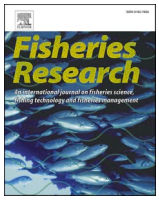
Fisheries Research 249 (2022) 106255

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



Influence of sample size and number of age classes on characterization of ageing error in paired-age comparisons

Geneviève Nesslage [a,*], Amy M. Schueller [b], Amanda R. Rezek [b], Raymond M. Mroch III [b]

[a] University of Maryland Center for Environmental Science, Chesapeake Biological Laboratory, Solomons, MD 20688, USA
[b] National Marine Fisheries Service, Southeast Fisheries Science Center, Beaufort Laboratory, 101 Pivers Island Road, Beaufort, NC 28516, USA

ARTICLE INFO

Handled by: A.E. Punt

Keywords:
Bias
Imprecision
Ageing error
Simulation
Symmetry

ABSTRACT

Diagnosis of ageing error is critical to the proper interpretation of age data used in fisheries science and management. However, the influence of sample size and number of age classes on the characterization of ageing error has not been thoroughly evaluated. We conducted a simulation study of ageing error diagnostics for paired-age comparisons across 648 scenarios differing in 1) number of age classes, 2) total number of samples aged, 3) trend in sample size by age, and 4) magnitude and type of imprecision and bias. Imprecision was identified by comparing average coefficient of variation (ACV) with two common thresholds. Bias was evaluated using maximally (McNemar's), diagonally (Evans & Hoenig), and unpooled tests of symmetry (Bowker's). Imprecision was identified less frequently at low to moderate ($\bar{x}$=6% of runs) levels of random vs high ($\bar{x}$=55% of runs) error, and ACV was artificially inflated in the presence of bias. McNemar's and Evans & Hoenig bias tests outperformed Bowker's ($\bar{x}$=4% vs 29% false positives), particularly at large sample sizes, and its use is strongly discouraged. This study can help guide the interpretation of ageing error studies and their products (e.g., ageing error matrices) used to inform stock assessment and management.

1. Introduction

The age of many fish can be determined by the examination of periodic growth increments in certain calcified structures, typically otoliths, scales, vertebrae, or spines (Anon, 2019). Fish age data collected from both fishery-dependent and fishery-independent sources can be used to characterize species life histories and inform the estimation of population dynamics (Fournier and Archibald, 1982; Quinn and Deriso, 1999). The majority of stock assessment models worldwide incorporate age structure (Ricard et al., 2012), and the resulting stock status and catch limit advice often rely on the use of catch-at-age data (Maunder and Punt, 2013; Methot and Wetzel, 2013). The ability to adequately detect and characterize ageing error is critical to informing the proper interpretation of age data used in fisheries stock assessments and management (Dorval et al., 2013; Punt et al., 2008). Ageing error matrices are often incorporated in stock assessment models to account for uncertainty in the catch-at-age data used to inform estimation of recruitment and overall age composition (Punt et al., 2008; Thorson et al., 2012). However, subsequent advice provided to fisheries managers from assessments that rely upon catch-at-age data can be highly uncertain and overly optimistic if error in age estimates is not properly identified and addressed (Beamish and McFarlane, 1995; Henríquez et al., 2016; Maunder and Piner, 2015; Reeves, 2003).

Best practices employed by ageing laboratories involve regular evaluation of ageing error, including accuracy (age estimates compared with true ages), precision (repeatability of age estimates), and bias studies (systematic error in age estimates; Campana et al., 1995; Morison et al., 2005). Adequate detection and characterization of ageing error is an essential component of providing age data for stock assessments and improving ageing laboratory performance through reference collection exchanges, workshops, training exercises, and routine QA/QC efforts. Thus, a suite of diagnostics has been developed to quantify accuracy, precision, and bias in estimated ages (McBride, 2015; Anon, 2019). Ageing laboratories routinely conduct visual inspection of their data in a variety of ways, including examination of age-bias plots and age frequency tables (Campana et al., 1995). Descriptive statistics and simple statistical tests are routinely calculated to provide a quantitative measure of accuracy and precision (Campana et al., 1995; Lai et al., 1996; Anon, 2019). Common measures of precision used in ageing error studies are average percent error (APE; Beamish and Fournier, 1981)

Abbreviations: ACV, average coefficient of variation; McN, McNemar's; E&H, Evans and Hoenig; Bowk, Bowker's tests of symmetry.

* Corresponding author.
E-mail address: nesslage@umces.edu (G. Nesslage).

https://doi.org/10.1016/j.fishres.2022.106255
Received 6 July 2021; Received in revised form 21 January 2022; Accepted 25 January 2022

and average coefficient of variation (ACV; Chang, 1982). Although there are numerous ways to examine and interpret these measures (Kimura and Anderl, 2005), many ageing laboratories characterize precision by comparing APE and ACV to a threshold representing a maximum acceptable level of imprecision (McBride, 2015). Although this imprecision threshold varies among laboratories, many use an ACV of seven (equivalent to APE of five for paired ages) or ten.

Many laboratories also use tests of symmetry to detect bias (Cailliet et al., 2006; Elzey et al., 2015; Liao et al., 2019; Matta and Kimura, 2012; Sutherland, 2020; Anon, 2019). Such tests diagnose bias by detecting asymmetry in the distribution of disagreements between paired-age comparisons (McBride, 2015). Common tests of symmetry include McNemar's maximally pooled (McNemar, 1947), Evans & Hoenig's diagonally pooled (Evans and Hoenig, 1998), and Bowker's unpooled (Bowker, 1948) tests of symmetry, which differ in their methods for combining elements along the off diagonals of the paired comparisons contingency table (see McBride, 2015 Supplementary Materials for a thorough review of bias testing methods).

