We used the Anderson-Darling test criterion as well as a chi-square test of fit. The Anderson-Darling criterion is more tail-sensitive than the ordinary chi-square goodness-of-fit test.

SOKAL & ROHLF (1969) state that the log transformation may be appropriate and useful when the means of the samples are proportional to the range or standard deviation of the respective samples. The biological questions we are asking of the *Vanzolinius* data require grouping of the data by locality. None of the variables, for the total sample or when organized by locality, show a relationship of mean with either standard deviation ($r = 0.06$ *ns*) or range ($r = 0.19$ *ns*). In addition, each raw variable plot shows approximate symmetry, lack of prominent skewness and unimodality (for example, snout-vent length as shown in fig. 3).

Thus, the data as analyzed by GASCON et al. (1996) can be appropriately analyzed as raw variable measurements, rather than log-transformed variables. It is not incorrect statistically to apply and use the logarithmic sample data for this problem. It is, however, unnecessary for the morphological variables being measured here.

The reason GASCON et al. (1996) used logarithmic transformation was to attack the problem of allometry effects in their data, which included both adults and juveniles. THORPE (1976) presented a procedure that uses a log transformation as an initial step toward eliminating the influence of allometry. We examined the application of this approach and found it inappropriate for the *Vanzolinius* data for the following reason.

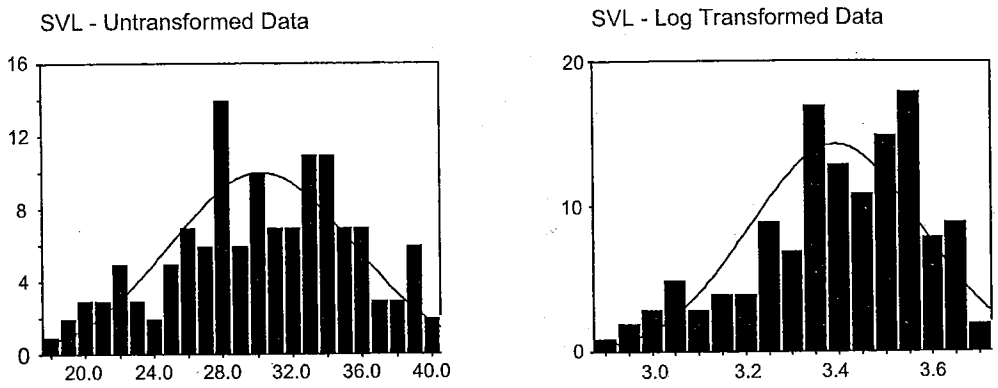

Fig. 3. – Histogram of SVL values measured by WRH with a normal distribution curve superimposed on both raw and log transformed data for 131 specimens.

Following THORPE (1975), GASCON et. al. (1996) used the allometric transformation for all variables to "remove size effects for the data". Raw measurements were adjusted using a common slope for all locality data sets and sexes combined. This procedure adjusts the allometric character or variable by using the slope of its regression against size. When there are multiple localities, as in the case of the present work, the pooled within-locality slope is used to make the adjustment.

This procedure can be applied appropriately only when the locality slopes are approximately parallel. That is, when a test of slope homogeneity (the first step in most packaged ANCOVA programs) shows no significance, the slopes from the separate localities can be pooled. For the 11 localities of this study, that is not the case. When there is heterogeneity, one can do the calculations to obtain a common within-locality slope, but the resultant number is meaningless. When the slope test indicates heterogeneity, as is the case for this data set (P = 0.001), there can be no one slope to describe the data (fig. 4). Therefore, the problem of size effects in the GASCON et al. (1996) data would remain.

We can eliminate the need to consider allometry by using only adults but we still need to consider sexual size effects. If size effects are not present or if they can be removed statistically, then male and female specimens can be pooled for analyses that can be more statistically powerful. As stated by GASCON et al. (1996), the transformation manipulations they (inappropriately) applied did remove size effects between males and females (which included both immature and mature individuals) for all variables except head length. They deleted this variable from their analyses and combined male and female data in their analyses. In our analyses, we examine the sex differences on both raw and transformed variables using adults only.

