Evolution, 57(1), 2003, pp. 18–26

PHYLOGENETICALLY NESTED COMPARISONS FOR TESTING CORRELATES OF SPECIES RICHNESS: A SIMULATION STUDY OF CONTINUOUS VARIABLES

NICK J. B. ISAAC,[1,2] PAUL-MICHAEL AGAPOW,[1,3] PAUL H. HARVEY,[4] AND ANDY PURVIS[1,5]
[1]*Department of Biological Sciences, Imperial College, Silwood Park, Ascot, Berks SL5 7PY, United Kingdom*
[4]*Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom*
[5]*E-mail: a.purvis@ic.ac.uk*

Abstract.—Explaining the uneven distribution of species among lineages is one of the oldest questions in evolution. Proposed correlations between biological traits and species diversity are routinely tested by making comparisons between phylogenetic sister clades. Several recent studies have used nested sister-clade comparisons to test hypotheses linking continuously varying traits, such as body size, with diversity. Evaluating the findings of these studies is complicated because they differ in the index of species richness difference used, the way in which trait differences were treated, and the statistical tests employed. In this paper, we use simulations to compare the performance of four species richness indices, two choices about the branch lengths used to estimate trait values for internal nodes and two statistical tests under a range of models of clade growth and character evolution. All four indices returned appropriate Type I error rates when the assumptions of the method were met and when branch lengths were set proportional to time. Only two of the indices were robust to the different evolutionary models and to different choices of branch lengths and statistical tests. These robust indices had comparable power under one nonnull scenario. Regression through the origin was consistently more powerful than the *t*-test, and the choice of branch lengths exerts a strong effect on both the validity and power. In the light of our simulations, we re-evaluate the findings of those who have previously used nested comparisons in the context of species richness. We provide a set of simple guidelines to maximize the performance of phylogenetically nested comparisons in tests of putative correlates of species richness.

Key words.—Comparative methods, MacroCAIC, macroevolution, phylogeny, sister-clade comparisons, species richness, validity.

Received December 3, 2001. Accepted September 10, 2002.

Contemporary clades differ so much in species richness that species must have differed significantly in their chances of speciating, their chances of going extinct, or of both (Dial and Marzluff 1989; Nee et al. 1996; Purvis 1996; Mooers and Heard 1997). What factors have made species more likely to speciate or to go extinct? This question, of central importance in evolutionary biology, is not simple to tackle (Schluter 2000). If we knew the full history of a clade—the complete pattern of speciation, extinction, and changes in the species' characteristics—it would be a relatively simple task to test hypotheses relating diversity to particular attributes of species. However, despite a few notable exceptions (e.g., Jablonski 1996; Smith and Jeffery 1998), the dearth of such direct information about the history of most clades means that biologists are typically restricted to testing hypotheses by correlating the current diversity of clades with the attributes of their extant species. Figure 1 illustrates the sort of dataset that might be available, with species richness (*S*) and trait values (*X*) for a number of related clades.

A common procedure has been to choose some taxonomic level for analysis, say the family level, and then compare species richness and the trait of interest across families, treating them as independent points for the purposes of analysis (e.g., Van Valen 1973; Dial and Marzluff 1988; Kochmer and Wagner 1988; Martin 1992; Ricklefs and Renner 1994). This practice has three main pitfalls: nonmonophyly, noncomparability, and nonindependence (Purvis 1996; Barraclough et al. 1998; Dodd et al. 1999). Comparisons among nonmonophyletic taxa are obviously hard to interpret. Nonmonophyly is always a potential problem, because estimation of phylogeny always runs the risk of error, but the problem is much more prevalent with the many taxonomic classifications that are not intended to reflect phylogeny accurately (Harvey and Pagel 1991). Even comparisons of monophyletic taxa of the same rank are problematic, however. They can differ greatly in age and therefore in the time they have had to diversify: the *melanogaster* subgroup of the genus *Drosophila*, for instance, is older than the most recent common ancestor of all extant apes (Avise and Johns 1999), and equal-aged primate lineages range in taxonomic level from subgenus to superfamily (Purvis et al. 1995). Lastly, related taxa may inherit determinants of diversification rate from a common ancestor, rather than evolve them independently; pseudoreplication is as much a problem here as in other sorts of comparative study (Felsenstein 1985; Harvey and Pagel 1991).

All three pitfalls can be avoided by basing comparisons on phylogeny (Barraclough et al. 1998), at least if the phylogeny is correct. Clades, unlike taxa, are monophyletic by definition. Likewise, sister clades, unlike taxa of a given rank, are the same age. In Figure 1, taxon C (which has a large value of trait *X*) has more species than either A or B, suggesting a positive correlation between trait *X* and diversity; but phylogenetic comparison shows that C actually has fewer species than its sister clade. Lastly, comparisons between sister families are sure to be mutually independent.

The recognition of the problems of nonphylogenetic analysis and the increasing availability of phylogenies have led to sister-clade comparisons now being routinely used when testing hypotheses linking diversity to particular values of discrete characters. Such studies proceed in one of two ways.

[2] Present address: Institute of Zoology, Zoological Society of London Regent's Park, London NW1 4RY, United Kingdom.

[3] Present address: Department of Biology, University College London, Gower Street, London WC1E 6BT, United Kingdom.

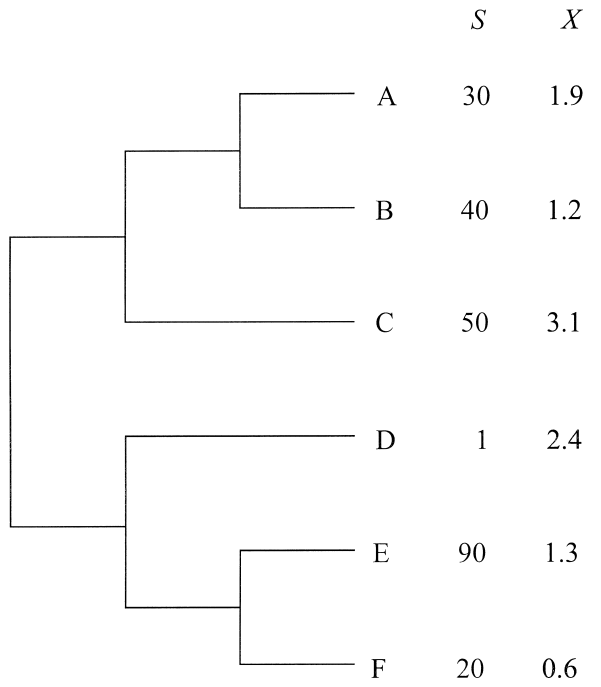

FIG. 1. Hypothetical dataset and phylogeny for six taxa (A–F) that differ in species richness (*S*) and a trait (*X*).

