RICHARD S. MCBRIDE, KRISTIN L. MAKI, and JANAKA DE SILVA
Florida Fish & Wildlife Conservation Commission
Fish & Wildlife Research Institute
100 Eighth Avenue SE
St. Petersburg, Florida 33701-5095 USA

ABSTRACT

When ages of fish are estimated via examination of their hardparts (i.e., otoliths, scales, rays, etc.), the precision (i.e., reproducibility) of those age estimates can be measured in several ways. Percent precision in our study represents the percent of replicated age estimates (i.e., for the same fish) that agree exactly or within some appropriately narrow age range (e.g., ± 1 year). The average percent error and coefficient of variation are slightly different formulas designed to express the uncertainty of the average estimated age. One (two, or even all three) of these measures can be calculated for many fish to develop an index (or indices) of precision, which is used to evaluate the consistency for which ages have been estimated within a sample of fish. The correlation coefficient can also be used to measure the association between replicate ages for a sample of fish. All four measures may be used to evaluate use of a particular hard part or preparation technique to age fish or to compare the 'ease' of ageing one species versus another. The use of these measures is problematic for a variety of reasons, as has been shown for a number of "real" data sets from fishery labs. We used a simulated data set with different levels of variance to evaluate the utility of these four measures of precision. Using simulated data with known patterns of precision and bias that represent a number of anticipated scenarios has been missing in the discussion of the relative efficacy of these different measures of precision.

KEY WORDS: Age and growth, ageing error, precision

Las Lecciones Aprendidas de la Medición de Edad Precisas para Poblaciones Simuladas de Pescados

La precisión de estimaciones de edad para un solo pez, cuando estimado con diferentes estructuras biológicas (como otoliths, escalas, rayos de aleta, etc.), puede ser medido en maneras diferentes. El porcentaje de acuerdo es el por ciento de las estimaciones de edad replicadas que concuerdan exactamente o dentro de algún estrecha apropiadamente la gama de la edad (por ejemplo, ± 1 año). El promedio por ciento error y coeficiente de la variación son fórmulas diferentes diseñado para ajustar una variación por el promedio edad evaluada. Uno (dos o todos los tres) de estas medidas es promediado a través de muchos pescados para desarrollar un índice (o los índices) de la precisión, que se utiliza para evaluar la consistencia de edades estimadas dentro de una muestra de

pescados. El coeficiente de la correlación se puede utilizar también para medir la asociación entre las edades replicadas para una muestra de pescados. Las cuatro medidas se pueden utilizar para evaluar una estructura biológica particular o diferentes técnicas de preparación para comparar 'la comodidad' de envejecimiento de una especie contra otro. El uso de estas medidas es problemático para una variedad de razones, como ha sido mostrado para varios conjuntos de datos de laboratorios de pesquería. Usamos un iueso de datos simulado con diferentes nievales de variación para evalúar completamente la utilidad de estas cuatro medidas de la precisión. Este enfoque – utilizando los datos simulados con pautas conocidas de la precisión y la tendencia que representan varios guiones anticipados – no ha tenido la discusión de la eficacia relativa de estas medidas diferentes de la precision.

PALABRAS CLAVES**:** La precisión de estimaciones, pescados

INTRODUCTION

Ageing of fish is a remarkably routine enterprise today. The most common method of ageing fish involves examining biological hardparts, particularly scales and otoliths, for bands that are laid down at regular time intervals (Campana and Thorrold 2001, Campana 2001). Confidence in age estimates determined via these methods is built by validating each method and using regular quality-control procedures to check for process errors and observation errors. Process error occurs when the selected hardpart contains an incomplete banding pattern; observation error occurs when the banding pattern cannot be unambiguously interpreted with a particular processing method. Process error is more likely to affect accuracy (i.e., difference between the estimated age and the known age); observation error is more likely to affect precision (i.e., agreement of replicate age estimates for the same fish) but can affect accuracy as well. Consideration of these errors is important because management and conservation policy decisions are increasingly formed based on age-structured data analyses of fish populations.