When the raw variables are examined using CG's classification of males and females (124 total) and his measurements: (1) all fourteen variables have non-significant univariate homogeneity of variance tests (an assumption for means tests); (2) all univariate F tests ($F_{1, 122}$) on means are significant (P = 0.000); (3) the multivariate $F_{14, 109}$ is significant (P = 0.000;

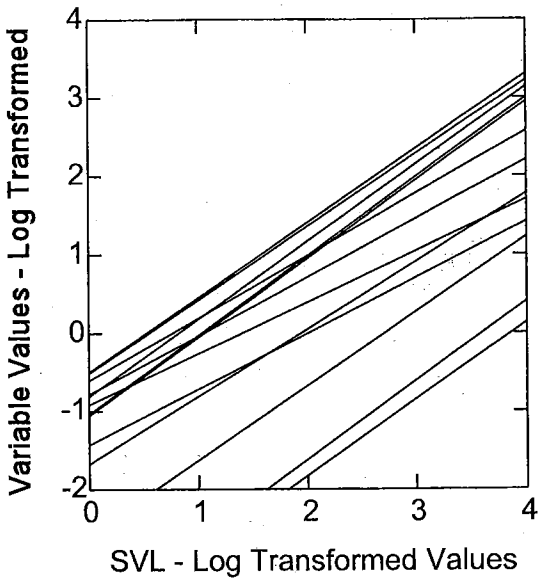

Fig. 4. – Fourteen regression slopes for morphometric variables, log transformed, CG measurements, for 131 specimens per variable.

Hotellings $T^2 = 1.075$); and, (4) homogeneity of slope is rejected ($P = 0.000$). No regression effect could be determined and removed. Similar results hold for the logarithmic-transformed data. Thus, it is inappropriate to assume that we can combine Gascon's raw male and female data in univariate or multivariate analyses. We know of no valid procedure to remove the sexual size differences under the conditions involved with this data set.

When WRH's raw data of 88 (57 female; 31 male) known adults are used: (1) all 14 variables have non-significant univariate homogeneity of variance test results; (2) all univariate $F_{1, 86}$ -tests (*t*-tests, *df* 86) are significant ($P = 0.000$); and, (3) multivariate $F_{14, 84}$ is significant ($P = 0.000$). Similar results held for the log-transformed variables. In practice then, because of equivalent results with this sample data, either the log-transformed or the raw data could be used for further testing. However, it is an unnecessary complication for both application and interpretation to transform a variable when the raw data can be used. We continue with the raw data results for the 88 adult specimens, for which males and females test significantly different on each of the measurements considered.

ANALYSIS OF MEASUREMENT DATA

EFFECT OF ROUNDING

There are two components to consider when rounding a raw measurement value that could impact amphibian data sets: (1) pseudo-precision, and (2) the number of decimal places used by computers in calculating statistical algorithms.

Pseudo-precision is using greater precision in calculations for measurements than can be justified in terms of the originally recorded accuracy of those measurements. For example, if multiple measures of the tympanum diameter of the same individual frog specimen are 2.165, 2.224, 2.187, 2.240, 2.193, the tympanum cannot be measured accurately beyond one decimal place. Using values with two or three decimal places for these values is pseudo-precision. Statisticians advise using precise measurements only (e.g., SOKAL & ROHLF, 1969: 13-16). Biological practitioners routinely ignore this advice. For example, although WRH uses mechanical dial calipers that record measurements to the nearest tenth of a millimeter, in the size range of *Vanzolinius discodactylus*, snout-vent length can be measured only to a precision of 0.6 (see tab. 4). Thus, this variable should be recorded to the whole number, not with one decimal place.

A second potential biological consequence results from the number of decimal places computers use in calculations. This is less of a problem now with recent computer advances in calculation. However, using pseudo-precise measurement data certainly can result in different numerical values for test statistics, which are summarizations. To test whether any biologically meaningful interpretations would be drawn from our data due solely to rounding errors, paired *t*-tests were computed on two sets of the data. We compared the CG and WRH measurements as recorded (WRH with one decimal place, CG with three) with both data sets recorded to one decimal place. The CG data set was rounded by the usual method of rounding up the i^{th} place when the $(1+i)^{th}$ place is 5 or more.

As expected, when different numbers of decimal places are used (rounded vs. not) for the data set, several of the resultant test values vary slightly. However, in no case are the decisions different for the selected test level (0.05, 0.01, 0.001), nor would any different biological inferences likely be drawn from the observed probability levels (tab. 1) from corresponding tests.

While pseudo-precision, as a consequence of computer generated or digital caliper induced values, is biologically and statistically offensive, it does not impact seriously the univariate descriptive or inferential results of real data sets such as ours for *Vanzolinius discodactylus*.

