Transactions of the American Fisheries Society 124:131–138, 1995
American Fisheries Society

Graphical and Statistical Methods for Determining the Consistency of Age Determinations

STEVEN E. CAMPANA, M. CHRISTINA ANNAND, AND JAMES I. MCMILLAN

Marine Fish Division, Bedford Institute of Oceanography
Post Office Box 1006, Dartmouth, Nova Scotia B2Y4A2, Canada

Abstract.—Many laboratories rely on periodic rereading of reference collections of scales or otoliths to ensure that their age readers remain consistent in their age interpretations, both through time and with other age readers. Measures of both systematic difference (bias) and precision are required for this purpose, because measures of bias are not suitable as measures of precision, and vice versa. Using data from an age comparison study of haddock *Melanogrammus aeglefinus* for demonstration purposes, we evaluated a variety of graphical and statistical approaches for making paired age comparisons from the standpoint of both detecting age differences and assessing precision. Parametric and nonparametric matched-pair tests, regression analysis, analysis of variance, and age difference plots were all capable of detecting systematic over- or underaging. However, only the age bias plot was sensitive to both linear and nonlinear biases. The coefficient of variation (CV = SD/mean) was a robust measure of precision, whereas the widely used percent agreement statistic was not. The combination of an age bias plot, an age frequency table, and a CV provides a powerful and easily prepared comparison of matched pairs of age determinations.

For the forseeable future, the process of determining the ages of fishes will retain an element of subjectivity that will contribute various degrees of error to all age determinations, whether at the annual or the daily level. This subjectivity often originates with the interpretation of periodic features in otoliths, scales, and other calcified structures, which tend to vary markedly in appearance and relative size among fish (Campana and Neilson 1985; Casselman 1987). Yet laboratories with ongoing responsibilities for age determination must be consistent in their age interpretations. An age determination method should first be tested for accuracy by an accepted age validation technique before it is adopted (Beamish and McFarlane 1983; Geffen 1992). However, age validation studies are generally too costly and time-consuming for routine use in monitoring consistency. Rather, many laboratories rely on reference collections of age determination structures, random subsamples of which are inserted periodically into regular samples to ensure that age readers remain consistent in their age interpretations. Repeated age interpretations of a given sample offer a better statistical ability to detect deviations through time than readings of independent samples.

Aging error can be of two forms: error that affects accuracy, or the proximity of the age estimate to the true value, and error that affects precision, or the reproducibility of individual measurements on a given structure (Wilson et al. 1987). The two forms of error are not necessarily linked. For example, consistent underaging of a sample by 1 year can yield the same measure of precision as a sample that is, on average, aged accurately. An ideal age reference collection contains structures of known age, thus allowing tests of both accuracy and precision. In practice, known-age samples of wild fish are seldom available. However, aging consistency can still be monitored by ensuring (1) that the age interpretations of individual age readers do not "drift" through time, introducing bias relative to earlier determinations; (2) that the age interpretations by different readers are comparable; and (3) that the precision of age interpretations by individual readers does not deteriorate through time. Items (1) and (2) actually refer to measures of bias relative to some reference value, rather than to accuracy. However, in the absence of a known-age reference collection, aging consistency is the best that can be achieved.

Age determinations are routinely compared among readers and for the same reader over time, but many published comparisons are of questionable value, perhaps because well-defined guidelines for such comparisons are lacking. The objective of this paper is to contrast the value of several graphical and statistical approaches for making paired age comparisons of a given sample, from the standpoint of both detecting differences in age estimates and assessing precision. The emphasis is on paired comparisons, although all of the approaches described here are adaptable to multiple comparisons. We demonstrate the techniques with data from our own laboratory, and we conclude by recommending a few easily implemented approaches useful in comparing multiple age determinations for any fish sample.