Routine quality-control procedures in production-ageing programs typically focus on measures of precision (e.g., Kimura and Lyons 1991) rather than measures of accuracy. The emphasis on precision has been criticized by Campana (2001) in cases where potential process errors have not been evaluated directly. In this paper we will be focusing on measures of precision, but we begin by stating our agreement with Campana (2001; p. 221) that 'precision cannot be used as a proxy for accuracy.' Accuracy issues must be dealt with first; otherwise, accounting for precision may not improve data quality. In routine production ageing, where there is (hopefully) little controversy about accuracy, measures of precision are used to compare the relative 'ease' of ageing one species over another or to compare the precision between readers within or between laboratories (Kimura and Lyons 1991, Campana 2001). A number of measures of precision are available (Campana 2001, Lai et al. 1996), but there is no consensus as to which one is most appropriate for determining precision. The index of percent precision (IPP) is the easiest

measure to calculate and understand:

$$\text{IPP} = 100 \times \frac{F}{N}$$

where F is the number of fish whose replicated, estimated ages agreed within some range, and N is the number of fish whose age were estimated. The IPP is the traditional index of precision but is falling out of favor according to Campana's (2001) review. This reversal of fortune for IPP appears to have resulted from two contrasting examples provided by Beamish and Fournier (1981; p. 982):

i) If nearly all ages estimated by two readers agree within $\pm$ 1 year this may still lead to poor precision if the species has only a few year-classes (i.e., < 10), whereas
ii) If a similarly high percentage of estimated ages agreed to within $\pm$ 5 years, this may be good precision for another species that has many more year-classes (i.e., 50-100).

Beamish and Fournier (1981) proposed the index of average percent error (IAPE) as an alternative index that should be less dependent on absolute age of the fish:

$$\text{IAPE} = 100 \times \frac{1}{N} \sum_{j=1}^{N} \left(\frac{1}{R} \sum_{i=1}^{R} \frac{\left| Y_{ij} - \overline{Y}_j \right|}{\overline{Y}_j} \right) \quad (1),$$

where N is the number of fish aged, R is the number of replicated age estimates per fish, Y_{ij} is the ith age determination of the jth fish, and $\overline{Y}_j$ is the average age for the jth fish. As an alternative to the IAPE, Chang (1982) proposed using the index of the coefficient of variation (ICV) –

$$\text{ICV} = 100 \times \frac{1}{N} \sum_{j=1}^{N} \frac{\sqrt{\sum_{i=1}^{R} \frac{\left(Y_{ij} - \overline{Y}_j \right)^2}{R-1}}}{\overline{Y}_j} \quad (2)$$

as a ratio index that should further eliminate the effect of fish age on measures of precision. The ICV should not be confused with an actual coefficient of variation (i.e., of replicate age estimates for an individual fish: $CV = s \times 100 / \overline{Y}$

where s is the standard deviation:

$$s = \sqrt{\sum_{i=1}^{R} \frac{\left(Y - \bar{Y}\right)^2}{R-1}} \quad (3)$$

.

The ICV is a summation of individual *CV* values for age estimates of individual fish so that variations in age estimates within fish, among fish, and among age classes are all combined; this 'oversummarization' of variance sources is one of the criticisms of ICV (and the IAPE; Hoenig et al. 1995). Nonetheless, the ICV has been recommended by Campana et al. (1995) and Campana (2001), and the IAPE continues to appear in recent literature (Sulikowski et al. 2003, McDougall 2004). Hoenig et al. (1995) proposed a 'test of symmetry' approach to avoid this oversummarization problem. This approach tests for asymmetrical bias between the replicate, estimated ages (i.e., away from the table diagonal in an age frequency table). It uses a chi-square (X^2) approach in the form of the test statistic:

$$X^2 = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{\left(n_{ij} - n_{ji}\right)^2}{n_{ij} + n_{ji}} \quad (4)$$

where n_{ij} is the observed frequency in the *i*th row and the *j*th column and n_{ji} is the observed frequency in the *j*th row and *i*th column. When systematic differences occur away from the diagonal, then the test statistic will become large and will be eventually rejected. The degrees of freedom are equal to only the number of paired cells with actual values to compare (i.e., nonzeros in either one or both paired [n_{ij} *vs* n_{ji}] cells). Finally, Campana et al. (1995) introduced the use of the correlation coefficient (r) as a measure of precision, and although it does not appear to be widely used, we include it in these comparisons.

Given this cacophony of choices, which approach should be used? Missing from this debate is an evaluation of these various approaches based on simulated data with known properties. In the studies cited above, a common theme was to calculate and compare these different measures of precision using "real" datasets, using ages that may or may not have had been based on a validated ageing method so there was some, unknown level of inaccuracy of the ages. In this study we present examples from a simulated dataset where accuracy and precision are known. We are specifically interested in how varying levels of imprecision affect our perception of an ageing method or of each metric of precision, even if it is without bias.