McBride (2015) conducted a simulation study to compare the characterization of ageing error by different diagnostics when presented with paired ages that differed in the amount of random and systematic error. In McBride's study, combinations of different types of error (both imprecision and bias) were applied to samples of five fish (ten in a smaller subset of tests) spanning 20 age classes. The ability of tests of symmetry to detect bias when present was evaluated and the magnitude of precision measures (APE and ACV) relative to imprecision thresholds was characterized. McBride thoroughly assessed these diagnostics; however, his conclusions were limited to the scenarios of five and ten samples aged uniformly across all age classes for a fish that lived 20 years. The question of how larger sample sizes, trends in sample size by age (e.g., decreasing sampling at older ages), and number of age classes in the population might affect characterization of ageing error was not explored. The influence of number of samples and trends in sample size on the performance of tests of symmetry has been discussed in the literature, but not simulation-tested across a wide range of sample sizes (Evans and Hoenig, 1998; McBride, 2015). Also, longevity of a species may impact the ability to characterize ageing error; for example, gathering adequate sample sizes per age class to reliably detect ageing error can be challenging for long-lived species and for age classes that are not vulnerable to fishing or sampling gear. If the potential to diagnose ageing error changes with age and sample size, biases can be introduced into the catch-at-age matrix, ageing error matrices, and life history estimates such as maturity-, weight-, length-, and mortality-at-age.

We expanded upon the simulation study of McBride (2015) to more broadly examine the impact of number of age classes and sample size on the interpretation of ageing error diagnostics in paired-age comparisons. Our objectives were to determine how sample size, trend in sample size, and number of age classes affect:

1) characterization of imprecision given the magnitude of ACV relative to alternative thresholds, and
2) relative performance of tests of symmetry in their ability to detect different types of systematic bias

across a range of random error levels. This simulation study aims to enhance our understanding of how common ageing error diagnostics should be interpreted in light of a wide range of ageing error and data collection scenarios for fish with different maximum ages.

2. Materials and methods

We characterized the influence of sample size, sample size trend by age, and number of age classes on the diagnosis of ageing error in paired-age comparisons by conducting a simulation study in R Version 4.1.0 (R Core Team, 2020) with ageing error diagnostics provided by the FSA package Version 0.8.32 (Ogle et al., 2021). We generated data sets of known ages for 54 different sampling scenarios to which 12 error scenarios were applied to generate a set of estimated ages (Fig. 1). Ageing error was then characterized by comparing ACV to imprecision thresholds and by performing tests of symmetry.

Known-age data sets differed with regard to 1) number of age classes (three scenarios), 2) total sample size (six scenarios), and 3) sample size trend by age (three scenarios). Three alternative scenarios for number of age classes were explored representing short-, medium-, and long-lived fish with maximum ages of 5, 20, or 50, respectively. Known-age sets were created that contained five samples of known ages uniformly distributed across all age classes for each of the maximum age scenarios with total sample sizes of 25, 100, and 250, respectively. To examine the impact of total sample size, five additional scenarios were explored for each number of ages such that the base number of samples ($n = 5$) was multiplied by a factor of 2, 5, 10, 20, or 40 per age class, representing a 100–3900% increase. Thus, the total number of samples ranged from 25 to 1000 for the maximum age-5 scenario, 100–4000 for the maximum age-20 scenario, and 250–10,000 for the maximum age-50 scenario. These scenarios were chosen because, while most stocks are minimally sampled near the low end of these sampling ranges, several high-profile stocks such as Gulf menhaden (*Brevoortia patronus*; SEDAR, 2018), Sablefish (*Anoplopoma fimbria*; Table 3.8; Goethel et al., 2020), and Atlantic Herring (*Clupea harengus*; Table A1–6; Northeast Fisheries Science Center, 2012) are sampled near the middle to high end of these ranges. Simulating a wide range of sample sizes allowed us to provide more comprehensive advice across a wide range of sampling situations.

To explore the impact of non-uniform collection of samples across all age classes for a given stock, we generated two alternative sample size trend scenarios in which the number of samples either decreased exponentially with age or peaked at the middle age. The declining trend in sample size with age mimicked an expected decrease in availability of older fish to sample if abundance declines exponentially with age. The peaked trend represented a common problem in which younger and older fish are not often encountered in the sampling program due to gear selectivity or lack of spatial overlap. For these trended scenarios, the total number of samples was kept the same as the uniform scenario, but the number of samples by age class varied (Fig. S1A).

A set of estimated ages were simulated to generate a paired read with error for each of the 54 known-age data sets using three levels of imprecision and four types of bias as described in McBride (2015) for a total of 648 total scenarios (54 sampling scenarios × 12 error scenarios). Imprecision in the estimated age for each pair of reads (one known and one estimated) was simulated by adding to the known age a normally distributed error with a mean = 0 and coefficient of variation (CVs) of either 0.05, 0.10, or 0.15 times the known age (termed precision levels "CV5″, "CV10″, and "CV15″, respectively). This approach assumes random ageing error increases with age. Bias scenarios included: 1) no bias, 2) persistent bias of true age with one additional year observed (termed "+1″), 3) 10% underestimation of each age class in units of years (termed "−10%") to simulate increasing bias with age (i.e., under-ageing), and 4) both of the previous two bias patterns combined (adding one year followed by 10% underestimation) to simulate over-ageing younger fish and under-ageing older fish (Robillard et al., 2009). Each of the 648 scenarios was repeated 1000 times (i.e., 1000 runs; see Fig. S1B for example of resulting error patterns generated).

To characterize ageing error for each run, we focused on ACV as a measure of precision and McNemar's (McN), Evans & Hoenig (E&H), and Bowker's (Bowk) tests of symmetry to diagnose bias because these tests are commonly used across most ageing laboratories around the world (Elzey et al., 2015; Liao et al., 2019; Matta and Kimura, 2012; Anon, 2019). ACV was calculated as:

$$ACV = 100 * \frac{1}{n} \sum_{j=1}^{n} \frac{sd_j}{\bar{x}_j} \quad (1)$$

where n was the number of times each fish was aged ($n = 2$), sd_j was the

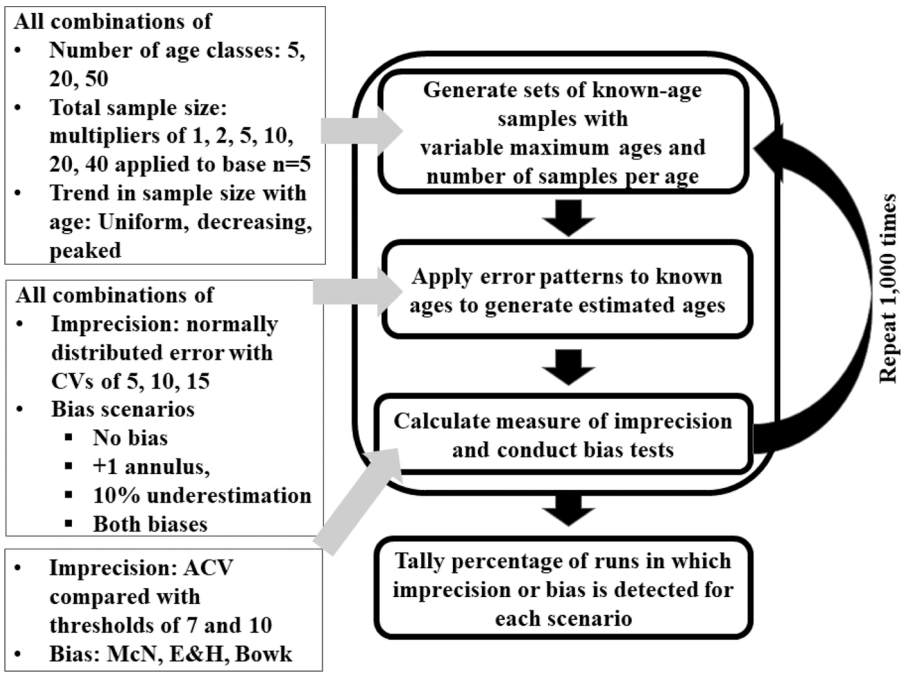


Fig. 1. Flowchart of simulation study steps and variable components. CV = coefficient of variation. Bias types included 10% underestimation of true age (−10%), overestimation by one year (+1), and both types (Both). Average CV = ACV, and tests of symmetry included McNemar's (McN), Evans & Hoenig (E&H), and Bowker's (Bowk).

standard deviation for the age estimates of the jth fish, and $\bar{x}_j$ was the mean age estimate of the j^{th} fish (Chang, 1982). APE was not included because it is redundant with ACV, which is 41% higher when the number of reads is two (Campana, 2001; Chang, 1982; Kimura and Anderl, 2005). ACV values were then compared with two commonly used imprecision thresholds of ACV of seven and ten (Campana, 2001; McBride, 2015), and the percentage of runs greater than each threshold was tallied for each scenario. The percentage of *P*-values for each test of symmetry that exceeded the significance level of $\alpha = 0.05$ was tallied for each run as well.

One additional set of simulations was run to examine the impact of pairing a known age with an estimated age on the magnitude of ACV and the characterization of precision. To do this, we repeated our simulations of random error by generating two estimated reads instead of just one, and calculated the resulting ACV for the two estimated ages. We then compared those ACV values with threshold values of seven and ten, as in our base set of simulations. This alternative simulation approach allowed us to explore the impact of number of age classes and sample size on ageing imprecision diagnostics when two read ages with error are evaluated.

3. Results

The number of age classes, total sample size, and trend in sample size, as well as the magnitude and type of ageing error simulated impacted the magnitude of ACV relative to imprecision thresholds and the ability of tests of symmetry to detect bias (Figs. 2 and 3). When interpreting Fig. 3, note that we expect bias to be detected based on chance alone for approximately 5% of the runs for each scenario given $\alpha = 0.05$. A set of figures containing detailed simulation results for all 648 scenarios can be found in Supplemental materials (Figs. S2–S5).

3.1. Imprecision

For scenarios in which random error was simulated without bias, mean ACV increased with increasing random error, and was higher for scenarios with peaked trends in sample size (Figs. S2–1–S2–9). As might be expected, ACV was higher in magnitude in our alternative simulation in which both reads in the pair were estimated (Fig. 4).

The median response relative to a given threshold was not influenced by sample size, but the distribution of ACV did narrow with increasing sample size (Figs. S2–1–S2–9). Although not used to detect bias, ACV increased when both types of ageing error (both random error and bias) were simulated (Fig. 2B and C). Thus, the tendency for ACV to exceed imprecision thresholds (either seven or ten) increased with the total amount of error, both random and systematic. In our base simulations in which paired ages included one known age and one estimated age, imprecision was only identified using a threshold of seven at higher levels of random error (CV15) for scenarios in which bias was not present ($\bar{x}$ = 20%; Fig. 2A). Imprecision was identified only rarely using a threshold of ten ($\bar{x}$ = 0.07%; Fig. 2A). However, in our alternate simulation in which both ages being compared were estimated, imprecision was identified at thresholds of seven and ten for a greater percentage of runs ($\bar{x}$ = 46% and $\bar{x}$ = 23%, respectively; Fig. 4). Imprecision was less likely to be identified when the trend in sample size was decreasing.

3.2. Bias

McNemar's and Evans & Hoenig tests of symmetry performed similarly across a wide range of age class and sample size scenarios simulated (Fig. 3B and C). The rate of false positive bias detection for McNemar's and Evans & Hoenig tests ($\bar{x}$=4%) was low and without pattern across all scenarios in which only random error was simulated (Fig. 3A). In contrast, Bowker's test of symmetry generated a large number of false positives ($\bar{x}$=29%) that indicated bias was present when it was not simulated, particularly when sample sizes were moderate to large (i.e., sample size multipliers of five and 20).