If the trait under consideration changes only rarely, its phylogenetic distribution is mapped to identify sister clades having different character states (e.g., Mitter et al. 1988; Zeh et al. 1989; Farrell 1998; Arnqvist et al. 2000). If the trait is more labile, a particular taxonomic level is chosen for analysis, and pairs of sister taxa compared to see whether the more diverse clade has a higher or lower prevalence of the character state proposed to increase diversity (e.g., Barraclough et al. 1995; Owens et al. 1999). Both approaches lead to straightforward tests of the null hypothesis of no association between character states and diversity. The replicates are sure to be independent, because they consider totally separate parts of the phylogeny, with no species being used in more than one comparison. There is no reason to expect homogeneity of variance among comparisons, however, so testing is typically nonparametric.

The use of sister-clade comparisons is less straightforward when *X* is a continuous variable like body size. Extant species all differ in their body mass because the trait has evolved along almost every branch of the phylogeny. This means that all pairs of sister clades differ and it is no longer obvious which comparisons to make. In Figure 1 there are five possible sets of independent comparisons. One choice, A versus B and E versus F, suggests a positive relationship between *X* and diversity. Another set, (A + B) versus C and D versus (E + F), suggests a negative relationship. The remaining three choices, A versus B and D versus (E + F), (A + B) versus C and E versus F, and (A + B + C) versus (D + E + F), are equivocal. Which set is to be preferred? As a result of such uncertainty, tests of hypotheses relating diversity to continuous variables have lagged behind those for discrete traits (Purvis 1996).

One resolution of the problem is to compare each pair of sister taxa; in Figure 1, the comparisons are A versus B, (A + B) versus C, E versus F, D versus (E + F), and (A + B + C) versus (D + E + F). Unlike the sets listed above, these comparisons are nested, that is, comparisons are made both within and between clades, using information from the tips to make comparisons deeper in the phylogeny. This process involves making assumptions about how evolution has proceeded (see below) but nested comparisons have two obvious advantages: there is no need for an arbitrary decision among possible sets of comparisons, and more comparisons are made. One comparison is made at each bifurcation in the tree that has more than two species ultimately descended from it: polytomies typically represent ambiguity about branching structure (Maddison 1989) so should not yield comparisons, and comparisons between sister species cannot bear on the hypothesis.

To compare clade (A + B) with clade C, we need to compute the values of *S* and *X* for clade (A + B). S_{A+B} is simply the sum of the species richnesses of A and B, but estimation of X_{A+B} requires a model of character evolution. Brownian motion (BM) is the most commonly used model in comparative tests of correlated evolution among traits (Felsenstein 1985), although more complex and realistic models are increasingly being used (Martins 1998; Pagel 1999). The model assumes that each lineage has evolved in isolation from others. If this is an adequate description of trait evolution, the resulting ΔX values will be mutually independent (Felsenstein 1985). Furthermore, the model specifies how variance accumulates over branch length, permitting scaling of the ΔX values to give constant variance and thus allowing parametric statistics to be used.

If, in the absence of any effect of *X*, all species have equal chance of speciating and there is no extinction (the Yule process; Yule 1924), then the rate of clade growth can be estimated simply as $r = \ln(S)/t$, where *t* is the age of the clade (for other ways of estimating speciation rates see Nee 2001). For the comparison between clades A and B, for instance, we can compute the absolute rate difference (ARD) as $r_A - r_B = \ln(S_A/S_B)/t$. Under the Yule model, the ARDs are mutually independent, with a variance that scales approximately as $1/t^2$ when *t* is large. If a linear relationship between *r* and *X* is posited, such that $r = a + bX + \varepsilon$ (where ε is random error), then a suitably weighted linear regression through the origin of ARD on ΔX might be used to assess whether *b* is significantly different from zero. Under these circumstances, the nested comparisons in both *S* and *X* are independent under the null hypothesis.

However, real datasets are likely to violate the assumptions of Brownian motion and the Yule process, on which independence and homogeneity of variance depend. Such violations will affect the validity of using nested comparisons.

First, the models assumed for character evolution may be inadequate. If so, the ΔX will lose both independence and homogeneity of variance (Pagel and Harvey 1992). The effects of misspecifying the model of character change have been investigated thoroughly in the context of tests of correlated evolution among characters (e.g., Martins and Garland 1991; Díaz-Uriarte and Garland 1996; Harvey and Rambaut 2000). Type I error rates (probability of rejecting a true null hypothesis) are indeed elevated, but only moderately under

most models considered, especially when statistical model criticism is applied when analyzing the comparisons (but see Harvey and Rambaut 2000).

Second, phylogenetic branch lengths are not known without error and may not be known at all. Errors in branch lengths effectively equate to misspecification of the model of character change, with the effects noted above. Simulations (Grafen 1989; Martins and Garland 1991; Purvis et al. 1994; Díaz-Uriarte and Garland 1996; Ackerly 2000) again indicate only moderate elevation of Type I error rates. Errors in estimates of when lineages split will also cause heterogeneity of variance in ΔS.

Third, extinction rates are probably seldom if ever zero. If extinction is nonzero, r will tend to be larger when t is small than when t is large (Harvey et al. 1994; Kubo and Iwasa 1995), leading to a similar trend in the variance of ARD. Extinction need not affect the independence of the ΔS: they are still independent under a constant-rates birth-death process, which we henceforth refer to as the Markov model.

Fourth, it is unlikely that the trait under study will be the only factor that has shaped diversification rates. If diversification probabilities have differed among clades, the ΔS may lose independence.