METHODS

Ages of a sample of 60 fish were simulated. In the sample there were six fish for each of 10 age classes: 1 through 10. The age of each fish was estimated twice, so there were a total of 120 (i.e., 60 paired) estimated ages for each sample. Seven sample cases were examined: a null model and six simulated cases that contained increasing amounts of random variation in the estimated ages. In the null model, the estimated ages simply agreed with the

known age of all fish. For the simulations, the "random" component varied fish ages between -1 and +1 year of the known age, and for each case, 30 runs were made. For each run, ICV, IAPE, IPP, and *r* statistics were estimated. IPP was calculated as the percentage of age estimates that agreed exactly. In case 1, both pairs of the estimated ages for one fish per age class were set to vary in a random manner; in case 2, both pairs of the estimated ages for two fish per age class were set to vary; this pattern was followed up to case 6, for which variations of all ages of all fish were randomly assigned (-1, 0, +1). In summary, the null model represented no aging error, and the alternative cases represented unbiased ageing, with increasing amounts of imprecision ranging from 17% to 100% of the ages being potentially incorrect (Figure 1).

Each of these six cases could illustrate a real situation in which most of the annuli are easy for an observer to read, but there might be one annulus that is difficult to read, so the observer is as likely to miss it as to see it or to add another annulus count. The nature of the variance in this example is homogeneous. The ageing method is not specified here but can be considered to result in direct annual ages from any one of a variety of biological hardparts.

RESULTS

In the null model (i.e., no ageing error), ICV and IAPE = 0, IPP = 100, and r = 1.0 (Figure 1). As simulated ageing error increased (i.e., from case 1 to case 6), ICV and IAPE increased and IPP and r decreased (Figures 1, 2). Average values of ICV for each case were two to three times higher than IAPE. In case 6 the average ICV was 19.7 (vs. IAPE = 7.0). Average values of IPP decreased across the widest range of absolute values: from 88.7 in case 1 to 32.1 in case 6. Average values of r decreased across a very narrow range of absolute values: from 0.986 in case 1 to 0.926 in case 6. IAPE had the greatest dispersion around each mean (CV ranged from 28.1 to 70.9% in each of the six cases), ICV had less variation (CV: 15.6 - 30.8%), IPP had even less (CV: 2.4-21.4%), and r had the least (CV: 0.464 - 1.25%).

Age-specific trends in ICV, IAPE, and IPP were evident in a number of simulations (Figure 3). Although these indices could be remarkably even across age-classes in some simulated runs (e.g., Figure 3I), ICV and IAPE usually demonstrated some declining trend with respect to age-class, and IPP was typically highly variable.

All four precision measures evaluated showed a strong linear relationship with each other, explaining as much as 90% of the variation between them (Figure 4). Coefficients of determination (r^2)were lowest when comparing IAPE to other indices (0.59 - 0.76), as might be expected because IAPE had the highest dispersion (see above regarding CV values). The relationship between the ICV and IAPE was very different for these simulated data than their relationship for real data plotted by Campana (2001; p. 223) (Figure 5).

In the simulated data, the inaccuracies in all six cases were designed to not have any bias with respect to repeated estimates of age, and the use of a test of symmetry confirmed that there was, in fact, no bias between paired age estimates (Table 1).