All three tests of symmetry failed to detect bias in some situations depending on the type of bias simulated, the number of age classes, and the total number and trend in sample size with age (Fig. 3B and C). Also, all three tests of symmetry had trouble detecting the bias pattern with an underestimation of − 10%, particularly when the number of age classes and sample sizes were smaller. Bias was detected in almost all runs when present by both McNemar's and Evans & Hoenig tests for scenarios in

A

Max Age	Precision Level	ACV (7) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	ACV (10) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20
5	CV5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV15	17	2	0	2	0	0	94	100	100	1	0	0	0	0	0	1	0	0
20	CV5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV15	77	97	100	29	11	1	98	100	100	0	0	0	4	0	0	0	0	0
50	CV5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV15	100	100	100	43	32	21	100	100	100	0	0	0	0	0	0	0	0	0

B

Max Age	Bias Type	ACV (7) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	ACV (10) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20
5	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0
	-10%	0	0	0	0	0	0	77	96	100	0	0	0	0	0	0	0	0	0
	Both	100	100	100	100	100	100	0	0	0	100	100	100	100	100	100	0	0	0
20	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0
	-10%	21	3	0	0	0	0	35	18	2	0	0	0	0	0	0	0	0	0
	Both	99	100	100	100	100	100	98	100	100	0	0	0	100	100	100	0	0	0
50	+1	0	0	0	100	100	100	0	0	0	0	0	0	100	100	100	0	0	0
	-10%	77	96	100	1	0	0	97	100	100	0	0	0	0	0	0	0	0	0
	Both	1	0	0	100	100	100	0	0	0	0	0	0	100	100	100	0	0	0

C

Max Age	Bias Type	ACV (7) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	ACV (10) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20
5	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	76	93	100
	-10%	64	78	96	24	10	1	100	100	100	22	4	0	4	0	0	85	100	100
	Both	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	19	3	0
20	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	-10%	100	100	100	80	97	100	100	100	100	66	85	98	38	30	17	88	100	100
	Both	100	100	100	100	100	100	100	100	100	98	100	100	100	100	100	100	100	100
50	+1	100	100	100	100	100	100	100	100	100	78	94	100	100	100	100	1	0	0
	-10%	100	100	100	100	100	100	100	100	100	95	100	100	38	28	13	99	100	100
	Both	100	100	100	100	100	100	100	100	100	97	100	100	100	100	100	39	30	13

Fig. 2. Summary of average coefficient of variation (ACV) results for scenarios which simulated A) only random ageing error (imprecision), B) both random error with CV= 5 and three types of bias, and C) both random error with CV= 15 and three types of bias. Columns group results by 1) ACV, using thresholds for detecting imprecision of seven and ten as noted in parentheses, 2) trends in sample size (uniform, decreasing, and peaked), and 3) a subset of low, medium, and high sample size multipliers of 1, 5, and 20 simulated. Rows group results by the number of age classes ("Max Age" of 5, 20, or 50). In A, precision levels indicate CVs used to generate random error of 5 (CV5), 10 (CV10), or 15 (CV15). In B and C, bias types include overestimation by one year (+1), 10% underestimation of true age (−10%), and both types (Both). Each box contains the percentage of 1000 runs for each scenario that were greater than the ACV imprecision threshold. Shading represents the percentage of 1000 runs in each of the following categories: 100% (black), > 5 and < 100% (light gray), or ≤ 5% (white).

which error was applied by adding one year to the known age (+1). Bowker's test failed to detect bias in a greater number of scenarios, particularly when sample sizes were low. At low levels of random error, rate of bias detection generally increased with the number of age classes and sample size (Fig. 3B).

4. Discussion

This study expands upon the work of McBride (2015) by highlighting the importance of considering the number of age classes, sample size, and trend in sample size when characterizing ageing error. Although sampling programs may be robustly designed, sample size targets are often not met due to various factors such as the realities of field work and sample preparation, fishery regulations, spatial movement of fish relative to the sampling program, and selectivity of the fishery or survey gear. Our simulations can be used by agers and assessment scientists to guide the selection and interpretation of ageing error diagnostics given the longevity (short-, medium-, long-lived) of the fish of interest and sampling levels and patterns achieved. For example, one might anticipate a greater chance of failing to detect bias when present for a species with approximately five age classes if the trend in sample size decreases with age or is peaked and overall sample size is small (Fig. 3B and C).

ACV is often compared to an ad hoc threshold as a diagnostic tool for characterizing imprecision (Campana, 2001; McBride, 2015). We have demonstrated how this approach may not result in the detection of imprecision, when present, at low to moderate levels of random error (e.g., CV5 and CV10; Fig. 4). In particular, comparison of ACV with a threshold of seven or ten is less likely to identify imprecision when present if there are a small number of age classes and if the trend in sample size decreases with age (Fig. 4; Worthington et al., 1995). Improved detection of imprecision for longer-lived fish and the inability to detect imprecision with decreasing sample size at age is to be expected given we simulated ageing error that increased with age. For situations in which this is not the case, detection of imprecision using ACV ad hoc thresholds of seven and ten may be less reliable. In addition, we demonstrated how ACV is artificially inflated (Fig. 2B and C) in the presence of different types of bias (Campana, 2001). Given these challenges in interpreting ACV relative to an ad hoc threshold, we echo the recommendations of McBride (2015) in suggesting that precision metrics such as APE and ACV not be used as the sole or primary diagnostic of ageing error, but instead as a rough indicator of the presence of ageing error. Visualization of the data (e.g., age-bias plots) and tests for bias

A

Max Age	Precision Level	McNemar's Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	Evans & Hoenig Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	Bowker's Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20
5	CV5	0	1	5	0	0	0	5	5	6	0	1	5	0	0	0	3	5	4	0	0	18	0	0	0	0	3	57
	CV10	4	5	4	0	5	6	4	5	4	3	5	3	0	5	6	3	3	4	0	25	100	0	0	26	0	9	99
	CV15	4	4	7	3	5	5	5	5	6	4	3	6	3	5	4	3	3	3	0	44	100	0	3	83	0	18	100
20	CV5	4	6	4	0	0	2	6	6	5	3	3	4	0	0	2	4	4	3	0	8	100	0	0	0	0	0	46
	CV10	4	5	5	2	5	6	5	5	5	4	3	5	2	5	6	2	2	4	0	14	100	0	16	96	0	1	96
	CV15	4	5	6	5	4	4	4	5	5	2	3	4	5	4	3	4	4	3	0	20	100	1	56	100	0	2	100
50	CV5	5	5	5	4	5	5	5	7	5	3	5	4	4	4	3	5	4	3	0	2	100	0	1	50	0	0	35
	CV10	5	4	5	5	5	4	4	5	5	2	4	5	3	3	4	3	3	4	0	1	100	0	2	93	0	2	96
	CV15	5	6	5	5	6	5	5	5	6	2	3	3	3	3	4	3	4	4	0	4	100	0	4	100	0	3	100

