Notes and Comments

An Improved Procedure for Testing the Effects of Key Innovations on Rate of Speciation

Jérôme Goudet*

Institut d'Ecologie-Zoologie et Ecologie Animale, Bâtiment de Biologie, Université de Lausanne, CH-1015-Lausanne, Switzerland

Submitted February 23, 1998; Accepted December 16, 1998

Keywords: Fisher procedure, randomization tests, phylogeny, statistical power, tests of symmetry.

The effect of key innovations (e.g., phytophagy in insects, Mitter et al. 1988; viviparity in fishes, Slowinski and Guyer 1993) on speciation rates has been a major focus of evolutionary biology in recent years (Sanderson and Donoghue 1996). Much of the apparent variability in species number might be consistent with simple stochastic models of phylogenesis, since all degrees of species diversity are equally likely under a null model of random speciation and extinction (Farris 1976; Slowinski and Guyer 1993; but see Losos and Adler 1995). Therefore, invoking a key innovation to explain the diversity of one single group possessing the innovation against a sister group not possessing it is not tenable. However, if several to many sister groups are considered, tests of positive association between the possession of the trait and species number can be designed (Slowinski and Guyer 1993).

The first to follow this path were Mitter et al. (1988), who found a significant positive association between phytophagy and number of insect species. To test the association, they used a sign test. Slowinski and Guyer (1993) suggested an improved, less conservative method, based on a null model of random extinction and speciation. Under this model, the probability (P value or p_i) of observing disparity as large or larger in species number between each group possessing the innovation and its sister group is first calculated. The different P values obtained are then combined using Fisher's widely advocated procedure to combine probabilities (Fisher 1970; Manly 1985; Sokal and Rohlf 1995). Under the model of random speciation and extinction, the distribution of Slowinski and Guyer P values follows a uniform distribution. Applying Fisher's procedure to these P values will therefore test the null hypothesis that the distribution of sister-group sizes follows a model of random speciation and extinction. The alternative hypothesis Slowinski and Guyer wish to test is that there are more species in groups possessing the innovation. Here I argue that the null hypothesis to be tested needs to include all cases where the distribution of P values is symmetrical about .5, of which the uniform distribution is only a special case. I suggest three randomization tests for this modified null hypothesis and estimate their respective power. One of the tests emerges as the most powerful and is applied to a body of key innovation data obtained from the literature.

Presentation of the Tests

I first summarize the principle of Fisher's procedure (FP): to combine results from independent tests of the same hypothesis, Fisher (1970) suggested use of the property that a sum of a number of values of χ^2 is itself a χ^2 with appropriate degrees of freedom. The P values for each of the n tests can be transformed into a χ^2 with 2 df using the relation $\chi^2_2 = -2\ln(p_i)$. If the null hypothesis is true for all n tests, the n p_i are independent samples from a uniform distribution between 0 and 1, and the sum $X^2 = -2\sum \ln(p_i)$ of these n values is approximately distributed as a χ^2 with $2n$ df. The null hypothesis is rejected at the α level when the probability associated with the calculated X^2 is $<100\alpha\%$. This procedure implies the p_i to be drawn from a uniform distribution when the null hypothesis is true.

When testing for the effect of a supposed key innovation, one wants to know whether there are more species in groups possessing the key innovation. The null hypothesis can be formulated as H_{0g}: "the presence of a key innovation has no effect on group size." This means that

* E-mail: jerome.goudet@ie-izea.unil.ch.

Am. Nat. 1999. Vol. 153, pp. 549–555.