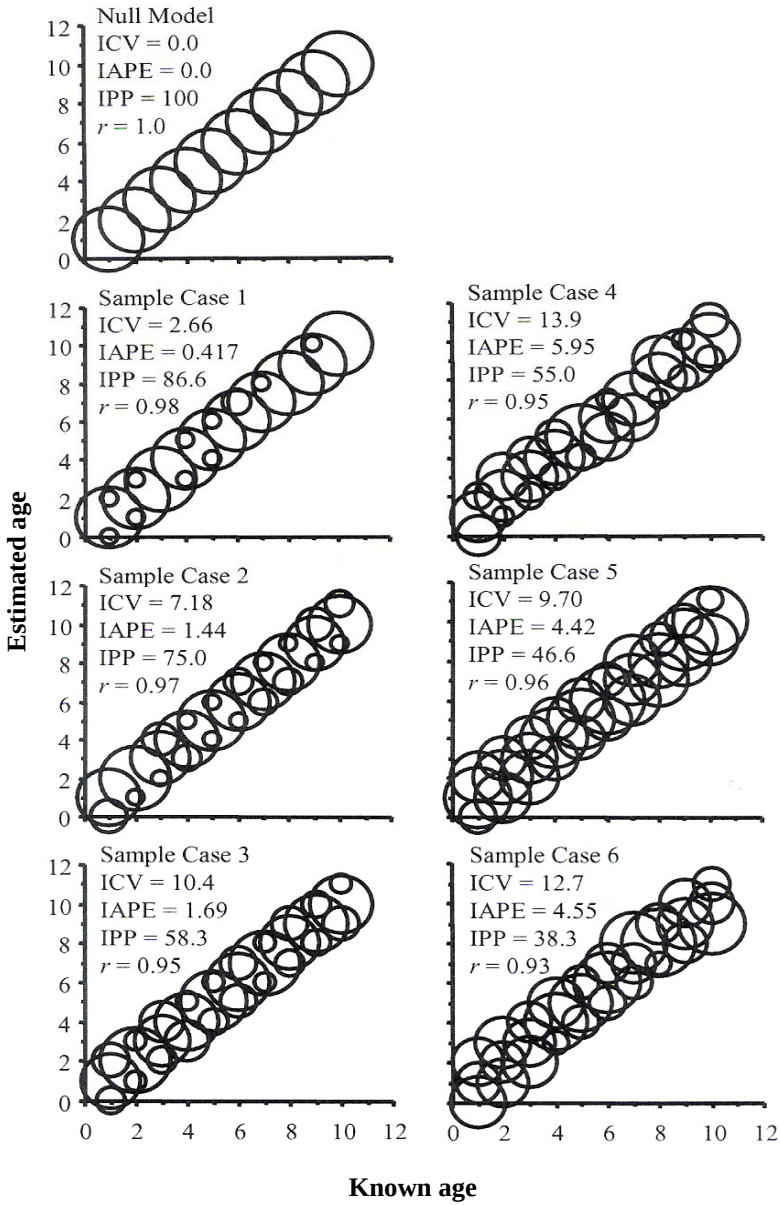

Figure 1. Bubble plots of the null model (i.e., no ageing error) and example runs of the six alternative cases, with increasing variability (increasing ageing error / decreasing ageing precision). The data plotted are for the last simulated run of 30 total runs per case (1-6) and are meant to be random examples (actual precision values for that particular run are also listed).

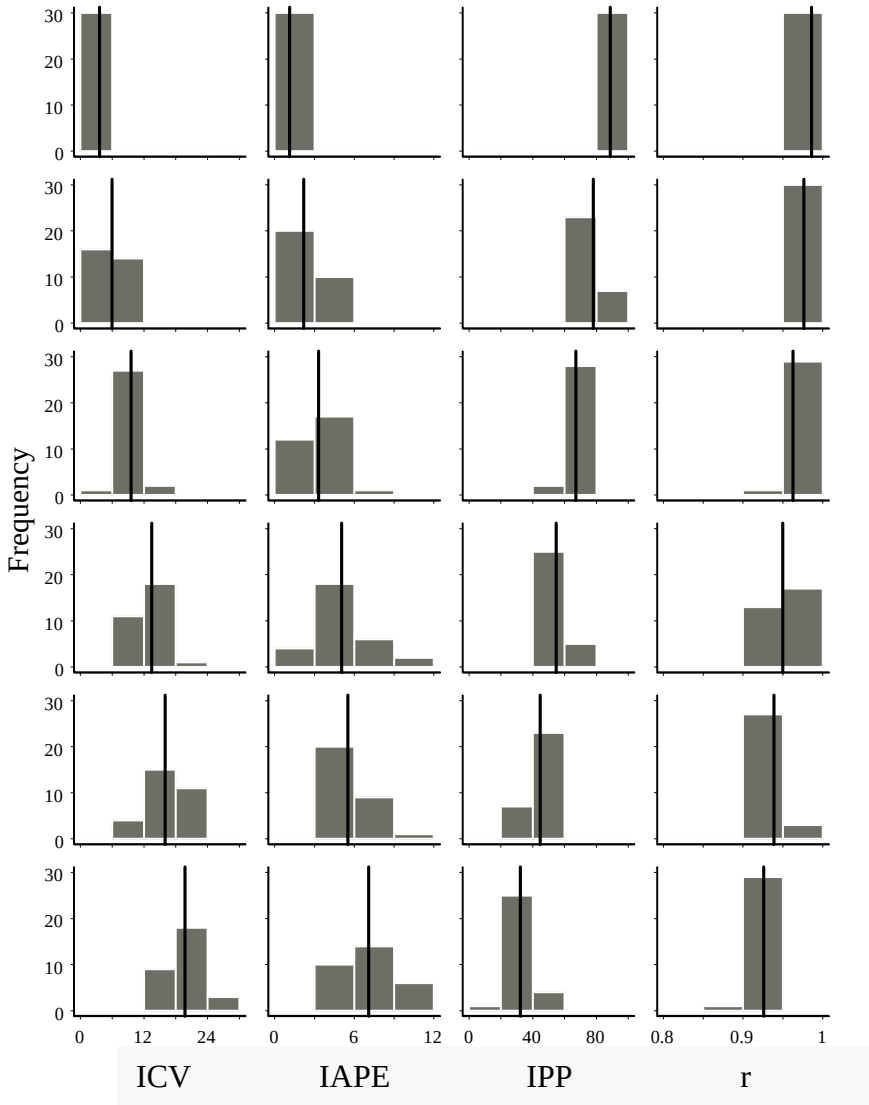

Figure 2. Histograms of three indices of precision and *r* generated from 30 simulated runs for each of 6 cases of increasing ageing error. The four indices are labeled (ICV, IAPE, IPP, *r*; left to right) and the cases are in order of increasing ageing error (case 1-6; top to bottom). The vertical bars represent the mean value for each case. See Figure 1 for scattergram examples of each case and the Introduction section for calculations of each index.