In this paper, we report simulations designed to assess the validity of nested comparisons when these assumptions are violated singly and in combinations. We also report simulations to assess power under one scenario in which the trait being tested does indeed shape S. Our simulations examine the performance of a range of measures of ΔS and ΔX and two ways of testing for an association between them to assess their robustness against violations. Our aim is to determine which method of making nested comparisons is most robust and, of the robust methods, which has greatest power to detect true correlations. We then discuss previous uses of nested comparisons (Gittleman and Purvis 1998; Gardezi and da Silva 1999; Desdevises et al. 2001; Katzourakis et al. 2001; Orme et al. 2002) to assess whether the methods used were the best available and, if not, whether the significant results of those studies are robust to reanalysis.

Simulations to Compare Type I Error Rates

The basic simulation procedure, outlined in greater detail below, was to generate phylogenies and data according to specific evolutionary models, then analyze these using a variety of indices and statistical tests. One thousand phylogenies and datasets were generated for each combination of three models of clade growth and two models of trait evolution (6000 datasets in total). Each dataset was analyzed by the nested sister-clade method under a range of data treatment options. Associations between a trait, X, and species richness, S, were tested for by comparing the diversity difference, ΔS, with phylogenetically independent contrasts in the trait, ΔX.

Models of Evolution

Three models of stochastic clade growth were used to produce phylogenies of uniform size (250 species) that varied greatly in symmetry. The degree of symmetry affects both the distribution of ΔS (however measured) and the proportion of comparisons that are informative (i.e., the proportion of nodes with more than two descendant species). The models were the Markov model and two variants in which the speciation rate of each lineage was a step function based on lineage age. In all cases the probability of speciation was independent of the lineage's value of X.

The instantaneous speciation rate under the Markov model was set at 0.1 per time unit for all extant lineages. Mean speciation rate in the first variant was also 0.1. However, this was distributed such that the youngest 5% of lineages had a speciation rate eight times higher than the remainder, with both daughters of a speciation event considered to be newly born. The values were chosen such that the mean balance score (B_1 statistic, Shao and Sokal 1990; Kirkpatrick and Slatkin 1993) of phylogenies produced by this model was approximately six standard deviations lower than that produced by the Markov model. The third model incorporated a refractory period (Losos and Adler 1995) to produce phylogenies that were more balanced than those from the Markov model: the instantaneous speciation rate for lineages younger than the refractory period was zero, rising to 0.1 in a single step thereafter. Again, both daughters of a speciation event were considered to be newly born. A refractory period of six arbitrary time units was found to increase the mean balance score by six standard deviations compared to Markov model.

All three clade growth models included a regime of background extinction. We present results in which the instantaneous extinction rate for all lineages in all models was 0.075 per unit time. Results for other extinction rates (0.0, 0.025, 0.05 and 0.9) were not qualitatively different, and are not described further.

The stochastic process of speciation, extinction, and trait evolution was terminated immediately before the birth of the 251st species, thus allowing closely related species some time for phenotypic divergence. This produced phylogenies that were comparable in size and balance score with those used previously for nested comparisons with species-level phylogenies (271 carnivore species, 203 primate species; Gittleman and Purvis 1998).

Species values of X were generated by two models of trait evolution: BM and speciational Brownian (SB; Garland et al. 1993). BM was modeled by simulating the effect of numerous small but independent changes such that the change along any given branch is drawn from a normal distribution with a mean of zero and variance equal to the branch length (Felsenstein 1985). Under SB, there are saltatory changes in both daughter lineages at speciation events, the magnitude of change being drawn from a standard normal distribution. Generally, the character X therefore evolves down the tree according to $X_{\text{daughter}} = X_{\text{ancestor}} + c\varepsilon$, where ε is a standard normal deviate and c is the square root of the branch length in BM and unity in SB.

Measuring ΔS

We present four candidate indices of ΔS, all symmetrically distributed around zero under the null hypothesis (Fig. 2). All are based upon sister clades containing S_i and S_j species (no comparisons are made at polytomies) and are calculated relative to the predictive trait, X, such that that S_i identifies the clade with the larger value of X at any given node.

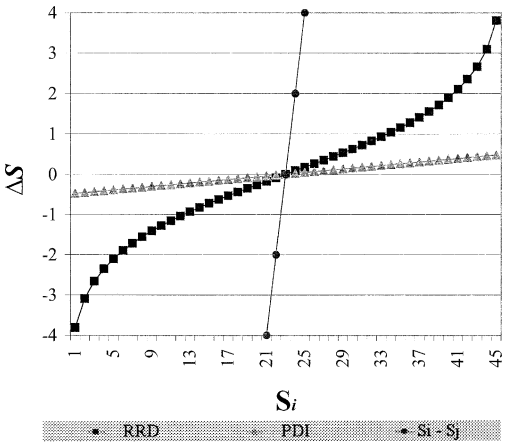

FIG. 2. Distribution of relative rate difference (RRD, square symbols), proportional dominance index (triangles), and $S_i - S_j$ (circles) at a node of 46 species. At any given node, absolute rate difference (ARD) = RRD/t, where t is the age of the node.

As explained in the introduction, our first index is ARD, given by $\ln(S_i/S_j)/t$, with units of time^{-1}. ARD is not calculable when node ages are unknown. Our second index is the relative, rather than absolute, difference in the rate of clade growth between sister clades. By removing the focus on absolute rates, we relax the assumption of density independent clade growth and obviate the need for branch length information. Relative rate difference (RRD) is dimensionless and given by $\ln(S_i/S_j)$. RRD has approximately constant variance when $S_i + S_j > 10$, provided that the values of S_i and S_j are drawn from a geometric distribution (S. Nee, pers. comm.), as they are under the Markov model.

Our third index, which we term the proportional dominance index (PDI), is given by $S_i/(S_i + S_j) - 0.5$. This takes discrete values but approximates to a uniform distribution between -0.5 and $+0.5$ under the null hypothesis and the Markov model, because the possible splits at a node containing S species (i.e., $S - 1$:1, $S - 2$:2 . . . 1:$S - 1$) are all equally probable in this scenario (Farris 1976). PDI gives equal weighting to every informative node in the phylogeny and is dimensionless.