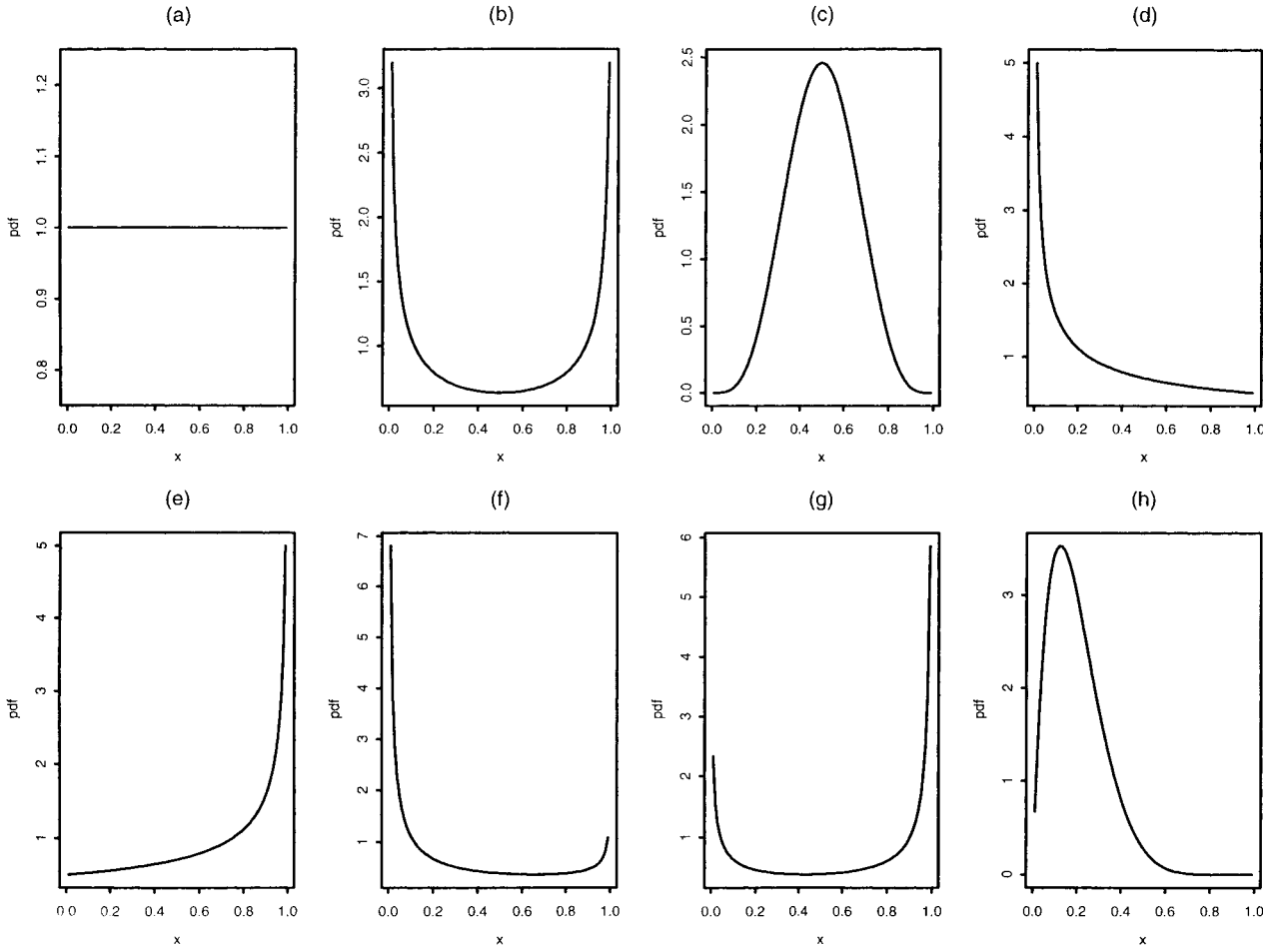

Figure 1: Alternative probability density functions (*pdf*) of the β distributions. *a*, $p = q = 1$ (uniform distribution). *b*, $p = q = .5$ (arcsin(x), U-shaped). *c*, $p = q = 5$ (bell-shaped). *d*, $p = .5$, $q = 1$ (L-shaped). *e*, $p = 1$, $q = .5$ (J-shaped). *f*, $p = .2$, $q = .6$ (U-shaped). *g*, $p = .4$, $q = .2$ (U-shaped). *h*, $p = 2$, $q = 20$.

