sampling intervals. In subsequent analyses, about 10,000 records at the genus level replaced the family data, and the time scale was refined to contain 51 sampling intervals (*4*).

The refinement of the time scale from 39 to 51 sampling intervals was not done arbitrarily. Rather, recognized substage boundaries were used to break up the longer sampling intervals of the more primitive Harland scale. The new generic data were placed in the substages as accurately as possible. That is, the refinement was not done merely by interpolation from the old time scale. Inevitably, the 51-interval scale includes a higher proportion of minor boundaries defined on criteria other than major extinction events. This should have the effect of diluting the 26-my signal seen in the original time scale (Fig. 1A). This prediction was tested by applying the Stigler and Wagner simulation procedure to the 51-interval scale, and the result is shown in Fig. 1B. The tendency for the time scale to produce a 26-my periodicity has disappeared. Thus, a 26-my signal can be seen in the coarse time scale because it contains a substantial number of boundaries defined by major extinctions. The finer time scale contains 36 of the original 39 boundaries, but the addition of 12 minor boundaries masks the periodic signal.

If our original finding of periodicity had been spurious or a statistical fluke, it is likely that increasing the data base by a factor of 20 would have destroyed or severely altered the signal. Instead, the periodic signal has been considerably strengthened, as shown in Fig. 2, which is based on 9773 generic records and the full 51-interval time scale. The last six events (150 my before present to the present) are clearly delineated and match the 26-my periodicity almost perfectly, although the radiometric dating of the sixth (Tithonian) is uncertain. Earlier events show a poorer fit, as is reasonable in view of the weaker biological and temporal control in the older record.

The case for periodicity in the extinction record is based on statistical inference with messy data, and thus it cannot be proved or disproved in a truly satisfactory manner. Because acceptance of periodicity (and some of its suggested causes) would entail a major shift in the way geologists look at the history of the earth and of life, it is proper that the hypothesis be evaluated as toughly as possible. To this end, the past three years have seen a number of published reinterpretations of the extinction data. Some of these have supported periodicity (*9*) and others, such as the Stigler and Wagner effort, have not (*10*). Some of the negative criticisms have been constructive and have led to important improvements in the testing procedures.

The question of periodicity will not be settled completely until we have new data independent of the extinction record. A number of laboratories are working intensively to provide independent tests, including broad sampling for evidence of climatic changes, meteorite impact, and other signals in environmental history that may corroborate periodicity. Not until these databases are fully developed will we know for sure whether extinction is periodic and, if so, whether the signal is simple or complex. In the meantime, the periodicity idea is a hypothesis being testing in the best tradition of science.

DAVID M. RAUP
J. J. SEPKOSKI, JR.
Department of Geophysical Sciences, University of Chicago, Chicago, IL 60637