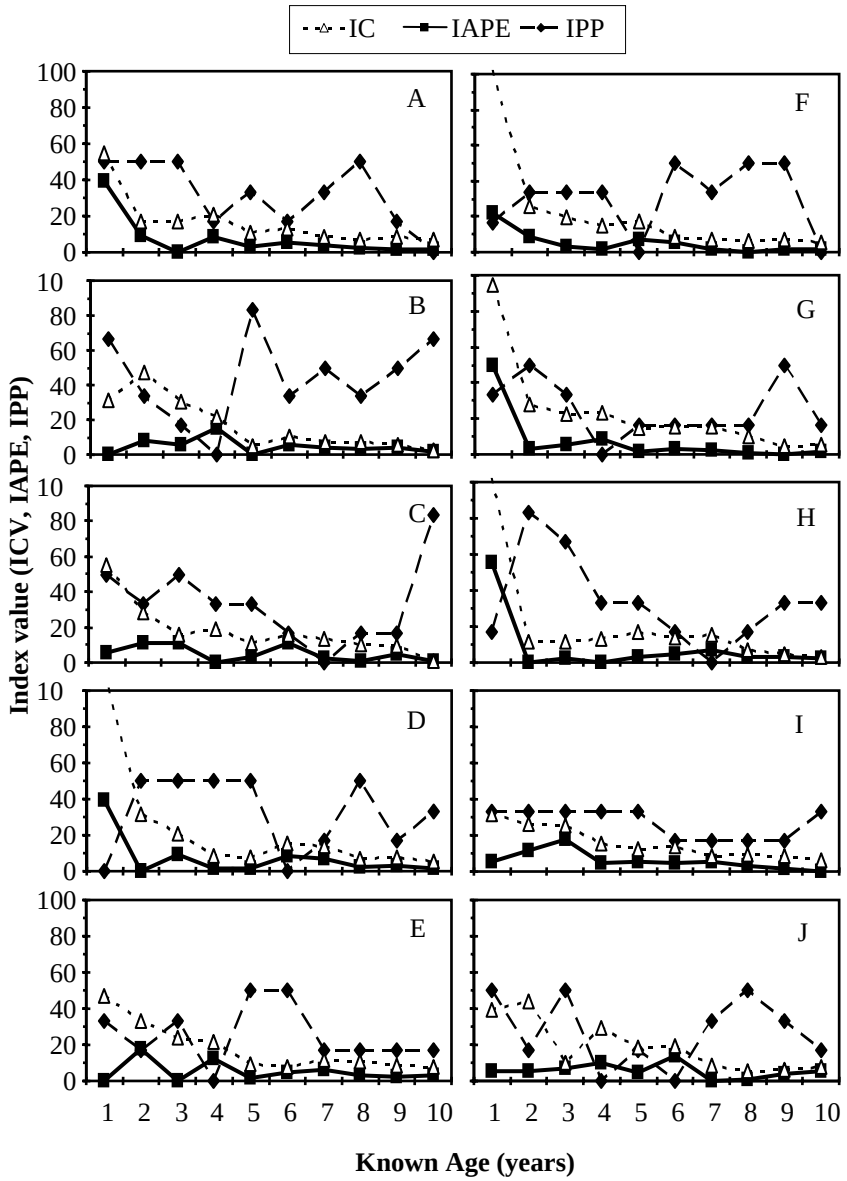

Figure 3. Ten random simulation runs (A-J) of age-specific patterns of ICV (triangles, short dash line), IAPE (squares, solid line), and IPP (diamonds, long dash line).