the joint distribution of (X, Y), where X represents the number of species in groups with the key innovation and Y represents the number of species in groups without it, is the same as the joint distribution of (Y, X). Therefore, it implies that the null distribution of the P values as defined by Slowinski and Guyer is symmetrical about .5. Note that the null hypothesis tested at the level of each sister group is different (as pointed out by Slowinski and Guyer [1993], tests on individual groups should not be applied, but to obtain the P value of the group, it is necessary to specify under which null hypothesis this P value is obtained) and can be formulated as H_{0f}: "The number of species in the two sister groups is compatible with the random model of speciation and extinction." Here I show that, when the actual distribution of P values is U-shaped, using Fisher's procedure to test H_{0g} gives an unduly large Type I error, while, when it is bell-shaped, the procedure is highly conservative. I suggest overcoming this problem by using a randomization test. Its algorithm allows us to test whether the observed p_i are drawn from a symmetrical distribution with a mean of .5, be it uniform, U-shaped, or bell-shaped. Letting G be a statistic used to rank the vectors of p_i: step 1, calculate G_{Obs} on the observed set of p_i; step 2, subtract .5 from the observed p_i to give cp_i; step 3, assign a sign at random to the cp_i using a Bernoulli distribution with mean .5 (shuffling signs among the cp_i is not appropriate here since if all cp_i are <0, only one vector state exists); step 4, add .5 to the randomly signed cp_i to obtain a distribution of random P values conditional on the observed array of P values; step 5, let G_j be the statistic used to rank the jth vector of randomized P values, and test whether G_j is less than or equal to G_{Obs}, the statistic calculated on the observed data set; step 6, repeat steps

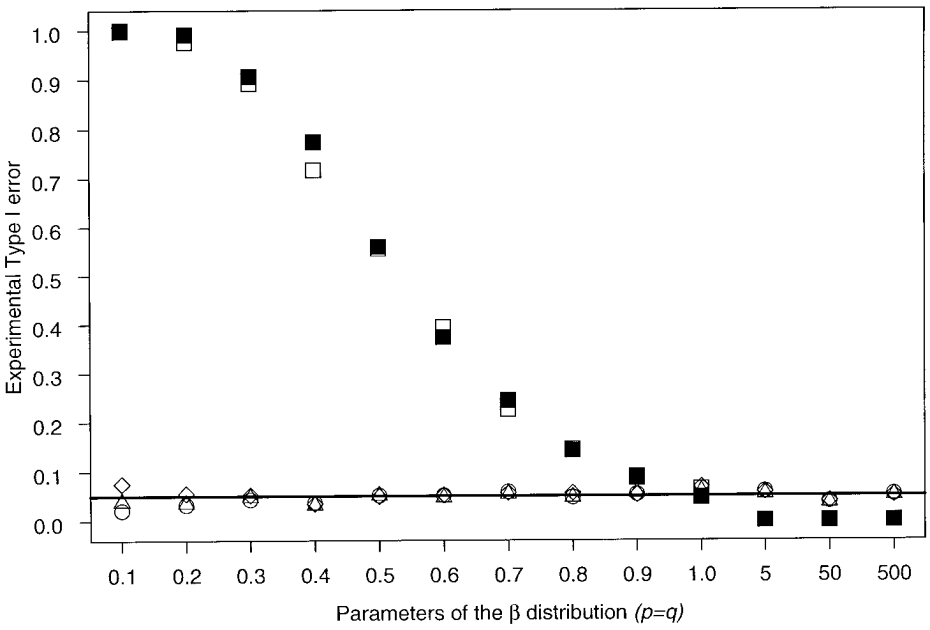

Figure 2: Experimental Type I error of Fisher's procedure (FP), and of the three tests proposed here—arithmetic (PA), geometric (PG), and harmonic (PH)—when the underlying distributions of P values are symmetrical about .5 ($p = q$). *Empty squares*, FP; *filled squares*, FP on the complement to one of the P values; *triangles*, PA; *circles*, PG; *lozenges*, PH. The horizontal line represents the nominal level fixed at 5%.

2–5 a large number (n) of times (e.g., 5,000). An unbiased estimate of the sought probability is

$$P = \frac{1 + \sum_{G_i \geq G_{\text{Obs}}} 1}{n + 1}$$

(Dwass 1957; Hope 1968).

Three statistics arise naturally to rank the vectors of P values: the arithmetic mean of the P values, $\sum_{np} p_i/np$, where np is the number of sister groups; their geometric mean,

$$\sqrt[np]{\prod_{np} p_i};$$

and their harmonic mean, $np/\sum_{np}(1/p_i)$. The test based on the geometric mean is a "distribution-free" equivalent to FP since a sum of logarithms is equivalent to a logarithm of products. Tests based on the arithmetic, geometric, and harmonic means will be called, respectively, PA, PG, and PH in the following.