Data

Otolith-based age comparisons for adult haddock *Melanogrammus aeglefinus* from the Scotian Shelf (1985 and 1992 samples from North Atlantic Fisheries Organization Division 4X) were drawn from a series conducted as part of the training of five new age readers (1, 2, 3, A, and B). All agers had received training prior to the exercise, and all interpreted their prepared otolith cross sections in isolation from the remaining agers. Experience levels and the type of microscope or image analysis system varied among age readers, but such differences are irrelevant in the context of this analysis. None of the age interpretations were validated. However, through comparison with experienced age readers from other laboratories, we believe that agers 1 and 3 provided unbiased age estimates. The data for agers A and B were modified slightly in order to illustrate a point.

Detection of Bias

Age frequency tables for each of three pairwise age comparisons are presented in Table 1. Such tables form a central component of many paired age comparisons, but they are not particularly well suited to the detection of age differences between readers. Close inspection of Table 1 suggests some underaging of older fish by ager 2 relative to the readings of ager 1, but the relationship otherwise appears linear. No bias is obvious in the other pairwise comparisons. It is evident that a graphical scatterplot of one age reader versus another would be difficult to interpret in all but the most extreme of age differences; coincident points would not be evident, and viewers would have to conduct the equivalent of a weighted regression in their heads.

Age difference plots, in which the difference between the two age readers is plotted as a function of one of the sets of ages, have often been used for comparing pairs of age determinations (Figure 1). These plots highlight major systematic differences between two sets of age readings, as indicated by a distribution of points centered on something other than the zero line, as in the top panel of Figure 1. More subtle discrepancies are difficult to discern; if systematic differences exist in the middle and bottom panels of Figure 1, they are not immediately obvious. Some of this difficulty arises because coincident points are not shown; a data point could represent either 1 or 100 age readings. In principle, color-coded or three-dimensional plots could be used to overcome this constraint. With our data, however, such plots did not necessarily show any additional bias. An age difference plot, which shows deviations from the line ager X = ager Y, is not equivalent to the residuals from a linear regression, which shows deviations from any straight line that may fit the data. For example, a relationship in which ager X counted two annuli for each one counted by ager Y would not show up in a plot of the residuals from a regression.

Both parametric and nonparametric statistical tests have been proposed for comparing ages, and all have advantages and disadvantages. Simple linear regression, the most frequently applied of the parametric tests, is normally interpreted in terms of significant differences from a slope of 1 and an intercept of 0. A slope other than 1 suggests inconsistency in the interpretation of annuli by one of the agers; an intercept other than 0 suggests a systematic difference between the two agers, perhaps due to different interpretations of the first annulus. Both these conditions were evident in the regression of ager 2 on ager 1, summarized in Table 2, which supports the conclusion of bias derived from Figure 1. In contrast, the regression results did not suggest the presence of bias between agers 1 and 3 or between agers A and B (Table 2).

Two standard tests for systematic differences between matched pairs of ages are the parametric paired *t*-test and nonparametric Wilcoxon matched-pairs rank test (Conover 1980). These tests are as effective as regression at detecting systematic differences between pairs of agers, as indicated by the significant differences between agers 1 and 2 (Table 2). The Wilcoxon test also is insensitive to assumptions of normality and homoscedasticity, unlike the *t*-test and linear regression. However, neither of these matched-pair tests is well suited for detecting situations in which one ager underages at one end of the age range and overages at the other end. Such a situation is more readily detected through estimation of a regression slope. Neither regression analysis nor either matched-pair test is satisfactory when the relationship between two agers is nonlinear but centered on the 1:1 line (see the comparisons of ager A with ager B in Table 2).

To overcome the constraints imposed by age difference plots and standard statistical tests, we propose the use of the age bias plot (Figure 2). This type of graph plots one age reading versus another, and can be interpreted through reference to the equivalence line ager X = ager Y. However, in this representation, the age readings of ager Y are pre-

TABLE 1.—Age frequency tables summarizing pairwise comparisons of age estimates from three different otolith samples. Data are numbers of fish.