INTER-OBSERVER DIFFERENCES

A battery of descriptive statistics was run on the raw measures of WRH-defined adults to evaluate the nature of differences between the CG and WRH measurements (tab. 2). The mean for each observer was calculated for each measurement. The usual assumption for a

Table 1. – Comparison using CG measurements at three decimal places and rounded off to one decimal place to WRH measurements at one decimal place. Mean values reported at statistically inappropriate 4 decimal place level to demonstrate effect of computation results.

Variable	Data set	Means		Coefficient of variation		T - statistic		T - significance	
		0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
Snout-vent length	CG	30.1270	30.1282	5.85	5.85	- 2.85	- 2.82	0.005	0.005
	WRH		30.2260		5.77				
Nostril separation	CG	2.7058	2.7053	6.87	6.78	5.98	5.92	0.000	0.000
	WRH		2.5908		6.76				
Eye width anterior	CG	5.8029	5.8000	7.06	7.12	- 7.67	- 7.52	0.000	0.000
	WRH		6.0130		7.02				
Eye width posterior	CG	8.1690	8.1710	6.82	6.84	- 9.79	- 9.79	0.000	0.000
	WRH		8.4771		7.14				
Head width	CG	10.5286	10.5321	5.76	5.76	- 0.57	- 0.48	*ns*	*ns*
	WRH		10.5496		6.17				
Head length	CG	11.0501	11.0430	5.58	5.59	- 14.57	- 14.66	0.000	0.000
	WRH		11.9618		6.80				
Eye-nostril distance	CG	3.5326	3.5290	6.30	6.29	- 3.21	- 3.36	0.002	0.001
	WRH		3.5901		6.34				
Eye length	CG	3.7019	3.7046	7.11	7.15	8.70	8.80	0.000	0.000
	WRH		3.4917		8.02				
Tympanum diameter	CG	1.9437	1.9450	5.40	5.39	- 16.67	- 16.09	0.000	0.000
	WRH		2.1962		5.53				
Thigh length	CG	12.8507	12.8489	5.50	5.50	- 3.99	- 4.02	0.000	0.000
	WRH		13.0870		5.60				
Shank length	CG	14.3533	14.3542	6.16	6.16	10.94	10.81	0.000	0.000
	WRH		14.1160		6.00				
Foot length	CG	15.4560	15.4550	5.97	5.98	- 12.87	- 12.86	0.000	0.000
	WRH		16.0511		5.86				
Third finger disk width	CG	0.6413	0.6427	5.13	5.06	- 4.83	- 4.39	0.000	0.000
	WRH		0.6710		6.10				
Fourth toe disk width	CG	0.8235	0.8260	5.18	5.04	- 10.44	- 9.48	0.000	0.000
	WRH		0.8962		5.46				

Table 2. – Descriptive statistical differences between CG and WRH measurement data on the same specimens of adult *Vanzolinius discodactylus* ($n = 88$).

Variable	Data set	Mean	Standard deviation	Variance	Standard - error	Coeff. variation	Hartley test	*t*	Observed *P* (2 tail)	*r*	*P* (2 tail)	Coeff. determin.
Snout-vent length	CG	32.54	3.90	15.19	0.42	8.35	1.03	- 3.15	0.002	1.00	0.000	0.99
	WRH	32.67	3.95	15.63	0.42	8.26						
Nostril separation	CG	2.87	0.32	0.10	0.03	8.88	0.81	5.24	0.000	0.75	0.000	0.57
	WRH	2.75	0.29	0.08	0.03	9.47						
Eye width anterior	CG	6.16	0.64	0.41	0.07	9.60	0.99	- 6.47	0.000	0.87	0.000	0.76
	WRH	6.38	0.63	0.40	0.07	10.12						
Eye width posterior	CG	8.71	0.92	0.85	0.10	9.45	0.97	- 7.30	0.000	0.92	0.000	0.84
	WRH	9.01	0.91	0.83	0.10	9.91						
Head width	CG	11.38	1.36	1.85	0.14	8.36	0.88	0.66	0.508	0.95	0.000	0.90
	WRH	11.35	1.28	1.64	0.14	8.88						
Head length	CG	11.98	1.33	1.77	0.14	9.01	0.94	- 12.42	0.000	0.89	0.000	0.80
	WRH	12.79	1.29	1.66	0.14	9.91						
Eye – nostril distance	CG	3.76	0.42	0.18	0.04	8.87	1.02	- 3.66	0.000	0.88	0.000	0.77
	WRH	3.84	0.43	0.18	0.05	8.98						
Eye length	CG	3.91	0.43	0.18	0.05	9.03	0.50	7.120	0.000	0.73	0.000	0.54
	WRH	3.69	0.31	0.09	0.03	12.01						
Tympanum diameter	CG	2.08	0.30	0.09	0.03	6.96	1.20	- 14.12	0.000	0.89	0.000	0.79
	WRH	2.35	0.33	0.11	0.04	7.18						
Thigh length	CG	13.87	1.87	3.51	0.20	7.40	1.01	- 3.14	0.002	0.94	0.000	0.89
	WRH	14.08	1.88	3.54	0.20	7.49						
Shank length	CG	15.38	1.75	3.07	0.19	8.77	1.05	8.74	0.000	0.99	0.000	0.98
	WRH	15.14	1.80	3.23	0.19	8.43						
Foot length	CG	16.61	1.96	3.82	0.21	8.49	1.05	- 11.86	0.000	0.96	0.000	0.93
	WRH	17.28	2.01	4.03	0.21	8.61						
Third finger disk width	CG	0.70	0.10	0.01	0.01	6.96	0.79	- 2.83	0.006	0.73	0.000	0.53
	WRH	0.72	0.09	0.01	0.01	8.06						
Fourth toe disk width	CG	0.89	0.12	0.02	0.01	7.15	1.08	- 8.11	0.000	0.77	0.000	0.59
	WRH	0.97	0.13	0.02	0.01	7.51						