The power of these different tests can be analyzed by generating data from known alternative hypotheses H_1. Since we want to generate P values from distributions that can be U-, L-, J-, or bell-shaped or flat (the uniform distribution), the appropriate distribution to use for this purpose is the β distribution (fig. 1), whose probability density function is given by

$$p(x) = \frac{1}{\beta(p,\ q)} x^{(p-1)}(1 - x)^{(q-1)}, \quad (0 \leq x \leq 1),$$

where

$$\beta(p,\ q) = \int_0^1 x^{(p-1)}(1 - x)^{(q-1)} dx$$

(Johnson and Kotz 1970). Figure 1 shows the different shapes that this distribution can take, depending on the parameters p and q. When $p = q = 1$, the β distribution simplifies to the uniform distribution (fig. 1*a*). When $p = q$, the distribution is symmetrical about .5, with a U shape when $p = q < 1$ (fig. 1*b*) and a bell shape when $p = q > 1$ (fig. 1*c*). Note that data generated from these symmetrical β distributions should be nonsignificant, since the mean of the distribution is .5. Nonsymmetrical β distributions are generated by setting $p \neq q$. If both p and q are <1, the distribution is bimodal (fig. 1*f* and *g*), while if at least one is >1, the distribution is unimodal (fig. 1*d*, *e*, and *h*). When $p < 1$ and $q \geq 1$, the distribution is L-shaped (fig. 1*d*), while it is J-shaped when the reverse is true (fig. 1*e*). In all cases, the mean of the distribution is $p/(p + q)$.

The statistical power of the different tests was assessed

Table 1: Power of Fisher's procedure (FP) and the three tests presented (PA, PG, and PH) as a function of the two parameters of the β distribution, p and q, estimated over 1,000 replicates

	Bimodal β distributions											Unimodal β distributions					
	$p = .8$ $q = .9$	$p = .7$ $q = .9$	$p = .6$ $q = .9$	$p = .5$ $q = .9$	$p = .7$ $q = .8$	$p = .6$ $q = .8$	$p = .5$ $q = .8$	$p = .6$ $q = .7$	$p = .5$ $q = .7$	$p = .5$ $q = .6$	$p = .3$ $q = .4$	$p = 1$ $q = 4$	$p = 2$ $q = 8$	$p = 20$ $q = 80$	$p = 2$ $q = 3$	$p = 4$ $q = 6$	$p = 40$ $q = 60$
$n = 30$:																	
FP:																	
H_1	.258	.534	.778	.961	.385	.681	.913	.552	.857	.715	.977	1.000	1.000	1.000	.028	.000	.000
H_1	.049	.014	.006	.004	.072	.033	.009	.127	.045	.215	.496	.000	.000	.000	.000	.000	.000
PG:																	
H_1	.142	.319	.548	.825	.162	.361	.659	.173	.428	.200	.312	1.000	1.000	1.000	.844	.971	1.000
H_1	.014	.001	.000	.000	.015	.001	.001	.009	.001	.009	.002	.000	.000	.000	.000	.000	.000
PA:																	
H_1	.140	.294	.511	.773	.153	.325	.608	.149	.392	.175	.264	1.000	1.000	1.000	.832	.970	1.000
PH:																	
H_1	.142	.282	.457	.720	.142	.311	.569	.164	.372	.190	.250	1.000	1.000	1.000	.819	.968	1.000
$n = 100$:																	
FP:																	
H_1	…	.904	…	…	.728	…	…	…	…	.987	1.000						
H_1	…	.005	…	…	.074	…	…	…	…	.344	.765						
PG:																	
H_1	…	.667	…	…	.319	…	…	…	…	.414	.697						
H_1	…	.000	…	…	.004	…	…	…	…	.002	.000						
Mean	.471	.438	.4	.357	.467	.429	.385	.462	.417	.455	.429	.2	.2	.2	.4	.4	.4

Note: PA, test based on arithmetic means; PG, test based on geometric means; PH, test based on harmonic means.

Table 2: Tests of positive association between number of species and key innovation