Age (years) estimated by:	Age (years) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
Estimated by ager 2																
Ager 1																
1			1													1
2		12	29	5	1											47
3			59	15	1											75
4			4	48	17	3	1									73
5				35	98	5	1									139
6				2	35	16	3	1								57
7					4	18	7									29
8					2	20	13	1	1							37
9				1	1	3	10	8	5							28
10						1	3	4	1	2						11
11						3	2	2		1						8
12									2	1						3
13										1						1
Estimated by ager 3																
Ager 1																
2		2	2													4
3		1	4	2												7
4			3	55	10	1	1									70
5				16	52	10	1	1								80
6					6	5	2									13
7						9	5	5	1							20
8			1			1	5	8	1	4						20
9							1	3	5	5	1					15
10					1				1	7	2					11
11										1	3	1	1			6
12										1	4	5		1		11
13												1	1		1	3
14													1			1
Estimated by ager B																
Ager A																
1	1															1
2	29		5		1											35
3	59		15		1											75
4		4		48		17		3								73
5				35		98		5	1							139
6				2		32		15	3	1						53
7						4		18	7							29
8					2		20	13	1	1						37
9					4		3	10	8	5	1	1				32
10							1	3	4	1	2					11
11							4	1	1		1	1	1			9
12										4	1		1			6

sented as the mean age and 95% confidence interval corresponding to each of the age categories reported by ager X. The intent of the confidence intervals is not to assign statistical significance to the comparison, but to allow informed interpretation of any difference between the observed line and the equivalence line. Either parallel but separated lines or increasing divergence as the lower or upper age range is approached indicates systematic differences between the two age readers. As was the case with the age difference plot, the selection of ager for the abscissa is arbitrary. In the example of Figure 2, there is very obvious bias between ager 1 and ager 2 but none between ager 1 and ager 3. For the first time, however, there is a clear indication of systematic error between ager A and ager B. This bias was not obvious in the other graphical or statistical tests presented, largely because the relationship between the two agers was nonlinear. Thus the age bias plot was the only means of age comparison tested that allowed clear visual detection of the systematic age differences

FIGURE 1.—Differences (years) in age estimates between paired age readers for each of the three pairwise age comparisons in Table 1. Each point represents one or more observations of an individual fish. **(Top)** Age estimation bias present. **(Middle)** No age estimation bias present. **(Bottom)** Nonlinear age estimation bias present.

TABLE 2.—Statistical tests for comparing pairs of age determinations, applied to the data of Table 1. The slope and intercept of simple linear regressions are tested for significant differences ($\alpha = 0.05$) from 1.0 and 0, respectively. The parametric paired *t*-test and the nonparametric Wilcoxon matched-pairs rank test are used to detect significant differences from a paired difference of 0. Error terms are 95% confidence limits.

Statistic	Ager 2 versus ager 1 ($N = 509$)	Ager 3 versus ager 1 ($N = 261$)	Ager A versus ager B ($N = 508$)
	Regression		
Slope	0.62 ± 0.03	0.96 ± 0.04	1.01 ± 0.05
P	0.000	0.12	0.83
Intercept	1.66 ± 0.16	0.25 ± 0.30	−0.18 ± 0.33
P	0.006	0.27	0.51
	Paired *t*-test		
Mean paired difference	0.34 ± 0.10	0.00 ± 0.11	0.09 ± 0.13
P	0.000	0.95	0.13
	Wilcoxon test		
Positive ranks	84	53	200
Negative ranks	177	56	189
Ties	248	152	119
P	0.000	0.67	0.07

between agers A and B. Although statistical tests exist for quantifying discrepancies between empirical and equivalence lines, we suggest that visual inspection is more appropriate; when both age readers are very precise, statistically significant differences may arise that have no practical significance (e.g., a bias of 0.1 year between readers). In addition, nonlinear bias patterns are most efficiently detected with the eye.