Our final index is the simple difference in species numbers, $S_i - S_j$. This gives the same weight to nodal split of 11:1 as to 100:90, even though the two splits represent a 10-fold difference in the relative rates of cladogenesis. It will tend to give higher weight to more basal splits, as the expectation of $S_i - S_j$ obviously increases with $S_i + S_j$.

Statistics, Analysis, and Calculation of the Error Rates

Differences in trait values (ΔX) were calculated as in Felsenstein's (1985) method of independent contrasts. Contrasts were generated both using branch lengths proportional to time (henceforth ''real'') and set equal to unity (''equal''; these were computed for the phylogeny linking extant species only, the nodes and branches leading to extinct species having first been collapsed). Equal branch lengths are commonly used in comparative studies when dates of lineage splits are unavailable (Purvis and Webster 1999; Ackerly 2000), so their performance is of interest. Furthermore, the length of internal branches is irrelevant in the evolution of trait values under the SB or a punctuational model of trait evolution with no extinction. Equal branch lengths are expected to outperform real branch lengths in that case.

Statistical associations between ΔX and ΔS were sought by least-squares regression through the origin (Garland et al. 1992), which uses the magnitudes of both sets of differences, and by one-sample t-test (against $\Delta S = 0$), which uses the magnitudes of only the ΔS. Each dataset was analyzed using all four measures of ΔS, both statistical techniques and both branch length options (16 combinations). The Type I error rate for each combination of evolutionary model and analysis options (120 in all) was calculated as the proportion of 1000 replicates in which a significant relationship between ΔX and ΔS was detected at $\alpha = 0.05$.

TEST OF POWER

A modified simulation procedure was used to compare the power of the various options for data analysis under one scenario, chosen for simplicity of simulation. An artificial correlation between species richness differences and a test character, Z, was generated using a biased sampling event that is analogous to a contemporary mass extinction event. One thousand replicates were performed as follows. A Markovian phylogeny of 500 species was generated, and two traits, X and Z, evolved by BM along it. Species were then selected at random for possible extinction. If selected, their probability of extinction depended linearly on their rank for Z; thus, the species with the largest value of Z had an extinction probability of 1.0, the median species had probability 0.5, and the species with the lowest value of Z had probability zero. Selection continued until only 250 species remained, at which point the contrasts were computed for ΔS, ΔX, and ΔZ. The power was measured as the proportion of replicates in which ΔS was significantly correlated with ΔZ minus the proportion in which ΔS was significantly correlated with ΔX. Each replicate was analyzed using all combinations of the four indices of ΔS, the two statistical tests, and the two branch length settings to estimate the relative power of each.

RESULTS

All four indices of ΔS give the nominal Type I error rates when the tree is Markovian, X evolves by BM, and the data are analyzed by regression through the origin using real branch lengths (binomial test $P > 0.5$ in all cases). However, the four indices perform very differently when these assumptions are violated and when data are analyzed in different ways. Comparison of the coefficient of variation in error rates across all evolutionary models and analysis options (Table 1) reveals that two indices of ΔS are highly susceptible to the simulation parameters and that two are relatively robust.

The susceptible pair, ARD and $S_i - S_j$, have Type I error rates that are elevated ($P < 0.01$ in both cases) and highly variable (coefficient of variation > 0.5), with some very high error rates. These properties make them unsuitable for testing hypotheses about species richness. The robust pair, RRD and PDI, have error rates that overall are not significantly different from $\alpha = 0.05$ on average ($P > 0.5$ in both cases) and

TABLE 1. Comparative performance of the four indices of ΔS. ''Best case'' is the Type I error rate when all assumptions are met (Markovian clade growth and gradual trait evolution) analyzed with branch lengths proportional to time and regression through the origin. ''Mean'' refers to the mean error rate across all combinations of evolutionary models and analysis options, and ''CV'' is the associated coefficient of variation. There are 24 combinations for relative rate difference, proportional dominance index, and $S_i - S_j$ (three clade growth models, two models of trait evolution, two statistical procedures, and two branch length settings) and 12 combinations for absolute rate difference (which can only be calculated when branch length information is used).

ΔS index	Best case	Highest	Mean	CV	Diagnosis
ARD	0.052	0.511	0.084	1.44	susceptible
RRD	0.051	0.106	0.051	0.35	robust
PDI	0.050	0.109	0.048	0.34	robust
$S_i - S_j$	0.047	0.166	0.063	0.83	susceptible

that are much less variable (Table 1). The most elevated error rates among PDI and RRD occurred when the *t*-test was used to analyze data that had evolved by the SB model. The highest error rate found when using regression with the robust indices was 0.076 using equal branch lengths and 0.063 when set proportional to time (Table 2).

The effects of the simulation parameters on the error rates were explored using generalized linear models with binomial error structure, analyzed using GLMStat 5.6.1 (Beath 2001). The full model for RRD is presented in the Appendix; that for PDI is qualitatively identical (not shown). The only parameter that increased the Type I error rate of RRD as a main effect was the choice of branch lengths; setting all branch lengths equal resulted in a higher error rate than using real branch lengths ($Z_{15} = 6.46$, $P < 0.0001$). Other factors affected the error rates in combination with one another and greatly contributed to the variation in error rates. These interaction terms indicate conditions under which the Type I error rates departed from the mean value of 0.051 (Table 1). The model of trait evolution interacts with the model of clade growth ($Z_{15} = 3.68$, $P < 0.001$) and whether analysis is by *t*-test or regression ($Z_{15} = 3.52$, $P < 0.001$). Positive *Z*-scores reflect the fact that the most unbalanced (asymmetrical) phylogenies and analysis by *t*-test both had inflated Type I error rates under the SB model but were conservative under BM. Both clade growth model and the choice of statistics also interacted with the choice of branch lengths to affect the variability in error rate ($P < 0.0001$ in both cases).

TABLE 3. Power of different analysis options at $\alpha = 0.05$ using a biased sampling regime. R, regression through the origin; T, one-sample *t*-test (vs. mean $\Delta S = 0$).