Clade	Key innovation	n	FP	PG	Estimated parameters of β distributions and SEs				Reference
					p	SE	q	SE	
Insects	Phytophagy	13	.000	.003	.53	.05	3.04	.42	Mitter et al. 1988
Arthropods	Ovipositor, amnion, and complex chorion	14	.000	.004	.58	.05	8.65	1.10	Zeh et al. 1989
Insects	Carnivorous parasitism	15	.790	.948	.52	.07	.33	.03	Wiegman et al. 1993
Fish	Viviparity	10	.268	.310	.60	.11	.94	.22	Slowinski and Guyer 1993
Fish	Viviparity	10	.045	.177	.41	.07	.75	.17	Slowinski and Guyer 1993
Birds	Dichromatism	31	.064	.092	.78	.03	1.18	.06	Barraclough et al. 1995
Angiosperms	Floral nectar spurs	6	.000	.017	.90	.26	24.65	1.70	Hodges and Arnold 1995
Angiosperms	Branch length	39	.001	.084	.44	.01	.61	.02	Barraclough et al. 1996
Angiosperms	Branch length	56	.000	.020	.46	.01	.68	.02	Savolainen and Goudet 1998
Monocotyledons	Branch length	27	.005	.510	.35	.02	.40	.03	Savolainen and Goudet 1998
Angiosperms	Branch length	39	.007	.301	.59	.03	.85	.06	Savolainen and Goudet 1998

by generating 1000 replicates of a number of P values drawn from β distributions, and it is reported in the following as the proportion of replicates rejecting the null hypotheses at the nominal $\alpha = .05$ level.

Results and Discussion

To check whether the different tests suggested are suitable for testing the null hypothesis H_{0g} that the distribution of p_i is symmetrical about .5 against the alternative hypothesis H_{1+} that it is skewed to the right (implying a mean P value <.5), it is first necessary to show that when the null hypothesis is true, no more than 100α% of the replicates give significant results at the 100α% level. The results are depicted in figure 2, in which the X-axis represents increasing values of $p = q$.

While Type I errors for PA, PG, and PH are close to the nominal 5% in all cases, the Type I errors for FP change from 5% when the distribution of P values is uniform to 100% when $p = q = .1$ and to <5% when $p = q$ is >1. Furthermore, when the tests are carried out against the alternative hypothesis H_{1-} (that there is an excess of P values >.5), Type I errors for PA, PG, and PH are again close to the nominal 5% as expected (data not shown), while for FP they are too high when the distribution of P values is U-shaped and too low when it is bell-shaped. Using FP, therefore, we would conclude from the same data set that when the distribution is U-shaped, the mean P value is both significantly less than and more than .5 (unduly large Type I error), while we would never conclude to a significant result if the distribution is bell-shaped (too conservative, since the experimental Type I error should be close to the nominal level α).

Table 1 shows the results of power analysis when the underlying β distribution is asymmetrical and either bimodal or unimodal. Again FP is too powerful when the distribution of P values is bimodal (e.g., when $p = .3$ and $q = .4$, 977 replicates out of 1,000 gave a significant result when H_{0g} was tested against H_{1+}, and 496 replicates gave a significant result when tested against H_{1-}), while it fails to reject the null hypothesis when the distribution is bell-shaped with the mean equal to .4. However, the randomization tests appropriately reject the null hypothesis (among them, PG is nearly always the most powerful; table 1). Power increases as the mean of the distribution of the P values gets farther from .5 and as the modes get narrower. It also increases as the number of P values used gets larger (bottom half of table 1), as expected.

Eleven published data sets on key innovations were analyzed using both FP and PG. Additionally, estimates of the β distribution parameters p and q best fitting the observed P values were obtained by the procedure NLS (nonlinear least squares) of the statistical computer package S-Plus (Statistical Sciences 1995). The results are presented in table 2. Out of the 11 data sets, seven yield estimates of the two β-distribution parameters <1 (thus, a U-shaped

distribution of *P* values), while the remaining four are L-shaped. In four of these seven U-shaped data sets, FP gives an overall significant positive association between number of species and key innovation, while PG does not. In many of these data sets, therefore, increased species richness cannot be shown to result from a key innovation. In the four remaining L-shaped data sets, both FP and PG give significant or marginally significant results, as expected.

Why are there so many U-shaped distributions? There is likely no simple answer to this question. For the effect of gene sequence evolution on rates of speciation in flowering plants, Savolainen and Goudet (1998) suggested three possible explanations. One has to do with the intensity of the species sampling, the second with the taxonomy employed, and the third with the accuracy of the phylogeny.

The tests presented here are by no means to be considered as the state of the art in testing symmetry. The statistical literature on this topic is a highly active and debated area (e.g., Antille et al. 1982; Dykstra et al. 1995; Bhattacharya 1997). It remains that PG is able to maintain the nominal level over a large class of symmetric distributions and seems quite powerful against asymmetric alternatives.