REFERENCES AND NOTES

1. D. M. Raup and J. J. Sepkoski, Jr., *Proc. Natl. Acad. Sci. U.S.A.* **81**, 801 (1984).
2. J. J. Sepkoski, Jr., and D. M. Raup, in *Dynamics of Extinction*, D. K. Elliott, Ed. (Wiley, New York, 1986), pp. 3–36.
3. D. M. Raup and J. J. Sepkoski, Jr., *Science* **231**, 833 (1986).
4. J. J. Sepkoski, Jr., in *Global Bio-Events*, O. Walliser, Ed. (Springer, Berlin, 1986), pp. 47–61.
5. S. M. Stigler and M. J. Wagner, *Science* **238**, 940 (1987).
6. S. J. Gould [*Nat. Hist.*, **93**, 14 (August 1984)] expressed this point well: "Our geological time scale depends on those events of mass extinction since they set the boundaries of major divisions. My standard response to generations of student groans (at the imposed necessity of memorizing all those funny names from Cambrian to Pleistocene) reminds my charges that they are not learning capricious words for the arbitrary division of continuous time, but rather the dates of major events in the history of life."
7. A. G. Fischer, oral presentation, Princeton University Research Symposium, Princeton, NJ, 9 May 1985.
8. U. Bayer, *Geol. Rundschau* **76**, 485 (1987).
9. The following papers have presented statistical analyses supporting periodic extinction: J. A. Kitchell and D. Pena, *Science* **226**, 689 (1984), but see (*10*); M. R. Rampino and R. B. Stothers, *Nature* **308**, 709 (1984); *Science* **226**, 1427 (1984); R. A. Muller, in *The Search for Extraterrestrial Intelligence: Recent Developments*, M. D. Papagiannis, Ed. (Reidel, Dordrecht, 1985), pp. 233–243; E. F. Connor, in *Patterns and Processes in the History of Life*, D. Raup and D. Jablonski, Eds. (Springer, Berlin, 1986), pp. 119–147; N. L. Gilinsky, *Nature* **321**, 533 (1986); J. A. Kitchell and G. Estabrook, *ibid.*, p. 534 (1986); W. T. Fox, *Paleobiology* **13**, 257 (1987).
10. Although many negative comments on the question of extinction periodicity have been published since 1984, the following contributions presented statistical arguments against periodicity: A. Hoffman, *Nature* **315**, 659 (1985); A. Hoffman and J. Ghiold, *Geol. Mag.* **122**, 1 (1985); S. Tremaine, in *The Galaxy and the Solar System*, R. Smoluchowski, J. N. Bahcall, M. S. Matthews, Eds. (Univ. of Arizona, Tucson, AZ, 1986), pp. 409–416; A. Hoffman, *Nature* **321**, 535 (1986); E. Noma and A. L. Glass, *Geol. Mag.* **124**, 319 (1987); C. Patterson and A. B. Smith, *Nature* **330**, 248 (1987). In addition, the paper by J. A. Kitchell and D. Pena (cited in *9*) has been interpreted by some as critical of periodicity.
11. Supported by NASA grants NAG-2-237 and 2-282.

3 December 1987; accepted 23 March 1988

Response: Fours years ago, Raup and Sepkoski (*1*) created an immense stir with their detailed examination of the bold hypothesis that extinction rates were periodic with a period of 26 million years (my). A major component of that paper was a significance test they performed, decisively rejecting the alternative hypothesis that extinctions have occurred as a totally random process. In our report (*2*), we examined the statistical properties of the test they used (*3*); we replicated their analysis and confirmed their main result (that the recorded series of extinction rates was inconsistent with the hypothesis of a totally random process), but we discovered two things about the test that led us to conclude that the apparent periodicity could well be a statistical artifact. We remain convinced that our conclusion was correct.

In our examination of the significance test, we discovered that the Harland time scale (*4*) as used in the original paper by Raup and Sepkoski (*1*) exerted a peculiar bias toward a best fitting period of 26 my and that the test was as sensitive to measurement error of a type known to be present in the data as it was to truly periodic signals, given the noise levels expected with these data. We speculated that the two factors working together could well have produced an artifactual, statistically significant, "period" of 26 my in the original study. We noted that, even if the time scale were refined to the point of being equally spaced (with a stage duration equal to the average stage duration for the Harland scale), the second factor could produce an artifactual "period" in the range from 25 to 30 my—the strong preference for exactly 26 my would disappear, but the tendency of such models to produce artifactual periods would persist. Our results imply that no valid demonstration of periodicity is possible without allowing for this tendency.

In their comment, Raup and Sepkoski note [as they already had in (*1*)] that the boundaries of several strata are determined at least in part by the fossil record itself, and they suggest that the strong preference for 26 my we found in the Harland time scale might be a reflection of this connection, and indeed that it might therefore be taken as itself evidence of periodicity. We note first that these patterns in the time scale, whatever their nature (whether they are the result of a numerical quirk or a consequence of a true periodicity), are irrelevant to an important part of our analysis. The patterns in the time scale were not responsible for the statis-

tical significance that Raup and Sepkoski found and we corroborated. This is because the test was conditional on the time scale, calibrated to be an equally valid test of the null hypothesis for any time scale. That is, the calculation of the P values was based on the given time scale in such a way that the chance a totally random series would produce a statistically significant result is the same for any time scale. We refer here to the test they performed that did not randomly rearrange the time scale; we shall briefly comment later on the reasons we believe the test that randomized the time scale, while valid, does not test an interesting null hypothesis and thus does not address the scientific question at issue.