B

Max Age	Bias Type	McNemar's Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	Evans & Hoenig Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	Bowker's Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20
5	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	-10%	71	100	100	2	97	100	100	100	100	71	100	100	2	97	100	100	100	100	14	100	100	0	54	100	100	100	100
	Both	100	100	100	100	100	100	2	2	7	100	100	100	100	100	100	11	97	100	100	100	100	100	100	100	13	100	100
20	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	-10%	100	100	100	46	100	100	100	100	100	100	100	100	46	100	100	100	100	100	100	100	100	6	100	100	100	100	100
	Both	3	23	90	100	100	100	100	100	100	66	100	100	100	100	100	100	100	100	99	100	100	100	100	100	100	100	100
50	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	-10%	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Both	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

C

Max Age	Bias Type	McNemar's Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	Evans & Hoenig Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	Bowker's Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20
5	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	18	100	100
	-10%	62	100	100	39	99	100	100	100	100	51	100	100	33	97	100	100	100	100	7	100	100	0	90	100	26	100	100
	Both	100	100	100	100	100	100	6	20	61	100	100	100	100	100	100	10	60	100	98	100	100	100	100	100	0	45	100
20	+1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	100	100	62	100	100	100	100	100	81	100	100
	-10%	100	100	100	69	100	100	100	100	100	99	100	100	63	100	100	100	100	100	4	100	100	25	100	100	7	100	100
	Both	25	88	100	100	100	100	12	50	99	51	100	100	100	100	100	100	100	100	8	100	100	97	100	100	100	100	100
50	+1	100	100	100	100	100	100	99	100	100	95	100	100	100	100	100	80	100	100	2	100	100	100	100	100	0	91	100
	-10%	100	100	100	95	100	100	100	100	100	100	100	100	98	100	100	100	100	100	0	100	100	12	100	100	3	100	100
	Both	83	100	100	100	100	100	100	100	100	97	100	100	100	100	100	97	100	100	0	100	100	88	100	100	0	100	100

Fig. 3. Summary of bias diagnosis test results for scenarios which simulated A) only random ageing error (imprecision), B) both random error (CV=5) and three types of bias, and C) both random error (CV=15) and three types of bias. Columns group results by 1) test of symmetry, 2) trends in sample size (uniform, decreasing, and peaked), and 3) a subset of low, medium, and high sample size multipliers of 1, 5, and 20 simulated; McN = McNemar's maximally pooled test, E&H = Evans & Hoenig's diagonally pooled test, and Bowk = Bowker's unpooled test. Rows group results by the number of age classes ("Max Age" of 5, 20, or 50). In A, precision levels indicate CVs used to generate random error of 5 (CV5), 10 (CV10), or 15 (CV15). In B and C, bias types include overestimation by on year (+1), 10% underestimation of true age (−10%), and both types (Both). Each box contains the percentage of 1000 runs for each scenario for which the p-value was less than 0.05. Shading represents the percentage of 1000 runs in each of the following categories: 100% (black), > 5 and < 100% (light gray), or ≤ 5% (white).

Max Age	Precision Level	ACV (7) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20	ACV (10) Uniform 1	5	20	Decreasing 1	5	20	Peaked 1	5	20
5	CV5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV10	3	0	0	0	0	0	85	99	100	0	0	0	0	0	0	0	0	0
	CV15	74	92	100	20	5	0	100	100	100	28	8	0	2	0	0	97	100	100
20	CV5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV10	51	50	52	6	0	0	73	94	100	0	0	0	0	0	0	0	0	0
	CV15	100	100	100	87	99	100	100	100	100	88	100	100	49	46	45	99	100	100
50	CV5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CV10	94	100	100	11	1	0	99	100	100	0	0	0	0	0	0	0	0	0
	CV15	100	100	100	100	100	100	100	100	100	100	100	100	65	79	94	100	100	100

Fig. 4. Summary of average coefficient of variation (ACV) results for an alternative simulation in which both ages in the pair were simulated with random error. Columns group results by 1) ACV, using thresholds for detecting imprecision of seven and ten as noted in parentheses, 2) trends in sample size (uniform, decreasing, and peaked), and 3) a subset of low, medium, and high sample size multipliers of 1, 5, and 20 simulated. Rows group results by the number of age classes ("Max Age" of 5, 20, or 50). Precision levels indicate CVs used to generate random error of 5 (CV5), 10 (CV10), or 15 (CV15). Each box contains the percentage of 1000 runs for each scenario that were greater than the ACV imprecision threshold. Shading represents the percentage of 1000 runs in each of the following categories: 100% (black), > 5 and < 100% (light gray), or ≤ 5% (white).

should always accompany the calculation of precision.

Our simulations demonstrate the relative frequency with which precision thresholds of seven and ten would lead to the identification of imprecision in a variety of situations. The selection of an appropriate threshold for a given ageing error study should ultimately depend on the maximum level of imprecision acceptable given the intended use of the data in analyses or modeling efforts. Thus, the appropriate threshold for a given set of data will be situation-specific (Campana et al., 1995; McBride, 2015). Although there is no definitive threshold that can be objectively adopted for all situations (Campana, 2001), readers may use our results to guide their interpretation of ACV relative to their chosen threshold given the potential influence of sample size trend and number of ages (Figs. S2–1–S2–9). Note that increasing sample size serves primarily to narrow the distribution of ACV simulated, but does not affect the median response relative to a given threshold (Figs. S2–1–S2–9).