Estimates of Precision

In the absence of demonstrable bias, a measure of precision is a valuable means of assessing the relative ease of determining the age of a particular structure, of assessing the reproducibility of an individual's age determinations, or of comparing the skill level of one ager with that of others. The traditional index of precision in aging studies, percent agreement, is gradually falling out of favor, because percent agreement may vary substantially among species and among ages within a species. Beamish and Fournier (1981) illustrated this point by noting that 95% agreement of two readers to within 1 year can be poor precision for aging Pacific cod *Gadus macrocephalus*, given the few year-classes in the fishery, but 95% agreement to within 5 years can be good precision for spiny dogfish *Squalus acanthias*, given this species' 60-

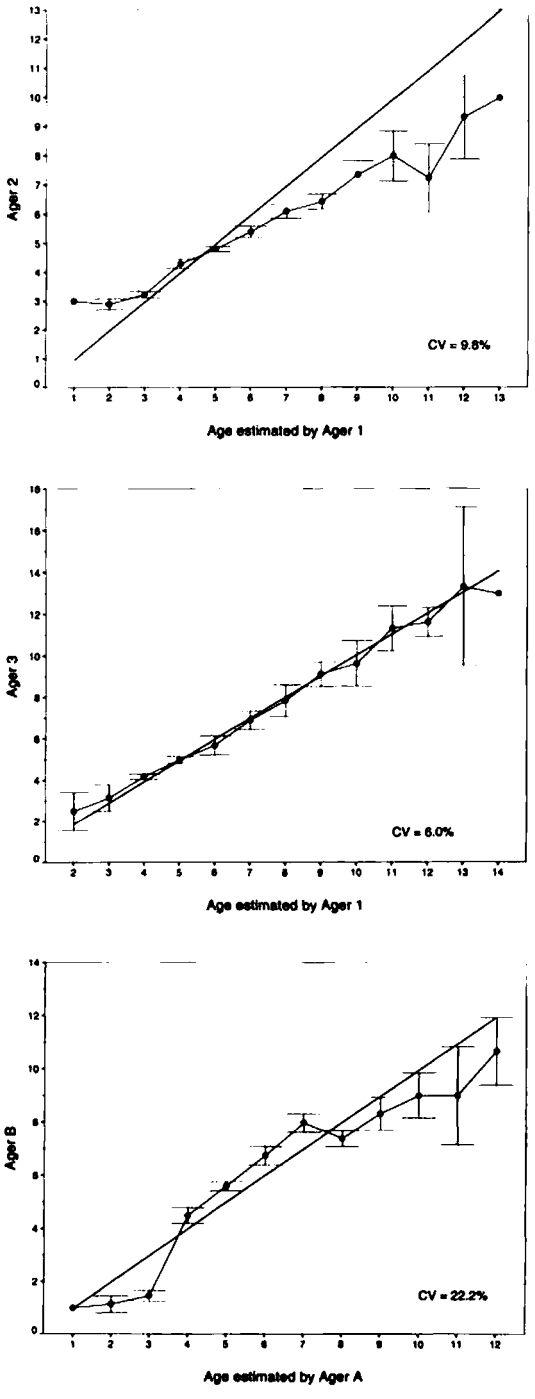

FIGURE 2.—Age bias graphs for each of the three pairwise age comparisons presented in Table 1. Each error bar represents the 95% confidence interval about the mean age assigned by one ager for all fish assigned a given age by a second ager. The 1:1 equivalence (solid line) is also indicated. **(Top)** Age estimation bias present. **(Middle)** No age estimation bias present. **(Bottom)** Nonlinear age estimation bias present.

TABLE 3.—Measures of precision for comparing pairs of age determinations.