I have emphasized in this note why FP is inadequate to test for the effect of key innovation on species richness and showed that testing symmetry is the appropriate way to verify or refute this hypothesis. Many other applications spring to mind. For example, population geneticists often score many loci to test whether there is an excess or a deficiency of heterozygotes. Exact tests for random association of alleles at each locus exist (e.g., Rousset and Raymond 1995). When combining the results of these tests across loci, two questions arise. The first one is concerned with the probability that all the loci are compatible with the null model of random mating. In this situation, applying FP is appropriate. The second pertinent question is whether the combined data give any evidence for a significant heterozygote deficiency. In this situation, FP would be inappropriate, and testing the symmetry of the *P* values distribution about .5 is adequate. It is therefore crucial, when combining probabilities, that investigators state their null hypothesis extremely carefully.

Acknowledgments

I am indebted to F. Balloux, T. DeMeeüs, T. Garland, C. Guyer, J. Hausser, L. Keller, Y. Naciri-Graven, N. Perrin, M. Raymond, J. Slowinski, and two anonymous referees for discussing these ideas and for their comments that greatly improved the manuscript. In particular, one referee was instrumental in allowing me to formulate more clearly the modified null hypothesis. This work was supported by grants 31-43443.95 and 31-55945.98 of the Swiss National Science Foundation.

Literature Cited

Antille, A., G. Kersting, and W. Zucchini. 1982. Testing symmetry. Journal of the American Statistical Association 77:639–646.

Barraclough, T. G., P. H. Harvey, and S. Nee. 1995. Sexual selection and taxonomic diversity in passerine birds. Proceedings of the Royal Society of London B, Biological Sciences 259:211–215.

———. 1996. Rate of rbcL gene sequence evolution and species diversification in flowering plants (angiosperms). Proceedings of the Royal Society of London B, Biological Sciences 263:589–591.

Bhattacharya, B. 1997. On tests of symmetry against one-sided alternatives. Annals of the Institute of Statistical Mathematics 49:237–254.

Dwass, M. 1957. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 28:181–187.

Dykstra, R., S. Kochar, and T. Robertson. 1995. Likelihood ratio tests for symmetry against one-sided alternatives. Annals of the Institute of Statistical Mathematics 47: 719–730.

Farris, J. S. 1976. Expected asymmetry of phylogenetic trees. Systematic Zoology 25:196–198.

Fisher, R. A. 1970. Statistical methods for research workers. 14th ed. Oliver & Boyd, Edinburgh.

Hodges, S. A., and M. L. Arnold. 1995. Spurring plant diversification: are floral nectar spurs a key innovation? Proceedings of the Royal Society of London B, Biological Sciences 262:343–348.

Hope, A. C. A. 1968. A simplified Monte-Carlo significance test procedure. Journal of the Royal Statistical Society Series B 30:582–598.

Johnson, N. L., and S. Kotz. 1970. Continuous univariate distributions. Vol. 2. Houghton Mifflin, Boston.

Losos, J. B., and F. R. Adler. 1995. Stumped by trees? a generalized null model for patterns of organismal diversity. American Naturalist 145:329–342.

Manly, B. F. J. 1985. The statistics of natural selection on animal populations. Chapman & Hall, London.

Mitter C., B. Farrel, and B. Wiegman. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? American Naturalist 132: 107–128.

Rousset, F., and M. Raymond. 1995. Testing heterozygote excess and deficiency. Genetics 140:1413–1419.

Sanderson, M. J., and M. J. Donoghue. 1996. Reconstructing shifts in diversification rates on phylogenetic trees. Trends in Ecology & Evolution 11:15–20.

Savolainen, V., and J. Goudet. 1998. Rate of gene sequence evolution and species diversification in flowering plants: a reevaluation. Proceedings of the Royal Society London B, Biological Sciences 265:603–607.

Slowinski, J. B., and C. Guyer. 1993. Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. American Naturalist 142:1019–1024.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. 3d ed. Freeman, New York.

Statistical Sciences. 1995. S-Plus user's manual, version 3.3 for Windows. Statistical Sciences, Seattle.

Wiegman, B. M., C. Mitter, and B. Farrell. 1993. Diversification of carnivorous parasitic insects: extraordinary radiation or specialized dead end? American Naturalist 142:737–754.

Zeh, D. W., J. A. Zeh, and R. L. Smith. 1989. Ovipositors, amnions and eggshell architecture in the diversification of terrestrial arthropods. Quarterly Review of Biology 64:147–167.

Associate Editor: Theodore Garland, Jr.