Tests of symmetry used to detect bias should be interpreted in light of their expected performance in similar age class and sample size scenarios. This study demonstrates that tests of symmetry have the potential to perform poorly at both low and high sample sizes (Figs. 2 and 3). Intuitively, most scientists expect that tests of symmetry may not detect ageing error (i.e., Type II error; fail to detect bias when present) if the number of samples per age class is low because the test would have low power (Evans and Hoenig, 1998; McHugh, 2013). However, in the case of high sample size, tests of symmetry may also perform poorly by falsely detecting the presence of bias when it is not present (i.e., Type I error; falsely detecting bias when it is not present). We demonstrated that large sample size is not always a guarantee of good diagnostic performance given tests of symmetry can generate "false positive" tests for bias, particularly when using Bowker's unpooled approach (Figs. 3 and 5). Bowker's test is highly sensitive and can result in a positive bias test even if only one pair of cells in a contingency table is sufficiently large (McBride, 2015). Although large samples sizes such as those simulated in this study are not typically achievable for fishery-independent sampling programs due to resource limitations, they may pose a problem when interpreting fishery-dependent samples collected for several large and valuable fisheries such as Gulf Menhaden, Sablefish, and Atlantic Herring (see Introduction). One empirical example of this was the Atlantic menhaden (*Brevortia tyrannus*) ageing error study conducted by Schueller et al. (2021) in which McNemar's and Evans & Hoenig's test results for paired ages of Atlantic menhaden were not significant for annual and total samples, but Bowker's test indicated bias. Large sample sizes, which were in the hundreds per age class, were high enough to trigger what is most likely a false positive Bowker's test (28% of runs). Although the simulation study presented here identified some situations in which McNemar's and Evans & Hoenig's tests also generated false positives for bias, these tests did so far less often (<4% of runs) and at frequencies similar to what would be expected by chance given $\alpha = 0.05$. Note that Bowker's also performed poorly at low sample size in the presence of high amounts of random error (Fig. 3C). Given the superior performance of McNemar's and Evans & Hoenig's tests, these tests should be used to diagnose bias, and we strongly recommend that Bowker's test not be used. The use of Bowker's test in many circumstances could lead to unnecessary concern about ageing uncertainty and the possible rejection of unbiased ageing data at high sample sizes or the acceptance of biased data at low samples sizes, which could hinder the generation of informative science for management.

Interpretation of the overall impacts of trend in sample size was complicated by the interaction of trend with number of age classes and total sample size; however, several general patterns emerged. First, mean ACV was generally lower for scenarios with decreasing sample size with age, making identification of imprecision less likely compared with uniform and peaked scenarios (Fig. 2A and Fig. 4). Also, the peaked trend in sample size interacted with the increase in total error when bias was present, resulting in less severe ACV inflation (Fig. 2B and C). Finally, tests of symmetry failed to identify the presence of bias more often for scenarios with peaked trends in sample size than uniform or decreasing trends. Peaked trends in sample size are quite common because younger fish are not susceptible to the fishing or sampling gear and older fish are fewer in number and less likely to be encountered; thus, ageing studies for these fish may fail to identify the presence of bias more often than fish with differing sample size patterns with age.

We found that, when ageing error increases with age as simulated in our "− 10%" and "Both" ("−10%" and "+1″) scenarios (Fig. S1B), the resulting ageing error pattern is not reliably detected even by the best-performing diagnostic tests, McNemar's and Evans & Hoenig (Fig. 3). All tests of symmetry had trouble detecting the − 10% bias pattern, particularly at a lower number of age classes and sample size. This result could be an artifact of our ageing error simulation approach given estimated ages that resulted in non-integer ages were rounded before diagnostic tests were run.

If the intended use of age data is to inform age composition in a stock assessment model, then bias, not random error, is the most important problem to diagnose (Chang et al., 2019). Although the accuracy of stock assessment models is generally robust to random error, model performance is negatively impacted by biased age data (Fournier and Archibald, 1982). Thus, even though it has been suggested that target sample size should increase as ageing error increases (Richards et al., 1992), there may be little benefit to increased sampling in some

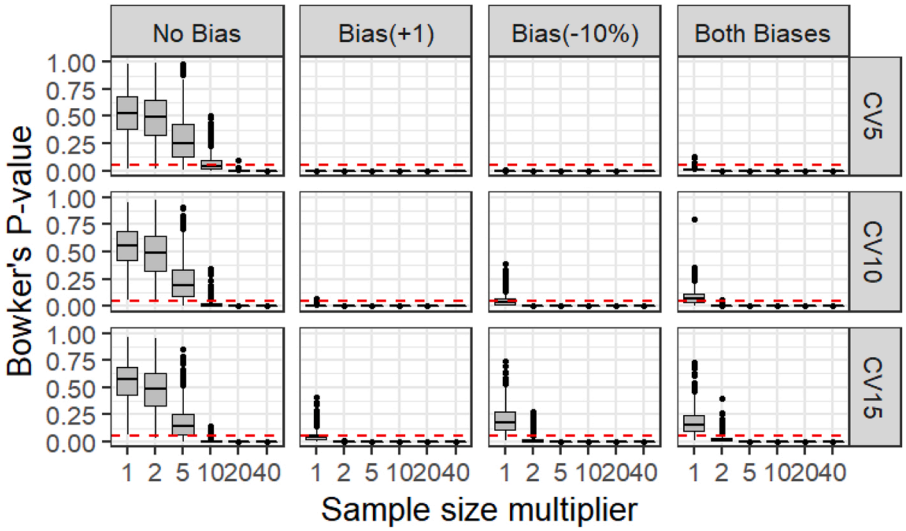


Fig. 5. Relationship between sample size (x-axis) and *P*-value for the Bowker's unpooled test of symmetry (y-axis) for paired comparisons between known and read ages simulated with random error (imprecision) and different types of bias (+1 year, −10%, Both Biases) for a fish that lives to age 20 and has no trend in sample size by age. Rows represent simulated imprecision levels generated using CVs of 5, 10, and 15 represented (CV5, CV10, and CV15, respectively). Columns represent alternative bias types. Red dashed line indicates $P = 0.05$.

situations, especially given that tests of symmetry perform reliably well at all but the lowest sample size categories in most of the age class and sample size/trend scenarios we simulated. Because decisions regarding sampling targets are often constrained by available resources, the goal of the ageing program should be carefully considered when evaluating tradeoffs between increased sample size and other aspects of fish stock monitoring and assessment. Note that, in our sample size scenarios, we do not make the distinction between how additional samples are collected (e.g., same or different trip), which could impact characterization of precision and the amount of additional independent information provided to the assessment model.