Table 3. – Performance rankings of measurement variables.

Variable	Mean difference/ mean	Coefficients of variation	Differences in coefficients of variation	Hartley test
Snout-vent length	Good	Best	Good	Good
Nostril separation	Moderate	Moderate	Good	Moderate
Eye width anterior	Moderate	Worst	Good	Good
Eye width posterior	Moderate	Moderate	Moderate	Good
Head width	Good	Best	Moderate	Moderate
Head length	Poor	Moderate	Poor	Moderate
Eye-nostril distance	Moderate	Moderate	Good	Good
Eye length	Poor	Worst	Poor	Poor
Tympanum diameter	Poor	Best	Good	Moderate
Thigh length	Moderate	Best	Good	Good
Shank length	Moderate	Moderate	Good	Good
Foot length	Moderate	Best	Good	Good
Third finger disk width	Moderate	Best	Poor	Moderate
Fourth toe disk width	Poor	Best	Moderate	Good

Student's *t*-test are not met because the observers measured the same sample and not samples independently chosen at random. Because these are repeated measurements, the test statistic denominator we use to test for a difference between each observer pair of measures is the formula for the standard error of a difference when samples are not independent. That formula is: $s_{m(1)\text{-}m(2)} = s_d = \text{sqrt}(s^2_{m(1)} + s^2_{m(2)} - 2s_{m(1)}\, s_{m(2)}\, r_{12})$, where m(i), i = 1,2, are the two observers means for the particular measurement, s is the standard deviation for each, and r is the correlation between the two sets of paired measures. Alternatively, for n pairs, the standard deviation of the d differences can be written, $s_d = \text{sqrt}((n\,\Sigma d^2 - (\Sigma d)^2/(n\text{-}1)))$, and the test statistic is $t = \Sigma d/s_d$.

The paired *t*-test results indicate that all variables differ significantly except for one (head width). The correlation coefficients are all statistically significant and most coefficients of determination are high. The correlation statistics, considered with corresponding coefficient of variation values, indicate that the two sets of observer measurements are consistent and generally comparable. The *t*-test results leave no doubt, however, that overall, our two sets of measurements differ statistically.

Given that our measurements are statistically different, we wish to explore our measurement performance on a variable by variable basis. To do this, various ways of describing performance are ranked and compared.

(1) Mean inter-observer difference of measures adjusted by magnitude of variable. The intent of this comparison is to evaluate how well the two sets of measurements agree with each other, specifically to see if the observers performed better on larger measurements than smaller (e.g., snout-vent length (SVL) vs. width of third finger disk). The smaller mean value

for the same individuals and for each variable was subtracted from the larger. That number was divided by the average value of the two means. The resultant values range from 0.002 to 0.126. For comparative ranking purposes, good is considered to be 0.000-0.005, moderate 0.005-0.050, and poor 0.050-0.150 (tab. 3).