Branch lengths	Statistics	ARD	RRD	PDI	$S_i - S_j$	Mean
Real	R	10	193	186	100	133.8
	T	20	128	127	65	90.4
Equal	R	n/a	209	192	96	173.5
	T	n/a	82	81	21	57
Mean		15	153	146.5	70.5	n/a

Results from the test of power (Table 3) revealed that the susceptible measures of ΔS have very low power. RRD is consistently more powerful than PDI, but this difference is not significant ($Z_3 = 0.82$, $P = 0.41$). Regression through the origin was more powerful than the *t*-test across both indices ($Z_4 = 5.34$, $P < 0.0001$). Interaction terms suggest that regression analysis is also insensitive to the choice of branch lengths ($Z_4 = 0.878$, $P = 0.38$), but the *t*-test has lower power with equal branch lengths ($Z_4 = -0.569$, $P < 0.0001$). The

TABLE 2. Type I error rates at $\alpha = 0.05$ for four indices of ΔS across all combinations of evolutionary models and analysis options. R, regression through the origin; T, one-sample *t*-test (vs. mean $\Delta S = 0$). Numbers in bold are elevated, numbers in italics are conservative; other rates are not significantly different from $\alpha = 0.05$ (binomial test).

Clade growth	Trait evolution	Branch lengths	Stats	ARD	RRD	PDI	$S_i - S_j$
Balanced	BM	real	R	0.047	0.039	0.043	0.049
			T	0.049	0.042	0.046	*0.035*
		equal	R	n/a	**0.076**	0.058	**0.160**
			T	n/a	*0.034*	*0.035*	*0.024*
	SB	real	R	0.05	*0.036*	*0.036*	*0.033*
			T	0.044	0.044	0.045	*0.031*
		equal	R	n/a	**0.072**	0.053	**0.114**
			T	n/a	*0.036*	0.038	*0.026*
Markovian	BM	real	R	0.052	0.051	0.05	0.047
			T	0.045	0.042	0.044	0.037
		equal	R	n/a	**0.064**	0.042	**0.141**
			T	n/a	*0.029*	*0.028*	*0.023*
	SB	real	R	**0.439**	0.040	0.043	*0.021*
			T	0.051	0.061	0.060	0.051
		equal	R	n/a	**0.065**	0.055	**0.151**
			T	n/a	0.052	0.043	*0.034*
Unbalanced	BM	real	R	0.042	0.043	0.048	0.043
			T	*0.026*	0.047	0.053	*0.033*
		equal	R	n/a	0.044	*0.029*	**0.162**
			T	n/a	*0.026*	*0.025*	*0.015*
	SB	real	R	**0.511**	0.047	0.067	*0.021*
			T	0.053	**0.106**	**0.109**	**0.067**
		equal	R	n/a	**0.067**	0.047	**0.166**
			T	n/a	0.046	0.048	*0.032*

complete minimum adequate model for the test of power is presented in the Appendix.

Discussion

Phylogenetically nested comparisons are commonplace in studies of character evolution, but their use in tests of hypotheses relating to species richness brings concerns about independence of comparisons to the forefront. Our simulations show that the use of phylogenetically nested comparisons is valid if the model of character change is accurate and if real branch lengths are available and used, at least where speciation and extinction rates do not change systematically through time. This conclusion holds for all four indices of species richness difference, ΔS, for both regressions and *t*-tests, for all three models of clade growth and under all background extinction regimes. However, only two of the indices, RRD and PDI, still perform reasonably when the assumptions about character evolution are violated. These two indices never gave Type I error rates greater than 0.087 in our simulations. This robustness may be because ΔS and ΔX are controlled by different processes: interaction terms without significant main effects in the model for Type I error rates suggest that the assumptions of both processes have to be violated before the error rate increases markedly.

One of the rejected indices, ARD, performed badly largely because it gives high weighting to very young nodes. We therefore investigated the effect of imposing a minimum value of $S_i + S_j$ at which comparisons were made, to see whether using such a threshold resulted in a useable index. The threshold reduced the high error rate under the SB model, but error rates remained significantly elevated even with a threshold of 15 (results not shown). The ratio of comparisons to species is then about 1:7, making ARD impractical for most real species-level phylogenies. Weighting each point by the inverse of the expected variance (i.e., t^2) did not improve the performance of ARD either, even under a Yule process (results not shown). The other rejected index, $S_i - S_j$, performs badly because its variance increases with $S_i + S_j$. We explored an alternative, $\ln|S_i - S_j|$ with the sign given by the sign of $S_i - S_j$, in which this heterogeneity of variance is reduced. This alternative index had lower Type I error rates than the two rejected indices, but was much more variable than RRD or PDI (results not shown), so we do not recommend its use.

RRD and PDI were found to be equally powerful in detecting that species-richness differences were associated with the magnitude of trait *Z*. This is surprising because RRD gives higher weight to the most unbalanced nodes, whereas PDI weights all nodes equally. The suspicion remains that RRD will be more powerful in detecting correlates of species richness when speciation probabilities are heritable and evolving. This is a priority for further research. The Type I error rates of PDI and RRD were always qualitatively similar when using real branch lengths (Table 2). Their performance under equal branch lengths was very similar when analyzing with the *t*-test, but RRD always returned inflated Type I errors using regression through the origin. This suggests that PDI should be preferred, at least when nodal ages are unknown. However, the distributions of the measures may be more important in determining which should be used. Whereas PDI is bounded, the fact that RRD can take on very large values when $S_i + S_j$ is large ($>10^6$) can lead to nonconstancy of variance when the total number of comparisons is small (C.D.L. Orme, pers. comm.). When many informative nodes are available, as in the simulated phylogenies, RRD is more normally distributed than PDI (Fig. 3) and is therefore more appropriate for use in parametric statistics.

The choice of statistical analysis and branch lengths affected both the validity and the power of all measures of ΔS. The *t*-test was both less powerful and more sensitive to violations of the evolutionary assumptions than was regression through the origin. Equal branch lengths increased the Type I error rate overall and did not reduce the errors introduced by the SB model (i.e., there was no ''right model'' effect). However, equal branch lengths were relatively conservative on the most unbalanced phylogenies. This highlights the importance of measuring the shape of the phylogeny under examination, especially because the phylogenies of the real world show a tendency to be unbalanced (Heard and Mooers 2000).