Statistic or index	Ager 2 versus ager 1 (N = 509)	Ager 3 versus ager 1 (N = 261)	Ager A versus ager B (N = 508)
Correlation coefficient (r)	0.88	0.94	0.85
Coefficient of variation (%)[a]	9.82	6.00	22.2
Average percent error[b]	6.94	4.24	15.7
Percent agreement	48.7	58.2	23.4

[a] From Chang (1982).
[b] From Beamish and Fournier (1981).

year longevity. Beamish and Fournier (1981) recommended the use of average percent error (APE), defined as

$$\mathrm{APE}_j = 100 \times \frac{1}{R} \sum_{i=1}^{R} \frac{|X_{ij} - X_j|}{X_j}, \qquad (1)$$

where X_{ij} is the ith age determination of the jth fish, X_j is the mean age of the jth fish, and R is the number of times each fish is aged. When APE_j is averaged across many fish, it becomes an index of average percent error. Chang (1982) agreed that APE is a substantial improvement over percent agreement but suggested that the standard deviation be used in equation (1) rather than the absolute deviation from the mean age. The resulting equation produces an estimate of the coefficient of variation (CV), expressed as the ratio of the standard deviation to the mean, and can be written as

$$\mathrm{CV}_j = 100 \times \frac{\sqrt{\sum_{i=1}^{R} \frac{(X_{ij} - X_j)^2}{R - 1}}}{X_j}. \qquad (2)$$

Equation (2) is the CV of the age estimate for a single fish (jth fish). As with equation (1), it can be averaged across fish to produce a mean CV. Equations (1) and (2) produce similar values (Chang 1982); however, the latter is statistically more rigorous and thus is more flexible. The same applies to the index of precision D (Chang 1982), which is similar to the CV (and identical to APE when $R = 2$), but is calculated as

$$D_j = \frac{\mathrm{CV}_j}{\sqrt{R}}. \qquad (3)$$

All of the above-mentioned measures of precision (more properly, of imprecision) are artificially inflated by any bias that exists between agers. For example, agers 1 and 2 had smaller errors associated with any age than did agers 1 and 3 (Figure 2), yet larger CV and APE (Table 3). In the absence