The time scale does affect the ability of the conditional test to detect alternative hypotheses; for the Harland scale the bias would make it difficult to detect any but a 26-my or 27-my period, if one existed, even though the sensitivity of the test to measurement error is little affected by the time scale. Indeed, the bias makes it likely that the test will interpret many appreciable departures from the null hypothesis (periodic or not) as approximately a 26-my period. But by itself, the pattern in the time scale will not tend to produce statistical significance with totally random series.

Raup and Sepkoski suggest that the preference of the Harland time scale for 26 my might itself be interpreted as evidence of periodicity in extinction rates. Unfortunately, this preference seems to be too fragile to be interpreted as support for that hypothesis. Indeed, to a surprising degree this preference for 26 my is due to what might be best termed a numerical quirk in the Harland scale that seems completely unrelated to any paleontological event.

Our study (*2*) followed (*1*) in concentrating on the period from 253 million years ago (Ma) to 5.1 Ma. More than half of the Harland scale over this period, namely the period from 238 Ma to 113 Ma is based on a simple linear interpolation of 19 stage boundaries to form 20 subintervals between these limits. However, there are two important senses in which the interpolation in the scale actually used was not simple. First, as we stated in our report, we followed (*1*) in omitting the boundary at 181 Ma and amalgamating the Bajocian and Aalenian into a single stage of duration 13 my. [The reason for this omission is not clear to us, but may have resulted from difficulties in dating extinctions of families in the 1982 Sepkoski Compendium that was the basis for (*1*); these difficulties have evidently since been overcome, because later studies by Raup and Sepkoski do not amalgamate these stages.] Second, the interpolation is not linear, but rather it is periodic, with a period of 25 my! If linear interpolation were employed, the interpolated stage duration would be 125/20 = 6.25 my. To provide round numbers for dates over this span, Harland *et al.* rounded to 6 my, but made every fourth duration 7 my in order to keep the total consistent. The sequence of durations over this span was thus the periodic sequence (in million years) 7, 6, 6, 6, 7, 6, 6, 6, 7, 6, 6, 6, 7, 6, 6, 6, 7, 6, 6, 6. The variation from uniformity is slight, but since it is extended over half the period of the study, its influence is appreciable. We find that these two changes (reintroducing a boundary at 181 Ma and using strict linear interpolation with all 20 of these stage durations equal to 6.25 my) account for about half the strong preference for 26 my we noticed, and if two other changes are made [namely the longest stage (duration, 15.5 my) is divided into two stages, and the shortest (duration, 1 my) is amalgamated with an adjacent stage], the strong preference we found essentially disappears. The effect of these changes is illustrated in Fig. 1, computed on the same basis as figure 7 of (*2*). These figures display the distance from a time scale to a perfectly regular grid; our Fig. 1A reflects the bias toward 26 my in the Harland scale that is at the bottom of the effect portrayed in their figure 1A [and in figure 4 of (*2*)]; our Fig. 1C shows how minor changes can produce near uniformity (like that reflected in their figure 1B); our Fig. 1B is intermediate.

The point is that the preference for 26 my is fragile, depending on a numerical quirk and the placement of two extreme stages. Indeed, the 51-interval time scale employed by Raup and Sepkoski in their recent study of generic level data employs only paleontologically recognized boundaries and shows no trace of periodicity.

Raup and Sepkoski also question whether our moving-average model adequately captures the stochastic structure of the errors in extinction series, and they quite correctly note that they had previously examined the effect that movement of two early mass extinctions would have upon the assessed significance in one application of the test (*6*), arguing that this confronted the measurement error problem. We discuss the adequacy of our model first, as that will permit us to explain why we believe that examination of alternative dates for two mass extinctions does not effectively confront the problem (although we were remiss in not addressing this question in the original report).