We have tested the method's behavior under several departures from the assumptions of the underlying models. Other departures from BM might have a more severe impact on the Type I error rate (Díaz-Uriarte and Garland 1996; Harvey and Rambaut 2000). Rates of speciation and extinction are often variable through time (Zink and Slowinski 1995; Schluter 2000), which might affect the variance of ΔS. Such variation might not matter, given the insensitivity of our results to the background extinction rates. However, we recommend that statistical model criticism be applied to all data derived from nested comparisons. BM can be tested using diagnostic tests of the branch length standardization procedures (Garland et al. 1992; Pagel and Harvey 1992). Likewise, homogeneity of variance in ΔS can be tested by regressing ΔS on node size or node age, although other methods provide more direct and probably more powerful tests for nonconstant rates of diversification (e.g., Pybus and Harvey 2000). Other factors not addressed in these simulations also have the potential to affect the method's validity. Like more traditional uses of independent contrasts, incomplete data is likely to affect the validity of this method when taxon sampling is not random (Ackerly 2000). More serious for comparisons involving species richness is the completeness and accuracy of the phylogeny. Missing clades will bias the values of S_i and S_j and are likely to generate spurious results. Likewise, results will obviously not be reliable if the taxa being compared are not in fact sister clades.

The MacroCAIC program (Agapow and Isaac 2002) implements both robust species-richness measures described above and is available from www.bio.ic.ac.uk/evolve. It permits the use of all the described analysis options and performs the assumption checks described above. It is also able to analyze phylogenies whose terminals are superspecific taxa (e.g., genera in Katzourakis et al. 2001). We have not explicitly tested the effects of this form of analysis, although our simulations show that the Type I error rate is not increased by raising the threshold value of $S_i + S_j$ at which comparisons were made (results not shown). We believe this approach to be valid so long as comparable species concepts are applied to describe the richness of taxa under consider-

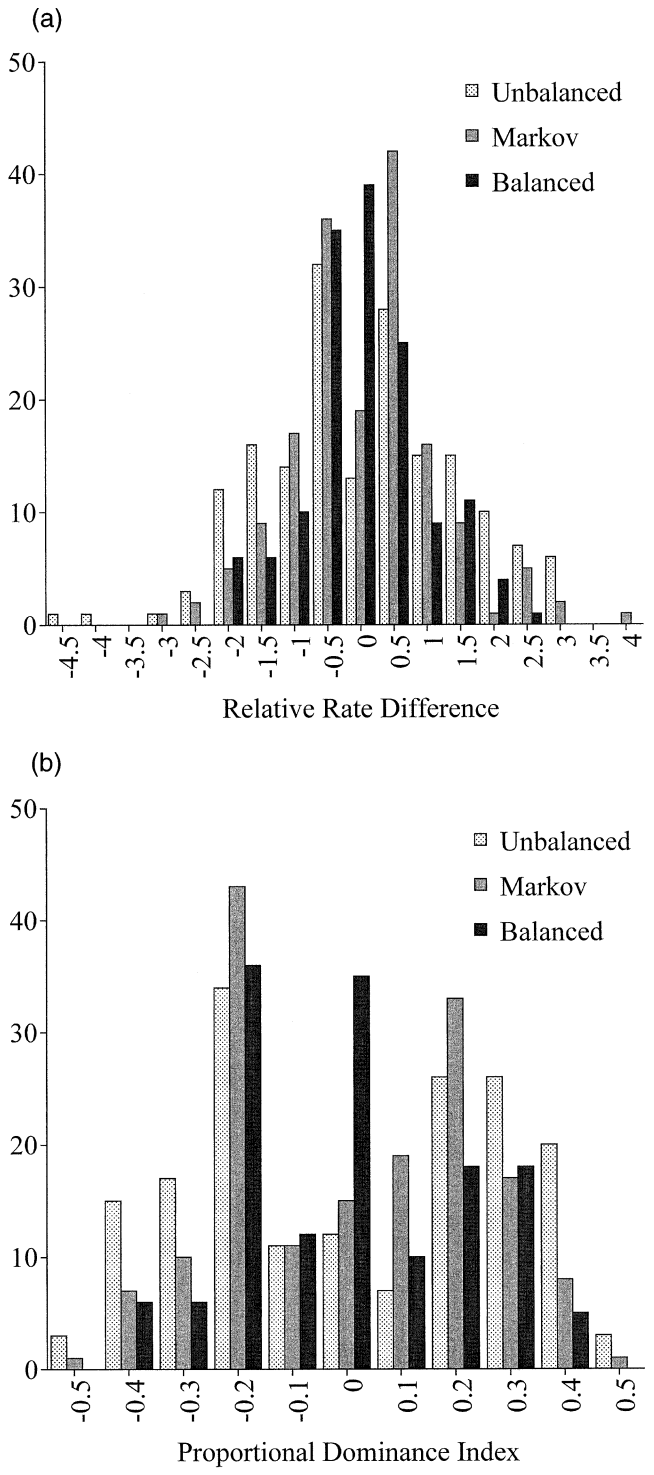

FIG. 3. Distribution of (a) relative rate difference and (b) proportional dominance index, relative to a randomly evolved Brownian motion character in balanced (n = 146 comparisons), Markovian (n = 165), and unbalanced (n = 174) simulated phylogenies of 250 species each.

ation. For instance, biases in systematic practices have been suggested as causes for the observed relationship between body size and species richness (Van Valen 1973; Kochmer and Wagner 1988).

We are now in a position to assess the robustness of published work that has used nested comparisons to search for continuous correlates of species richness. Gittleman and Purvis (1998) reported a significant negative correlation between RRD and body mass in the Carnivora using both branch length options and statistical treatments presented here. An important omission from their methods section, however, is that contrasts in body mass were not standardized. The authors reasoned that an extinction filter had shaped the carnivore phylogeny, and that the fate of sister-taxa in extinction events is likely to depend on the absolute difference in body mass rather than the rate at which that difference evolved. Simulations confirm that, under these circumstances, power was higher using raw (unstandardized) contrasts than standardized (results not shown). However, use of raw contrasts also significantly compromised the validity (results not shown). This is not surprising given the importance of branch length standardization to the validity of independent contrasts (Díaz-Uriarte and Garland 1996).