Our choice of how to simulate ageing error influenced interpretation of imprecision diagnostics. First, we chose to expand upon McBride's (2015) study and generate ageing error by pairing a known age with an estimated age to simulate an age validation process. Thus, our study simulated the use of ageing error diagnostics in the case where the reference collection is composed of ages known without error (e.g., tagging study-based reference collections). Because one of the simulated ages was known, our results do not represent the imprecision normally encountered in most ageing studies that characterize the differences among two or more estimated ages. The alternative simulation (Fig. 4) with paired estimated ages represents ACV levels for situations in which age estimates are compared across readers or laboratories. Both approaches highlighted the importance of accounting for number of age classes and trend in sample size when interpreting ACV levels relative to a threshold. Also, we chose to simulate scenarios that included the same number of samples across all ages (uniform) for comparison with scenarios in which sample size decreased with age or peaked at the middle age in order to highlight the impact of a trend in sample size on ageing error characterization. Although an assumption of complete uniformity in sample size by age is not realistic, some programs come close to achieving similar sample sizes across ages for fish with a small number of age classes; thus the results of our uniform sample size simulations may be most useful in the case of a well-sampled, shorter-lived fish.

Several important questions regarding the use and interpretation of ageing error diagnostics were beyond the scope of this paper, but would be important to pursue in future research. For example, this study was restricted to paired-age comparisons, but could be expanded to examine three or more reads from multiple readers or structures per fish; such a simulation study would allow for examination of other ageing error diagnostics such as sequential multinomial confidence intervals (Zar, 2013) and multivariate Euclidean distance-based techniques (Wakefield et al., 2016). Furthermore, identification of ageing error is an important first step, but quantifying that error and incorporating it into stock assessments (to the extent this is feasible in catch-at-age models) is also important (Clark, 2004; Punt et al., 2008; Richards et al., 1992). The impact of both random ageing error and bias on the accuracy of stock assessments is of utmost importance to sustainable fisheries management. Although the impact of ageing error on age-based assessments has been examined previously, this research question could be expanded to include a wider range of bias scenarios as well as age class and sample size scenarios.

Decisions regarding the utility of age estimates in ecological studies and stock assessments require a nuanced understanding of the way in which ageing error diagnostics are influenced by number of ages and sample size. This study should help facilitate communication between ageing experts and fisheries scientists who use age data when interpreting ageing error studies. Ageing error diagnostics are critically important for identifying within- and among-laboratory inconsistencies (Campana, 2001; Morison et al., 2005). When high imprecision or bias is identified, this study can be referenced to identify the potential impact of number of age classes and the sampling program under consideration. If substantial error is likely (i.e., diagnostics are reliable given the number of age classes and associated sample sizes), ageing experts can then work to pinpoint the problem and identify the solutions most likely to improve age estimates. Given sampling targets are often length-based, there will be differences between the general results from our simulation study and the actual sample sizes at age achieved in the field due to variability in the relationship between length and age. However, the overarching results of this simulation study can be used to more reliably interpret ageing error study results and the products generated (e.g., ageing error matrices), which are increasingly being used to inform stock assessment and management.

Funding sources

This project was supported by the Improve a Stock Assessment Program of the Office of Science and Technology, National Marine Fisheries Service (NA19OAR4320074), the National Science Foundation Science Center for Marine Fisheries (1266057), and through membership fees provided by the Science Center for Marine Fisheries Industry Advisory Board (8006298-03.02 UMCES).

CRediT authorship contribution statement

Geneviève Nesslage: Conceptualization, Methodology, Data curation, Formal Analysis, Funding acquisition, Writing – original draft, Software, Visualization. **Amy Schueller:** Funding acquisition, Conceptualization, Methodology, Visualization, Writing – review & editing, Software. **Amanda Rezek**: Funding acquisition, Visualization, Writing – review & editing. **Ray Mroch III:** Funding acquisition, Visualization, Writing – review & editing, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Richard McBride, Erik Williams, and two anonymous reviewers for helping to improve this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2022.106255.

References

Beamish, R., Fournier, D., 1981. A method for comparing the precision of a set of age determinations. Can. J. Fish. Aquat. Sci. 38, 982–983.
Beamish, R., McFarlane, G.A., 1995. A Discussion of the Importance of Aging Errors, and an Application to Walleye Pollock: The World's Largest Fishery. University of South Carolina Press, Columbia, SC, pp. 545–565.
Bowker, A.H., 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43, 572–574.
Cailliet, G.M., Smith, W.D., Mollet, H.F., Goldman, K.J., 2006. Age and growth studies of chondrichthyan fishes: the need for consistency in terminology, verification, validation, and growth function fitting. Environ. Biol. Fishes 77, 211–228.
Campana, S., 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. J. Fish. Biol. 59, 197–242.
Campana, S.E., Annand, M.C., McMillan, J.I., 1995. Graphical and statistical methods for determining the consistency of age determinations. Trans. Am. Fish. Soc. 124, 131–138.
Chang, W.Y., 1982. A statistical method for evaluating the reproducibility of age determination. Can. J. Fish. Aquat. Sci. 39, 1208–1210.
Chang, Y., Hsu, J., Shiao, J., Chang, S., 2019. Evaluation of the effects of otolith sampling strategies and ageing error on estimation of the age composition and growth curve for Pacific bluefin tuna Thunnus orientalis. Mar. Freshw. Res. 70, 1838–1849.
Clark, W.G., 2004. Nonparametric estimates of age misclassification from paired readings. Can. J. Fish. Aquat. Sci. 61, 1881–1889.
Dorval, E., McDaniel, J.D., Porzio, D.L., Felix-Uraga, R., Hodes, V., Rosenfield, S., 2013. Computing and selecting ageing errors to include in stock assessment models of Pacific sardine (Sardinops sagax). Calif. Coop. Ocean. Fish. Investig. Rep. 54, 192–204.