In order to properly address the issues involved here, it is important to keep in mind the fundamental logic of the significance test in question: a test statistic (we called the measure of fit "D") is computed for the data, and its value is compared with a

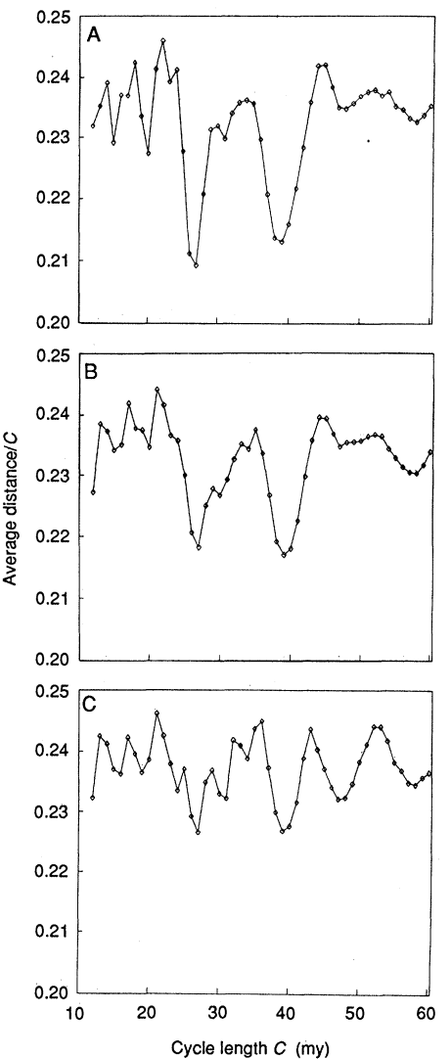

Fig. 1. This figure, computed on the same basis as figure 7 of (*2*), illustrates the fragility of the sharp bias toward 26 my in the original version of the Harland scale (**A**), by showing how it is reduced by about half by elimination of the periodic numerical quirk in the interpolation scheme in (**B**), and is largely eliminated by (**C**) changes of the boundaries of the two most extreme stages. The ordinates give the average distances from each of three time scales to the best fitting cycle C, for C = 12 to 60. For each of the time scales, the average distance from the n boundary points for that scale to the nearest point on the grid equally spaced by C (with phase chosen to minimize this average distance) was calculated; the vertical scale gives these averages divided by the cycle length C. The time scales are (A) the n = 40 point version of the Harland time scale over the period 253 Ma to 5.1 Ma, as used in (*1*); (B) the n = 41 point version of this scale, where the boundary at 181 Ma is reintroduced and the span from 238 Ma to 113 Ma is recalculated with strict linear interpolation to eliminate the periodic numerical quirk in (*1*) and (*4*); (C) the n = 41 point version of the same scale as (B), where the 15.5-my Albian stage is subdivided and the 1-my Coniacian stage is amalgamated with the adjacent 2.5-my Turonian stage.

reference distribution of values, values corresponding to a random sample of series of the same length from a model that describes a data-generating mechanism for a world where the hypothesis of periodicity does not hold. For the test to have credibility, the model must capture in at least a gross sense the characteristics of the data we would expect in a nonperiodic world (no straw men allowed), and the test must not be too sensitive to likely departures from that model that are not of interest (are not periodic). For example, a test that uses random rearrangements of the time scale to construct a reference distribution eliminates any time scale effect, but it has no obvious relevance to the geological world in question. It scrambles poorly estimated time spans with well-determined ones and confounds time scale effects, measurement errors that are serially dependent, species origination trends, and other factors related to time (for example, problems in resolving time of extinction to the stage level and time spans needed for ecosystem recovery after mass extinction).