We therefore reanalyzed the carnivore dataset using standardized contrasts. The reported correlation was found to be qualitatively similar but marginally weaker than using raw contrasts (two-tailed P = 0.032 vs. 0.021), although PDI is more conservative still (P = 0.07 with real branch lengths, P = 0.16 when set equal). Katzourakis et al. (2001) used RRD to search for correlates of species richness in the hoverflies. Their use of raw contrasts is less likely to be important because all their significant results were from t-tests, which are insensitive to standardization. The shape of the phylogeny is significantly more unbalanced than Markovian (Katzourakis et al. 2001), which suggests that their t-tests are likely to have been conservative for all indices of ΔS (Table 2). Desdevises et al. (2001) and Orme et al. (2002) used standardized contrasts but found no significant relationships. Gardezi and da Silva (1999) also standardized contrasts and measured ΔS as RRD/$\sqrt{t}$. We found this index to behave much like ARD, with high Type I and Type II error rates (results not shown). Their finding that body size predicts species richness in the three lightest mammalian orders is marginally supported using RRD instead (P = 0.048 vs. 0.01). The only other uses of nested comparisons of which we are aware (Marzluff and Dial 1991; Nee et al. 1992) did not make quantitative comparisons using the species richnesses of descendent clades and are therefore not directly comparable.

Conclusions and Recommendations

We have shown that the method of nested sister clade comparisons is a reasonable approach for detecting correlates of species richness and outlined the analysis procedures that will optimize validity under a range of scenarios. We recommend: (1) that ΔS be measured as RRD when branch length information is available and the phylogeny contains at least 30 informative nodes, otherwise as PDI; (2) that statistical analysis use regression through the origin rather than a t-test; (3) that branch lengths be set proportional to time where possible; (4) the assumptions of Brownian motion and constancy of variance in ΔS and ΔX be checked; and (5) phylogenies under study be tested against the null model of constant rates. If the evolution of trait X cannot be adequately

modeled or if X is a categorical trait, we recommend the use of nonnested comparisons rather than the methods described here. We have also demonstrated how to maximize the power under one particular nonnull scenario. Future work will explore the merits of nested and nonnested comparisons, using RRD and PDI, under models of variable diversification rates.

Acknowledgments

We are grateful to A. Ø. Mooers for providing the Pascal source code used to generate phylogenies by the Markov model. We thank T. G. Barraclough, M. J. Crawley, H. C. J. Godfray, A. Katzourakis, D. L. J. Quicke, C. D. L. Orme, and D. Stephens and two anonymous reviewers for discussion and comments on previous versions of this manuscript. This work was funded by Natural Environment Research Council grant GR3/11526.

Literature Cited

Ackerly, D. D. 2000. Taxon sampling, correlated evolution, and independent contrasts. Evolution 54:1480–1492.

Agapow, P. M., and N. J. B. Isaac. 2002. MacroCAIC: Revealing correlates of species richness by comparative analysis. Divers. Distrib. 8:41–43.

Arnqvist, G., M. Edvardsson, U. Friberg, and T. Nilsson. 2000. Sexual conflict promotes speciation in insects. Proc. Natl. Acad. Sci. USA 97:10460–10464.

Avise, J. C., and G. C. Johns. 1999. Proposal for a standardized temporal scheme of biological classification for extant species. Proc. Natl. Acad. Sci. USA 96:7358–7363.

Barraclough, T. G., P. H. Harvey, and S. Nee. 1995. Sexual selection and taxonomic diversity in passerine birds. Proc. R. Soc. Lond. B 259:211–215.

Barraclough, T. G., S. Nee, and P. H. Harvey. 1998. Sister-group analysis in identifying correlates of diversification: comment. Evol. Ecol. 12:751–754.

Beath, K. J. 2001. GLMStat: generalised linear modelling software. Kagi Software, Berkeley, CA.

Desdevises, Y., S. Morand, and G. Oliver. 2001. Linking specialisation to diversification in the Diplectanidae Bychowsky 1957 (Monogenea, Platyhelminthes). Parasitol. Res. 87:223–230.

Dial, K. P., and J. M. Marzluff. 1988. Are the smallest organisms the most diverse? Ecology 69:1620–1624.

———. 1989. Nonrandom diversification within taxonomic assemblages. Syst. Zool. 38:26–37.

Díaz-Uriarte, R., and T. J. Garland. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. Syst. Biol. 45:27–47.

Dodd, M. E., J. Silvertown, and M. W. Chase. 1999. Phylogenetic analysis of trait evolution and species diversity variation among angiosperm families. Evolution 53:732–744.

Farrell, B. D. 1998. ''Inordinate fondness'' explained: Why are there so many beetles? Science 281:555–559.

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

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Gardezi, T., and J. da Silva. 1999. Diversity in relation to body size in mammals: a comparative study. Am. Nat. 153:110–123.

Garland, T., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

Garland, T., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. Syst. Biol. 42:265–292.

Gittleman, J. L., and A. Purvis. 1998. Body size and species richness in carnivores and primates. Proc. R. Soc. Lond. B 265:113–119.

Grafen, A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. Lond. B 326:119–157.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford Univ. Press, Oxford, U.K.

Harvey, P. H., and A. Rambaut. 2000. Comparative analyses for adaptive radiations. Philos. Trans. R. Soc. Lond. B 355: 1599–1605.

Harvey, P. H., R. M. May, and S. Nee. 1994. Phylogenies without fossils. Evolution 48:523–529.

Heard, S. B., and A. Ø. Mooers. 2000. Measuring the loss of evolutionary history from extinction: phylogenetically patterned speciation rates and extinction risks alter the calculus of biodiversity. Proc. R. Soc. Lond. B 267:613–620.

Jablonski, D. 1996. Body size and macroevolution. Pp. 256–289 *in* D. Jablonski, D. H. Erwin, and J. H. Lipps, eds. Evolutionary paleobiology. Univ. of Chicago Press, Chicago, IL.

Katzourakis, A., A. Purvis, S. Azmeh, G. Rotherow, and F. Gilbert. 2001. Macroevolution of hoverflies (Diptera: Syrphidae): the effect of the use of higher level taxa in studies of biodiversity and correlates of species richness. J. Evol. Biol. 14:219–227.