(2) Coefficients of variation. Values of the coefficient of variation (CV) for each measure are often used to compare the variability of the variables. Adjustments for sample size and other factors have been suggested (e.g. DELAUGERRE & DUBOIS, 1985). We chose to use the original formula and to categorize the CV values because, regardless of adjustment, the CV remains extremely sensitive to errors in sample means. For evaluation and ranking purposes, the best category, 5.0-6.0, has the lowest variability in the attribute measured; moderate is 6.0-7.0, and the worst category is 7.0-8.0 (tab. 3). Most of the coefficient of variation values for each observer pair fall into the same categories (see tab. 2); in the few cases where our values fell in different categories, the average of our values was used for category placement.

(3) Difference in coefficients of variation. The intent of this comparison is to evaluate repeatability of our measurements. If each of us has the same degree of measurement repeatability, our coefficient of variation values should be identical. Therefore, how different these values are indicates degree of deviation from consistency of measurement for the variable involved. For ranking purposes, good is a difference of 0.0-0.2; moderate is 0.3-0.5; and poor is 0.6-1.5 (tab. 3).

(4) Hartley F-max test. The Hartley test statistic, which is the quotient of the larger and the smaller variance, provides another way to evaluate repeatability of measurements. A Hartley test value of 1 is not significant; values both larger and smaller indicate differences. For ranking purposes, good is 0.9-1.1, moderate 0.8-0.9 or 1.1-1.2, and poor < 0.8. (tab. 3).

From the above (tab. 3), it is apparent that CG and WRH measured one variable consistently and with the greatest precision: snout-vent length. There are five variables that we measured with reasonable consistency and precision: head width, eye-nostril distance, thigh length, shank length, and foot length. There are four variables that we apparently measured differently, but each of us with reasonable to good precision: head length, tympanum diameter, foot length, and width of fourth toe disk. Apparently we are using slightly different landmarks for these measurements. For the tympanum, it would seem that CG's description of tympanum height (GASCON et al., 1996) does in fact describe something different from WRH's definition of tympanum diameter. Once these results became known, CG confirmed that he always measured the vertical distance of the tympanum relative to head position and WRH took the measurement at the point of greatest tympanum diameter, irrespective of position of the tympanum relative to the head. For the width of the fourth toe, we obviously used different criteria of how much contact of the disk with the calipers was used. The most inconsistent measurement is eye length. That is, we measure the variable differently as well as imprecisely. This suggests that this variable should not be used for morphometric analyses in *Vanzolinius discodactylus*. We further suggest that because this variable is affected by preservation artifact to a great degree, it should probably not be included in any frog morphometric study.

There is one result we find surprising. Overall, we measured larger variables (such as snout-vent length) equally as well (or poorly, depending on perspective) as smaller variables (such as third finger disk width).

Table 4. – Descriptive statistics for 20 repeated measurements of a single specimen of *Vanzolinius discodactylus*.

Variable	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
Snout-vent length	26.1	26.7	26.4	0.16	0.01
Nostril separation	2.0	2.4	2.2	0.09	0.04
Eye width anterior	5.1	5.6	5.4	0.12	0.02
Eye width posterior	7.4	8.0	7.7	0.17	0.02
Head width	9.0	9.3	9.2	0.10	0.01
Head length	10.0	10.6	10.3	0.16	0.02
Eye-nostril distance	2.9	3.2	3.0	0.09	0.03
Eye length	3.0	3.8	3.4	0.19	0.06
Tympanum diameter	1.6	1.8	1.7	0.06	0.03
Thigh length	11.0	11.8	11.4	0.24	0.02
Shank length	11.8	12.1	12.0	0.07	0.00
Foot length	12.8	13.7	13.2	0.24	0.02
Third finger disk width	0.5	0.6	0.6	0.02	0.04
Fourth toe disk width	0.7	0.8	0.8	0.04	0.05

Intra-observer differences

Standard descriptive statistics for the twenty repeated measurements on each morphological variable of the single specimen (tab. 4) generally mirror inter-observer variation. That is, SVL, which CG and WRH measured with greatest precision, has a low intra-observer coefficient of variation. Eye length, which was the most imprecise inter-observer variable, has the highest intra-observer coefficient of variation.

Given the thousands of frogs that WRH has measured, one would predict that there would not be a change (improvement) in measurement accuracy from the first re-measure to the twentieth. Two sample *t*-tests of measurements 1-10 against measurements 11-20 were not statistically significant, except for posterior eye width. Given that the eye that was measured was misshapen with preservation, it is likely that the landmarks used by WRH changed over the re-measurement process.