We are not aware of any tests before our study of the periodicity of the extinction rate series that incorporated a stochastic component for measurement error (*5*). Our choice of a specific, moving average model was heuristic; it was not based on a detailed empirical investigation of dating errors in extinction times for either families or genera. Such an empirical investigation would be difficult, since it would require estimates of the survival frequency and probability of discovery of fossils of a large number of species at many geological levels and in many geographical locations. Nonetheless, we believe we can defend our choice of model as describing errors of a plausible magnitude. Our moving average model with parameter values $\theta = 0.5$ or 1.0 would describe errors of a magnitude appropriate for generic-level data if a sizable fraction (say 1/3 to 1/2) of generic-level extinctions were in error by one stage, which would correspond roughly to an average dating error for individual genera of at least a third to a half of an average stage, or about 2 to 3 my. For comparison, we note that Raup and Sepkoski have elsewhere described (*6, 7*) error bounds for dates of eight mass extinctions (which we would expect to be better dated than the extinctions of individual genera) as ranging from 1 my [for that at the Cretaceous-Tertiary (K-T) boundary] to 12 my (for that in the Rhaetian), averaging 6 my. Errors of this magnitude are sufficient to produce the effect we studied. That measurement errors of the moving average type can be as large or larger than our model specifies is also suggested from the prevalence of the phenomenon that Raup has given the evocative name "The Pull of the Recent"—the tendency for ancient taxa to be much better represented among currently extant species than past extinction rates would predict, a phenomenon that has been attributed in part to much better sampling among current species than in the fossil record (*8*). Indeed, our choice of model is certainly conservative in one respect, namely that it does not permit errors larger than a single stage.

We now turn to the relation of the dating of individual mass extinctions to our model for measurement error. We do not in the least dispute the statement by Raup and Sepkoski in their comment that the dates of several of the mass extinctions, such as that at the K-T boundary, are quite well determined. But it is the effect of measurement errors on the whole series that influences the test. The effect is that of a smoothing filter, one which reduces the propensity to peak or trough, and tends to space out peaks or troughs more evenly than would be the case for random series. The test we examined looked at all peaks, both major and minor, and the major impact of measurement error would be expected to be on the placement and size of minor peaks. Actual minor peaks may have been effaced or moved by this smoothing, particularly when they occur (as may be inevitable) in less studied strata. Indeed, the pronounced effect of our incorporation of measurement error in our model is to restrict the reference distribution for the statistic D to series that are smoother and more regular than totally random series. We grant their specialist knowledge of accurate dating of several extinction events, but it would require an extension of that knowledge to most of the entire series to have a large effect on our conclusions. To experiment with the dating of two uncertain, early major peaks, as Raup and Sepkoski did in (*6*), is to move toward the construction of a reference distribution that incorporates special knowledge about measurement errors (although it does not get to the important issue of the minor peaks), and even there the statistical significance weakens. We do not know how one might effectively incorporate their expert specialist knowledge on the different sizes of measurement errors in different strata into a significance test.

In their comment, Raup and Sepkoski argue that, if the apparent periodicity had been "spurious or a statistical fluke, it is likely that increasing the data base by a factor of 20 would have destroyed or severely altered the signal." To the contrary, as we noted in our report (*2*), it is likely that such an increase in the data base would enhance the effect, particularly if the augmented data base were more susceptible to measurement error than the original. Thus our model would predict that as the generic data base is enlarged, with the addition of less accurately resolved genera (more likely to be subject to the Signor-Lipps effect), the pseudoperiodicity would become more pronounced. The extinction series, anchored at a few well-determined mass extinctions (such as that at the K-T boundary) would be further smoothed by the new data, and the moving average effect, paradoxically, would become stronger. It is correct that a "statistical fluke" would not be expected to survive augmentation of the data, but the same would not hold for a persistent bias such as that we discussed.

On top of the Signor-Lipps effect, there is also the separate problem of dating fossils that are observed. For the full generic-level data set, no more than 67% of the extinctions have dates resolved to the level of the stage; the remainder are known only to lower resolution, and the data set was constructed by proportionally allocating them among possible stages (*9*). Yet while this proportional allocation is a quite sensible step from many points of view, it also increases the moving average component in the measurement error.