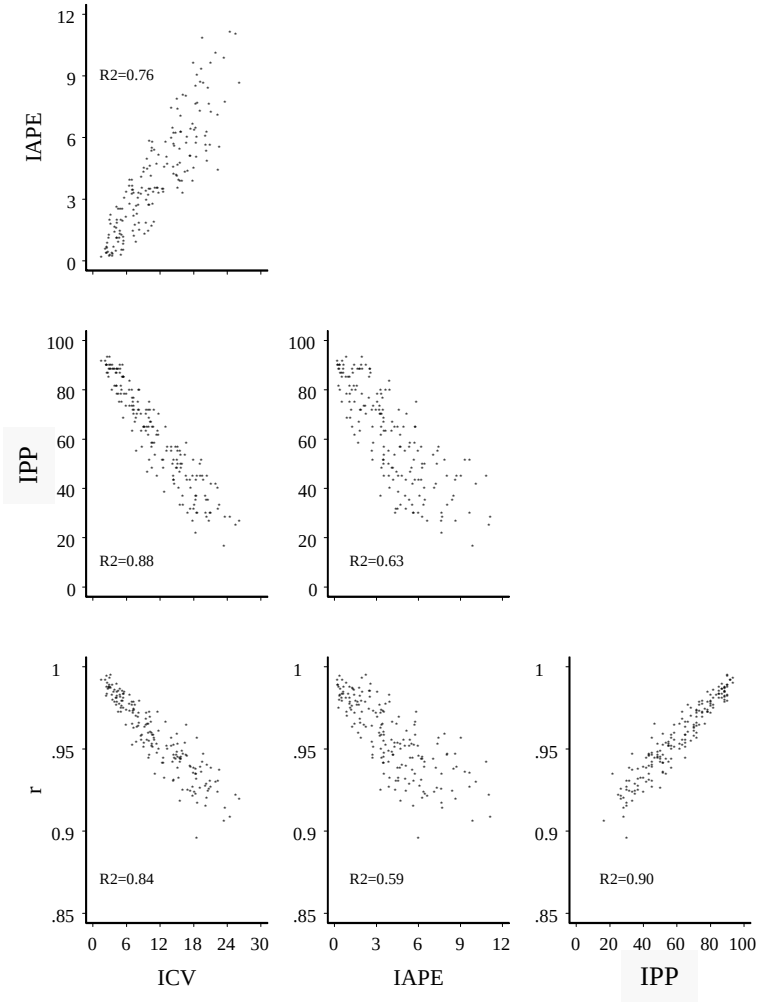

Figure 4. The association between all three indices and *r* with the corresponding coefficient of determination (r^2) between each variable. There are 180 data points per scattergram (30 simulations of six cases).

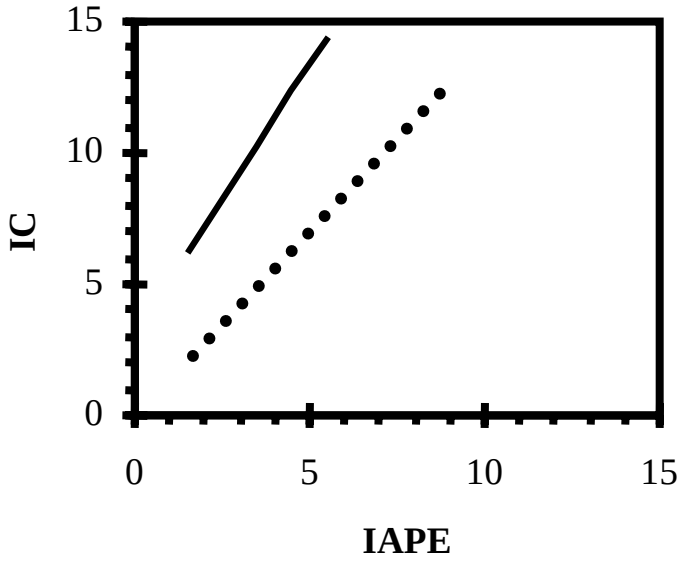

Figure 5. ICV regressed against IAPE as calculated for this simulated data (solid line) and for data from the literature (dotted line) as reported by Campana (2001; p. 223). The equation for the simulated data is ICV = 3.06 + 2.06 × (IAPE) and was calculated by least squared regression (see Fig. 4 [top panel] for actual data). The equation reported by Campana (2001) is ICV = -0.15 + 1.41 × (IAPE).

Table 1. (A) A random example of an age-frequency table that compares paired estimated ages for one simulated run of case 6 (see Methods for descriptions of the six cases).

	First estimated age (R1)											
R2	0	1	2	3	4	5	6	7	8	9	10	11
0	1	1										
1	2			1								
2	2		2	2	1							
3		2		1								
4			1	1	5	1	1					
5				1	2	1	3					
6						3	1	1	1			
7								1		1		
8								3	4		2	
9									2	3	2	
10											1	
11										3	1	

Table 1. (B) The chi-squared values (X^2) , degrees of freedom (df), and probability (P) for 20 simulated runs of case 6. In all 20 runs the null hypothesis (symmetry along the diagonal) was not rejected (α = 0.05). This was anticipated because there was no bias introduced into the simulated data. The test results for the example above (A) are in the lower right-hand corner of (B).