Biological interpretations of measurement differences

Inter-observer differences

Our inter-observer differences over sets of measurements are unarguably statistically different at highly significant levels, yet it does not necessarily follow that such inter-observer

measurement differences lead to different biological conclusions for the same set of specimens. For example, it seems likely that some of our measurement differences are due to consistent differences in the way we took the measurements. Given a large enough sample, such differences would be statistically significantly different. However, because the measurements would have been taken consistently by each observer, the variation described in the two sets of measurements would be equivalent, and, hence, lead to similar conclusions for any biological inferences drawn from the data. We test this idea using our measurement data in two analyses aimed at obtaining insight into biological processes through analyses of morphometric data.

Geographic variation

Multivariate discriminant function analyses are often used to analyze patterns of geographic variation in study organisms. For our purposes, we grouped the specimens from GASCON et al.'s (1996: 377, fig. 1) eleven numbered localities into four major groups, separated linearly along the Rio Juruá. Our Area 1 is GASCON et al.'s (1996) locality 1 ($n = 7$), Area 2 is localities 2+3+8+9 ($n = 20$), Area 3 localities 4+5+10+11 ($n = 50$) and Area 4 localities 6+7 (n = 11). We use only WRH-defined adult specimen raw data ($n = 88$) in the analyses.

As described previously, the data for males and females are significantly different ($P < 0.001$). The values for each of the variables are assumed to have a multivariate normal distribution with equal variance-covariance matrices (VCV) within the 4 areas. To decide whether to combine the sexes, locality tests should be performed. However, all tests of VCV equality are highly sensitive to normality. In addition, there is no practical, effective test for multivariate normality for our smaller-sized samples. We can hypothesize that since the sexes are highly significantly different over the entire sample then they should be different in and over each area. Alternatively, we might not.

Let us use the untransformed measurements to examine the results of a discriminant analysis by sex. We use WRH's designations of adult males and females and compare final results when using each observer's measurements.

For the female data (Area 1: $n = 5$; Area 2: $n = 12$; Area 3: $n = 34$; Area 4: $n = 6$; total: $N = 57$), the discriminant analysis results for each of the observer's data sets are far from identical (tab. 5, fig. 5). Of particular interest is that, in the stepwise procedure, the variable entered in the first step (that which explains the greatest amount of unconditional univariate variance among area samples) differed, as did the variables used in the final model. Since the variable impact differed between the two data sets in the discriminant model, it is not surprising that there were differences in the values for the canonical functions, first axis variable loading, and posterior classifications (tab. 5).

Male data (Area 1: $n = 2$; Area 2: $n = 8$; Area 3: $n = 16$; Area 4: $n = 5$; total: $N = 31$) results are similar (tab. 6, fig. 6) to the female results in the kinds of discrepancies that measurement differences caused in the discriminant function analyses for the two sets of measurement data.

Would the different results from these analyses result in different biological interpretations? One of the main methods for evaluating such geographic variation analyses is the plot of the first two canonical axes. The discriminant function program in SYSTAT 9 (ANONY-

Table 5. – Comparison of discriminant function analysis results for female data of *Vanzolinius discodactylus* by geographic regions, with two sets of measurements taken on the same individuals.

CG measurements	WRH measurements
Significant univariate F-test SVL Head length Head width Nostril separation Eye-nostril distance Eye width anterior Eye width posterior Tympanum diameter Thigh length Shank length Foot length	Significant univariate F-test SVL Head length Head width Nostril separation Eye-nostril distance Eye width anterior Eye width posterior Tympanum diameter Thigh length Shank length Foot length
Stepwise discriminant model First variable tried Thigh length Final model uses Posterior eye width Head width Thigh length Foot length Eye length Third finger disk width All groups separable at 0.001 level in final model	Stepwise discriminant model First variable tried Tympanum diameter Final model uses Posterior eye width Shank length Tympanum diameter Final model cannot separate Group 4 from Group 1
Canonical discriminant function F. # Eigenvalue % variation $\sim\chi^2$ 1 1.2801 0.55 0.000 2 0.8429 0.36 0.000 3 0.2028 0.09 0.052 *ns*	Canonical discriminant function F. # Eigenvalue % variation $\sim\chi^2$ 1 1.1674 0.77 0.000 2 0.2750 0.18 0.002 3 0.0718 0.04 0.056 *ns*
First axis explanation Thigh length (0.94) Head width (- 0.66) Foot length (0.63) Third finger disk width (- 0.52)	First axis explanation Tympanum diameter (0.93) Posterior eye width (- .058) Shank length (0.51)
Overall classification Group % 1 100 2 88 3 69 4 80 Overall 77.4	Overall classification Group % 1 80 2 83 3 83 4 67 Overall 80.7