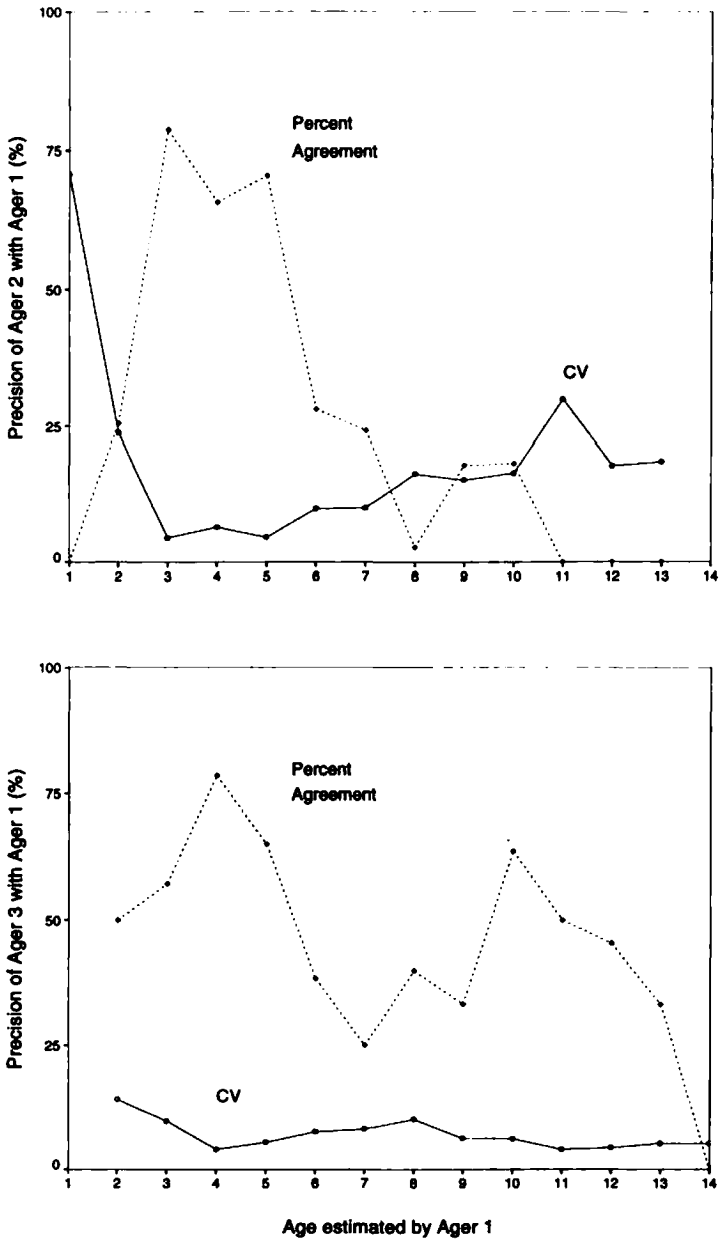

FIGURE 3.—Coefficient of variation (CV) and percent agreement for two of the pairwise age comparisons presented in Table 1. **(Top)** Bias present between the two age readers. **(Bottom)** Bias absent between the two age readers.

of bias, CV and APE were most sensitive and the correlation coefficient and percent agreement were least sensitive in documenting precision differences among agers. In addition, percent agreement varied substantially across ages, while CV did not (Figure 3). The instability of percent agreement across ages is largely responsible for the large artifactual variations in percent agreement that can occur when the precision of two samples of different age composition is compared.

Discussion

Repeated age determinations of a sample of fish are generally conducted for one of two reasons: to determine if there are systematic differences in age estimates between one or more age readers, meth-

odologies, or laboratories; or to estimate the precision (reproducibility) of age estimates. Of the two, it is the precision estimates that are most often reported, probably because the analysts implicitly assumed that no systematic differences existed. Yet the potential presence of systematic differences (bias) in interpreting age structures poses the most serious problem for those responsible for conducting age determinations. The presence of bias, in turn, will confound the interpretation of most measures of precision. Therefore, we suggest that bias should always be addressed before precision.

Several specialized tests have been proposed for use in determining if two or more sets of ages differ, but the circumstances for their use is somewhat limited. Hayes (1993) presented an approach based on the Fisher exact test for testing differences between age–length keys. Because it was designed for comparing the lengths at age of different samples of fish, it could be applied to the comparison of replicated age determinations, but it lacks the statistical power of a matched-pair test. It also requires a comparison of the age composition of each length interval, necessitating multiple comparisons and subjective criteria for evaluating the results. Bowker (1948) presented a test for symmetry in contingency tables that seems to be applicable to pairwise age comparisons. However, it has seen little or no use since it was published. Age comparisons have even been based on published growth curves. However, such comparisons are not ideal because they are based on random samples often drawn at different times of the year. Comparisons based on matched pairs (whereby the same structure is interpreted by each age reader) always provides the most statistical power. This report has focused on the analysis of paired comparisons, which are only a subset of the age comparisons possible. However, because most comparisons will be made on a reference set of otoliths or scales or in reference to another reader's estimates, we feel that pairwise comparisons will most often be the ones made. In any event, both the age bias plot and the CV can be used for multiple comparisons.

When systematic under- or overaging occurs, several linear models can be used to detect the problem. Approaches based on analysis of variance have successfully detected systematic differences between preparation techniques, laboratories, and age readers for analyses of both daily and annual ages (Boehlert and Yoklavich 1984; Baker and Timmons 1991; Campana and Moksness 1991). Indeed, when the differences are systematic and constant across ages (such as when one ager consistently misses the first annulus), most of the available techniques should detect the difference; age difference plots, paired *t*-tests, and Wilcoxon matched-pairs tests should be particularly good for this purpose. However, when bias varies across ages, such as when young to middle ages are aged without bias but older fish are underaged, these tests become much less sensitive. This is particularly true if fish at one end of the age range are overaged and fish at the other end of the age range are underaged. Under such a scenario, analyses of variance, paired *t*-tests, and Wilcoxon matched-pairs tests are all likely to fail. The age bias plot, on the other hand, effectively depicts both forms of bias. Use of the age bias plot also allows estimation of the magnitude of any bias that may exist, including (if precision is good) biases small enough to be safely ignored.