Kirkpatrick, M., and M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47: 1171–1181.

Kochmer, J. P., and R. H. Wagner. 1988. Why are there so many kinds of passerine birds? Because they are small. A reply to Raikow. Syst. Zool. 37:68–69.

Kubo, T., and Y. Iwasa. 1995. Inferring the rates of branching and extinction from molecular phylogenies. Evolution 49:694–704.

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

Maddison, W. P. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377.

Martin, R. A. 1992. Generic species richness and body mass in North American mammalian body mass and speciation rate. Hist. Biol. 6:73–90.

Martins, E. P. 1998. Adaptation and the comparative method. Trends. Ecol. Evol. 15:296–299.

Martins, E. P., and T. Garland. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534–557.

Marzluff, J. M., and K. P. Dial. 1991. Life-history correlates of taxonomic diversity. Ecology 72:428–439.

Mitter, C., B. Farrell, and B. Wiegmann. 1988. The phylogenetic study of adaptive zones: Has phytophagy promoted insect diversification? Am. Nat. 132:107–128.

Mooers, A. Ø., and S. B. Heard. 1997. Evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.

Nee, S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661–668.

Nee, S., A. Ø. Mooers, and P. H. Harvey. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl. Acad. Sci. USA 89:8322–8326.

Nee, S., T. G. Barraclough, and P. H. Harvey. 1996. Temporal changes in biodiversity: detecting patterns and identifying causes. Pp. 230–252 *in* K. J. Gaston, ed. Biodiversity: a biology of numbers and difference. Blackwell Scientific, Oxford, U.K.

Orme, C. D. L., D. L. J. Quicke, J. Cook, and A. Purvis. 2002. Body size does not predict species richness among the metazoan phyla. J. Evol. Biol. 15:235–247.

Owens, I. P. F., P. M. Bennett, and P. H. Harvey. 1999. Species richness among birds: body size, life history, sexual selection or ecology? Proc. R. Soc. Lond. B 266:933–939.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pagel, M. D., and P. H. Harvey. 1992. On solving the correct problem: wishing does not make it so. J. Theor. Biol. 156:425–430.

Purvis, A. 1996. Using interspecies phylogenies to test macroevolutionary hypotheses. Pp. 153–168 *in* P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith, and S. Nee, eds. New uses for new phylogenies. Oxford Univ. Press, Oxford, U.K.

Purvis, A., and A. J. Webster. 1999. Phylogenetically independent comparisons and primate phylogeny. Pp. 44–70 *in* P. C. Lee,

ed. Comparative primate socioecology. Cambridge Univ. Press, Cambridge, U.K.
Purvis, A., J. L. Gittleman, and H. K. Luh. 1994. Truth or consequences: effects of phylogenetic accuracy on two comparative methods. J. Theor. Biol. 167:293–300.
Purvis, A., S. Nee, and P. H. Harvey. 1995. Macroevolutionary inferences from primate phylogeny. Proc. R. Soc. Lond. B 260: 329–333.
Pybus, O. G., and P. H. Harvey. 2000. Testing macroevolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B 267:2267–2272.
Ricklefs, R. E., and S. S. Renner. 1994. Species richness within families of flowering plants. Evolution 48:1619–1636.
Schluter, D. 2000. The ecology of adaptive radiation. Oxford Univ. Press, Oxford, U.K.
Shao, K.-T., and R. R. Sokal. 1990. Tree balance. Syst. Zool. 39: 266–276.
Smith, A. B., and S. H. Jeffery. 1998. Selectivity of extinction among sea urchins at the end of the Cretaceous period. Nature 392:69–71.
Van Valen, L. 1973. Body size and numbers of plants and animals. Evolution 27:27–35.
Yule, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis F.R.S. Philos. Trans. R. Soc. Lond. A 213:21–87.
Zeh, D. W., J. A. Zeh, and R. L. Smith. 1989. Ovipositors, amnions, and eggshell architecture in the diversification of terrestrial arthropods. Q. Rev. Biol. 64:147–168.
Zink, R. M., and J. B. Slowinski. 1995. Evidence from molecular systematics for decreased avian diversification in the Pleistocene epoch. Proc. Natl. Acad. Sci. USA 92:5832–5835.

Corresponding Editor: J. Huelsenbeck

Appendix

Appendix Tables 1 and 2 present minimum adequate models (MAM) of the effects of simulation parameters on the error rates of ΔS. The error structure is binomial in both models with the scale parameter fixed at 1.0.

Appendix Table 1. Minimum adequate model of the validity results for relative rate difference (see Table 2). The statistical behavior of the most balanced trees was found to be statistically equivalent to Markovian. The model for the other robust index of ΔS, proportional dominance index, is qualitatively identical. Residual deviance = 17.41, df = 15. Model deviance = 122.29, df = 8.

Parameter	Estimate	SE	*Z*	*P*
Constant	−3.16	0.0857	−36.9	<0.0001
Equal vs. real branch lengths	0.603	0.0934	6.46	<0.0001
t-test vs. regression	0.0562	0.109	0.517	0.605
Unbalanced vs. other trees	0.0779	0.114	0.685	0.494
SB vs. BM	−0.114	0.0914	−1.25	0.21
2 × 3 (Blens = .*t*-test)	−0.874	0.121	−7.23	<0.0001
2 × 4 (Blens = .Unbalanced)	−0.504	0.125	−4.041	<0.0001
3 × 5 (*t*-test.SB)	0.423	0.120	3.52	0.0004
4 × 5 (Unbalanced.SB)	0.464	0.126	3.68	0.0002

Appendix Table 2. Minimum adequate model of the tests of power (see Table 3) for relative rate difference and proportional dominance index. Residual deviance = 1.072, df = 4. Model deviance = 155.0, df = 7.

Parameter	Estimate	SE	*Z*	*P*
Constant	−1.45	0.0571	−25.5	<0.0001
Equal vs. real branch lengths	0.0701	0.0798	0.878	0.38
t-test vs. regression	−0.470	0.0880	−5.34	<0.0001
2 × 3 (Blens = .*t*-test)	−0.569	0.133	−4.30	<0.0001