X^2=	**15.3**	**19.7**	**17.5**	**31.0**	**12.2**	**18.0**	**13.3**	**19.0**	**15.1**	**14.2**
df=	16	18	20	22	16	16	17	18	16	16
P=	0.500	0.352	0.622	0.096	0.730	0.324	0.714	0.392	0.514	0.584
X^2=	14.9	25.2	18.3	19.0	17.7	13.2	23.9	18.7	21.2	24.0
df=	15	18	17	20	18	18	18	16	18	19
P=	0.461	0.120	0.368	0.522	0.478	0.780	0.159	0.286	0.269	0.196

DISCUSSION

All three indices (ICV, IAPE, and IPP) and *r* were correlated with one another. Campana (2001) also demonstrated that ICV and IAPE were correlated with each other, which should not be surprising because ICV merely replaces the average absolute deviation used in the IAPE with a standard deviation in its formulation (Chang 1968). Still, although these two indices should be related to each other, we urge caution in applying Campana's (2001) regression relationship between ICV and IAPE, because using our data, this relationship was very different from that reported by Campana (2001) (Figure 5). Regression equations between these three indices and *r* may be highly dependent on the specific data sets used, so generalizations may be misleading.

The importance of the correlations between the three indices and *r* is that they all say more or less the same thing. If so, then what guiding principles are there to choose one over the other? First, there should be sufficient range in the absolute value of an index to allow one to distinguish low precision from high precision. The correlation coefficient (*r)* had very little range from case 1 to case 6. Even for case 6, where every estimate of age could randomly be incorrect, the average *r* was quite high (0.926). To an untrained scientist, a report of $r > 0.9$ might very well be misconstrued as indicating very good precision, when it could actually indicate very poor precision. The effect of bias between paired reads may actually lower *r* even more, but our example did not have any bias with respect to known age.

Another guiding principle is that the index chosen should not have a high variability inherent in its formulation. The index with the highest relative variability, as measured by *CV*, was IAPE. This high variability probably results in the use of absolute deviation in the IAPE formula. In fact, use of the standard deviation in the formulation of ICV does appear to lower the variability of the ICV, which makes this index statistically more robust. The formulation of ICV also increases the range of ICV values between the six different cases, which we regard as generally a good quality (see above). The variability of IPP was lower than the variability of either IAPE or ICV, which was somewhat surprising because IPP has been criticized as having age-related bias for certain species (Beamish and Fournier 1981). Our finding of lower variability in the IPP may simply be the result of the modest longevity (10 years) of our hypothetical fish population and the homogeneity of variance that we imposed on the simulated data. In the future, we intend to rerun simulations that extend the maximum age out by several decades and includes heteroscedasticity. One of the central tenets of the superiority of one index of precision over another is that there are no age-specific trends in an index. In this particular example, it is then relevant to note that the age-specific variances of IPP are not particularly worse than those of ICV and IAPE.

IPP has a real advantage over ICV and IAPE: it is easy to interpret. It has obvious boundaries (0 - 100), and a value of 80% clearly means that 80% of the paired reads agreed and 20% did not. Only experienced otolithologists have a feeling for various values of ICV and IAPE. Campana (2001) helps remedy this to some degree by summarizing published values of ICV; he reported a median of 7.6 and a mode of 5. Of course, these values may be

artificially low if researchers are reticent about publishing high values. The danger is that reviewers or editors may use these values as benchmark criteria for accepting or rejecting the quality of a study. In our simulated study, an ICV of about 5 was associated with case 2 (i.e., 1/3 of the otolith age estimates had the potential to vary $\pm$ 1 year). We agree with Campana (2001) that there is no *a priori* target value, and the criteria for evaluating a particular index value depend on the objective of each study.

Much has been said before about these indices, but most of this has been based on real datasets for which accuracy and bias were not known. Using simulated data with no bias with respect to known age, we varied the precision of paired age estimations. We included *r* in this comparison because it has a solid statistical basis, but the resulting values of *r* in this simulation strike us as misleading in terms of evaluating imprecision so we do not recommend its use. We find little difference between ICV and IAPE, except that ICV ranks higher because its formulation leads to greater values than the formulation for IAPE (i.e., greater absolute range of values between each case), and the statistical rigor of ICV accounts for greater stability of this index across age-classes. Finally, we conclude that much of the criticism directed at IPP may be unwarranted. All three indices, not just IPP, tend to oversummarize the data (Hoenig et al. 1995). IPP is easy to measure and easy to understand, which is an apparently unsung advantage. The cumulative effect of several papers (Beamish and Fournier 1981, Chang 1982, Kimura and Lyons 1991, Campana et al. 1995, Lai et al. 1996, Campana, 2001) criticizing IPP amounts to tossing the baby out with the bathwater.