Determination of bias may be somewhat problematic, but estimation of precision is not. Measures of precision are relative values only; no one value can be considered an "acceptable level" for all species. However, several authors have clearly documented the dangers of the percent agreement statistic and the superiority of both APE and CV as measures of precision (Beamish and Fournier 1981; Chang 1982; Kimura and Lyons 1991). Despite these reports, roughly 35% of 21 randomly sampled age comparison papers published since 1985 report only percent agreement. Use of this statistic is difficult to understand and, given the results reported here and elsewhere, difficult to justify.

To conclude, bias and precision can be assessed by a variety of statistical and graphical techniques. Several parametric and nonparametric tests can detect consistent over- and underaging, but no one test seems able to detect both this and age-related shifts in aging criteria. For this reason, we recommend use of easily prepared and interpreted age bias plots for diagnosing systematic differences between two sets of age determinations. A combination of a CV to measure precision, an age frequency table (Table 1) to document the matched observations, and an age bias plot provides all the information required for assessing the consistency of repeated age determinations.

Acknowledgments

We thank all of the participants in the age reading exercise: D. Beanlands, P. Comeau, J. Hamel, J. Hunt, B. MacEachern, C. Nelson, P. Perley, J. Simon, S. Wilson, and G. Young. We also thank

P. Comeau for helpful comments on the analysis and an anonymous reviewer for helpful comments on the manuscript.

References

Baker, T. T., and L. S. Timmons. 1991. Precision of ages estimated from five bony structures of Arctic char (*Salvelinus alpinus*) from the Wood River System, Alaska. Canadian Journal of Fisheries and Aquatic Sciences 48:1007–1014.

Beamish, R. J., and D. A. Fournier. 1981. A method for comparing the precision of a set of age determinations. Canadian Journal of Fisheries and Aquatic Sciences 38:982–983.

Beamish, R. J., and G. A. McFarlane. 1983. The forgotten requirement for age validation in fisheries biology. Transactions of the American Fisheries Society 112:735–743.

Boehlert, G. W., and M. M. Yoklavich. 1984. Variability in age estimates in *Sebastes* as a function of methodology, different readers, and different laboratories. California Fish and Game 70:210–224.

Bowker, A. H. 1948. A test for symmetry in contingency tables. JASA (Journal of the American Statistical Association) 43:572–574.

Campana, S. E., and E. Moksness. 1991. Accuracy and precision of age and hatch date estimates from otolith microstructure examination. ICES (International Council for the Exploration of the Sea) Journal of Marine Science 48:303–316.

Campana, S. E., and J. D. Neilson. 1985. Microstructure of fish otoliths. Canadian Journal of Fisheries and Aquatic Sciences 42:1014–1032.

Casselman, J. M. 1987. Determination of age and growth. Pages 209–242 *in* A. H. Weatherley and H. S. Gill, editors. The biology of fish growth. Academic Press, New York.

Chang, W. Y. B. 1982. A statistical method for evaluating the reproducibility of age determination. Canadian Journal of Fisheries and Aquatic Sciences 39:1208–1210.

Conover, W. J. 1980. Practical nonparametric statistics. Wiley, Toronto.

Geffen, A. J. 1992. Validation of otolith increment deposition rate. Canadian Special Publication of Fisheries and Aquatic Sciences 117:101–113.

Hayes, D. B. 1993. A statistical method for evaluating differences between age–length keys with application to Georges Bank haddock, *Melanogrammus aeglefinus*. U.S. National Marine Fisheries Service Fishery Bulletin 91:550–557.

Kimura, D. K., and J. J. Lyons. 1991. Between-reader bias and variability in the age-determination process. U.S. National Marine Fisheries Service Fishery Bulletin 89:53–60.

Wilson, C. A., Chairman, and Glossary Committee. 1987. Glossary. Pages 527–530 *in* R. C. Summerfelt and G. E. Hall, editors. Age and growth of fish. Iowa State University Press, Ames.

Received February 28, 1994
Accepted July 23, 1994