Elzey, S., Trull, K., Rogers, K., 2015. Massachusetts Division of Marine Fisheries Age and Growth Laboratory: Fish Aging Protocols. Massachuestts Division of Marine Fisheries, Gloucester, MA, USA.
Evans, G.T., Hoenig, J.M., 1998. Testing and viewing symmetry in contingency tables, with application to readers of fish ages. Biometrics 620–629.
Fournier, D., Archibald, C.P., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39, 1195–1207.
Goethel, D., Hanselman, D., Rodgveller, C., Fenske, K., Shotwell, S., Echave, K., Malecha, P. Siwicke, K., Lunsford, C., 2020. Assessment of the sablefish stock in Alaska, in: Stock assessment and fishery evaluation report for the groundfish resources of the GOA and BS/AI as projected for 2011. North Pacific Fishery Management Council, Anchorage, AK.
Henríquez, V., Licandeo, R., Cubillos, L.A., Cox, S.P., 2016. Interactions between ageing error and selectivity in statistical catch-at-age models: simulations and implications for assessment of the Chilean Patagonian toothfish fishery. ICES J. Mar. Sci. 73, 1074–1090.
Kimura, D.K., Anderl, D.M., 2005. Quality control of age data at the Alaska Fisheries Science Center. Mar. Freshw. Res. 56, 783–789.
Lai, H.-L., Gallucci, V.F., Gunderson, D.R., Donnelly, R.F., 1996. Age determination in fisheries: methods and applications to stock assessment. Stock Assess.: Quant. Methods Appl. small-Scale Fish. 82–170.
Liao, H., Jones, C., Gilmore, J., 2019. 2018 Final Report: Virginia and Chesapeake Bay Finfish Ageing and Population Analysis. Center for Quantitative Fisheries Ecology. Old Dominion University, Norfolk, VA.
Matta, M.E., Kimura, D.K., 2012. Age determination manual of the Alaska Fisheries Science Center Age and Growth Program. NOAA Alaska Fisheries Science Center, Seattle, WA.
Maunder, M.N., Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72, 7–18.
Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74.
McBride, R.S., 2015. Diagnosis of paired age agreement: a simulation of accuracy and precision effects. ICES J. Mar. Sci. 72, 2149–2167.
McHugh, M.L., 2013. The chi-square test of independence. Biochem. Med. 23, 143–149.
McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157.
Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.
Morison, A., Burnett, J., McCurdy, W., Moksness, E., 2005. Quality issues in the use of otoliths for fish age estimation. Mar. Freshw. Res. 56, 773–782.
Northeast Fisheries Science Center, 2012. 54th Northeast Regional Stock Assessment Workshop (54th SAW) Assessment Report. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 12–18; 600 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543–1026, or online at ⟨http://www.nefsc.noaa.gov/nefsc/publications/⟩.
Ogle, D.H., Wheeler, P., Dinno, A., 2021. FSA: Fisheries Stock Analysis. R package version 0.8.32.9000, https://github.com/droglenc/FSA.
Punt, A.E., Smith, D.C., KrusicGolub, K., Robertson, S., 2008. Quantifying age-reading error for use in fisheries stock assessments, with application to species in Australia's southern and eastern scalefish and shark fishery. Can. J. Fish. Aquat. Sci. 65, 1991–2005.
Quinn, T.J., Deriso, R.B., 1999. Quantitative fish dynamics. Oxford University Press, Oxford, p. 560.
R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL ⟨https://www.R-project.org/⟩.
Reeves, S.A., 2003. A simulation study of the implications of age-reading errors for stock assessment and management advice. ICES J. Mar. Sci. 60, 314–328.
Ricard, D., Minto, C., Jensen, O.P., Baum, J.K., 2012. Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish Fish 13, 380–398.
Richards, L.J., Schnute, J.T., Kronlund, A., Beamish, R.J., 1992. Statistical models for the analysis of ageing error. Can. J. Fish. Aquat. Sci. 49, 1801–1815.
Robillard, E., Reiss, C.S., Jones, C.M., 2009. Age-validation and growth of bluefish (Pomatomus saltatrix) along the East Coast of the United States. Fish. Res. 95, 65–75.
Schueller, A.M., Rezek, A., Mroch III, R.M., Fitzpatrick, E., Cheripka, A., 2021. Comparison of ages determined by using an Eberbach projector and a microscope to read scales from Atlantic menhaden (Brevoortia tyrannus) and Gulf menhaden (B. patronus). Fish. Bull. 119, 21–32.
SEDAR, 2018. Gulf Menhaden Stock Assessment Report available online at: https://sedarweb.org/sedar-63. SEDAR, North Charleston, SC, pp. 333.
Sutherland, S., 2020. How to use the Standard Precision Template. ⟨https://www.fisheries.noaa.gov/resource/tool-app/precision-templates-ageing⟩ (Accessed 2021).
Thorson, J., Stewart, I., Punt, A., 2012. nwfscAgeingError: a user interface in R for the Punt et al. (2008) method for calculating ageing error and imprecision. ⟨http://github.com/nwfsc-assess/nwfscAgeingError⟩.
AnonVitale, F., Worsøe Clausen, L., Ní Chonchúir, G. eds, 2019. Handbook of fish age estimation protocols and validation methods. ICES Cooperative Research Report No. 346. 180 pp. http://doi.org/10.17895/ices.pub.5221.
Wakefield, C.B., O'Malley, J.M., Williams, A.J., Taylor, B.M., Nichols, R.S., Halafihi, T., Humphreys Jr., R.L., Kaltavara, J., Nicol, S.J., Newman, S.J., 2016. Ageing bias and precision for deep-water snappers: evaluating nascent otolith preparation methods using novel multivariate comparisons among readers and growth parameter estimates. ICES J. Mar. Sci. 74, 193–203.
Worthington, D., Fowler, A., Doherty, P., 1995. Determining the most efficient method of age determination for estimating the age structure of a fish population. Can. J. Fish. Aquat. Sci. 52, 2320–2326.
Zar, J.H., 2013. Biostatistical Analysis, fifth ed. Pearson Prentice-Hall, Upper Saddle River, NJ, p. 944.