Researchers have not stopped using IPP completely. An interesting trend in the literature is that IPP is used to screen datasets prior to calculating either IAPE or ICV. For example, Sulikowski et al. (2002) calculated IAPE only for those cases where within-fish variability was $\leq$ 2 years. Simpfendorfer et al. (2000) used an ordinal system for characterizing the readability of individual vertebrae, and calculated IAPE only for those that ranked above a certain value. IPP is used in these publications to improve the index of precision actually reported (i.e., as a stealth index to weed out specimens with markedly poor precision).

In this study, a test of symmetry approach was briefly introduced. In our example, it is somewhat trivial because the simulated data had no bias, so finding no bias was not unexpected. Using a test of symmetry to check for asymmetry between paired (but independent) age estimates by the same reader is also likely to be trivial, unless this reader changes some interpretive criteria between paired estimates. A test of symmetry is more commonly used to compare the age estimates of two readers (e.g., expert *vs.* novice) or between two biological hardparts (otoliths *vs.* scales). The value of such a test is that it sets up an age-frequency table, which graphically displays the data and then tests it for departures from symmetry. In comparison, the various indices of precision (ICV, IAPE, and IPP) do not measure any such bias. On the other hand, the test of symmetry does not provide a simple value for precision; in fact its formulation does not include the diagonal cells. We recommend that fishery scientists should evaluate ageing error using both an index of precision together with tests of symmetry.

In this study we present an artificial, or at least a restricted, example, but one with heuristic value. As mentioned above, we anticipate extending this simulation approach to include a greater variety of underlying models, particularly for greater longevity and other types of variance and bias with respect to known age. Here we simply point out that:

i) IPP has perhaps been overcriticized for some simple shortcomings that other, more complicated indices do little to overcome, and

ii) Running a test of symmetry in addition to calculating an index of precision provides a useful way for researchers to evaluate their age data.

ACKNOWLEDGMENTS

We thank the FWRI stock assessment group for comments on an early synthesis of these ideas and B. Muller and J. Tunnell for their reviews of the manuscript. We are also grateful to E. Sosa for translating the English abstract to a Spanish version. Support for RSM was provided, in part, by the Department of Interior, U.S. Fish and Wildlife Service, Federal Aid for Sport Fish Restoration, Grant Number F-106. Funding for this study was also supported in part by the U.S. Department of Commerce, National Oceanic and Atmospheric Administration (DOC-NOAA) award numbers NA17FF2882 (support for KLM) and NA16FG1221 (support for JdeS). The views expressed herein are those of the authors and do not necessarily represent the views of the DOC-NOAA.

LITERATURE CITED

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

Campana, S.E. 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. *Journal of Fish Biology* **59**:197-242.

Campana, S.E., and S.R. Thorrold. 2001. Otoliths, increments and elements: keys to a comprehensive understanding of fish populations? *Canadian Journal of Fisheries Aquatic Sciences* **59**:30-38.

Campana, S.E., M.C. Annand, and J.I. McMillan. 1995. Graphical and statistical methods for determining the consistency of age determinations. *Transactions of the American Fisheries Society* **124**:131-138.

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

Hoenig, J.M., Morgan, M.J., and C.A. Brown. 1995. Analysing differences between two age determination methods by tests of symmetry. *Canadian Journal of Fisheries Aquatic Sciences* **52**:364-368.

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

Lai, H.-L., V.F. Gallucci, D.R. Gunderson, and R. F. Donnelly. 1996. Age determination in fisheries: methods and applications to stock assessment. Pages 82-178 in: V.F. Ballucci, S.B. Saila, D.J. Gustafson, and B.J. Rothschild (eds.). *Stock Assessment: Quantitative Methods and Applications for Small-scale Fisheries*. CRC Press, Lewis Publishers, Boca Raton, Florida USA.

McDougall, A. 2004. Assessing the use of sectioned otoliths and other methods to determine the age of the centropomid fish, barramundi (*Lates calcarifer*) (Bloch), using known-age fish. *Fisheries Research* **67**:129-141.

Simpfendorfer, C.A., J. Chidlow, R. McAuley, and P. Unsworth. 2000. Age and growth of the whiskery shark, Furgaleus macki, from southwestern Australia. *Environmental Biology of Fish* **58**:335-343.

Sulikowski, J.A., M.D. Morin, S.H. Suk, and W.H. Howell. 2003. Age and growth estimates of the winter skate (*Leucoraja ocellata*) in the western Gulf of Maine. *Fisheries Bulletin, U.S.* **101**:405-413.

BLANK PAGE