Other factors have been noted which, while we did not incorporate them in our model, would qualitatively have the same effect as our moving average model, namely a tendency to separate the peaks more regularly than would be the case for purely random series. These include a tendency for the ecosystem to require a recovery time after a mass extinction (*10*) and a possible tendency for taxonomists to group together species as in the same genera due to proximity of extinction time (*11*). In addition, if the analysis only treats as peaks those stages where the rate is judged to be statistically significantly higher than the neighboring minima, as was done in (*6*) and (*7*), this too will tend to separate the peaks more regularly and bias the test.

We emphasize that the issue we addressed was that of assessing the statistical significance of a visually appealing pattern. If a periodic appearance is advanced as a purely descriptive way of summarizing the pattern of extinction rates over the past 150 my in figure 2 of Raup and Sepkoski's comment, then no one with that figure in view could quarrel (just as no one, we feel, would see a visually appealing pattern over the first half of the same figure.) But the hypothesis of a periodic dynamic structure is so powerful in its implications, and so seductive in the ease with which it imposes itself on us with limited data sets such as this one, that it must be required to pass a stringent test. We

have *not* shown that extinction rates are not periodic. We *have* shown that periodicity cannot be validly demonstrated without incorporating a model that allows for measurement errors that have moving-average behavior. This result, that certain types of measurement error can enhance a periodic signal or cause a pseudoperiodic signal to emerge from nonperiodic data, is counterintuitive. As Raup and Sepkoski have elsewhere written: "Inaccurate geologic dates or nonexistent extinction events will degrade the sample in a direction toward randomness and away from any regular signal. Thus, to include uncertain data is to make statistical testing more conservative. To argue that uncertainty in the data explains the observed periodicity is illogical" (*6*). While this is undoubtedly true for many stochastic error structures, what we have found is the surprising result that it does not apply to reasonable models for exactly the type of measurement error that they and other paleontologists have long recognized as prevalent in such data.

STEPHEN M. STIGLER
MELISSA J. WAGNER
Department of Statistics, University of Chicago, Chicago, ILL 60637

REFERENCES AND NOTES

1. D. M. Raup and J. J. Sepkoski, Jr., *Proc. Natl. Acad. Sci. U.S.A.* **81**, 801 (1984).
2. S. M. Stigler and M. J. Wagner, *Science* **238**, 940 (1987).
3. We have since learned that essentially the same test was introduced in another context in 1956 by S. R. Broadbent [*Biometrika* **43**, 32 (1956)].
4. W. B. Harland *et al.*, *A Geologic Time Scale* (Cambridge Earth Sciences Series, Cambridge Univ. Press, Cambridge, England, 1982).
5. J. A. Kitchell and D. Pena [*Science* **226**, 689 (1984)] model the series with a time series model that permits serial correlation and indeed they find reasonable fits using nonperiodic models. They mention (their note 16) that a moving average model provided a good fit if the initial observation is omitted, but they do not tie that model in with a measurement error hypothesis. A recent study by W. T. Fox [*Paleobiology* **13**, 257 (1987)] using Fourier analysis makes no allowance for measurement error in posing a null hypothesis, errs in performing a significance test that is inappropriate for heavily linearly interpolated data, and does not allow for testing multiple hypotheses.
6. D. M. Raup and J. J. Sepkoski, Jr., *Science* **231**, 833 (1986).
7. J. J. Sepkoski, Jr., and D. M. Raup, in *Dynamics of Extinction*, D. K. Elliott, Ed. (Wiley, New York, 1986), chap. 1.
8. D. M. Raup, *Paleobiology* **4**, 1 (1978).
9. J. J. Sepkoski, Jr., *Lect. Notes Earth Sci.* **8**, 47 (1986).
10. S. M. Stanley, *Extinction* (Scientific American Library, New York, 1987), chap. 10.
11. C. Patterson and A. B. Smith, *Nature* **330**, 248 (1987).
12. Supported in part by National Science Foundation grant DMS-8601732. This manuscript was prepared with the use of computer facilities supported in part by NSF grants DMS-8601732 and DMS-8404941 to the Department of Statistics at the University of Chicago.

11 January 1988; accepted 6 June 1988