Quality Assurance and Quality Control Estimates for the Production Ageing of Northwest Atlantic Species

Rationale

It is important to ensure consistency in fish ages generated by a production ageing laboratory. There are three components to measuring this consistency, namely accuracy, intra-reader precision, and inter-reader precision. Accuracy is determined by how closely the ages generated in production ageing are to the known ages for a set of fish; this is a measure of whether the age reader applies ageing criteria correctly. Intra-reader precision is determined by how reliably an age reader will assign the same age to an individual fish; this is a measure of how consistently the ageing criteria are applied from day to day. Finally, inter-reader precision tests whether fish ages are comparable between different people; it is measured by two (or more) age readers examining the same set of fish independently. For all three components, age is determined multiple times for each fish, and a comparison of the resulting ages determines the level of consistency. These aspects of consistency may change over time or between age readers, so it is necessary to measure them regularly throughout the production ageing process.

The three components affect the production age data in different ways, but all may introduce errors into the data. Measurement of ageing error has two primary aspects: quantification of variability, and detection of systematic bias. Any significant ageing bias indicates that ageing criteria are not being applied properly, whether it occurs within an intra-reader precision test (indicative of a drift in how the person applies ageing criteria) or in an accuracy test (indicating that the person is applying incorrect ageing criteria). Either case implies that the most recent production ages may be inconsistent with past years' data. Variability in intra-reader precision levels will introduce random errors, and may reduce the apparent abundance of strong year-classes while making weak year-classes appear more abundant. Within accuracy tests, high variability indicates that ageing criteria are not being applied consistently. Finally, if two readers differ in their age determinations, it becomes more difficult to utilize data from both of them in one stock assessment.

Providing these measures of consistency allows assessment scientists to consider these sources of variability within stock assessments. These measures are regularly estimated within the Fishery Biology Program at the Northeast Fisheries Science Center

(NEFSC). Accuracy tests are conducted for those species that have reference collections already assembled. Intra-reader precision tests are conducted on each set of production ages generated. A test of inter-reader precision is completed when an inter-laboratory exchange is conducted or a change in age reader occurs (due to temporary substitution or when training a permanent replacement). This website is an effort to make the results of those tests easily available to assessment scientists and other interested parties.

Methods

All production ageing at the Fishery Biology Program at the NEFSC follows established ageing methods, as described in Penttila and Dery (1988). Tests of the various aspects of ageing consistency are regularly conducted as described below. In all tests, age readers have knowledge of the data normally available during production ageing (i.e. fish length, date captured, and area captured), but do not have knowledge of previous ages given to the fish. If two ages (e.g., test age and production age) are not assigned to a given fish for any reason, that fish is excluded from calculation of statistical measures.

Types of Test

Accuracy

For each accuracy test, age readers are asked to re-age a random subset (N = 50–100 fish) of the reference collection. Most tests are conducted after the completion of production ageing; for haddock, tests are usually conducted both before and after production ageing.

A prerequisite to conducting these tests is the establishment of a reference collection, composed of a few hundred fish of known age. However, it is very difficult to obtain a large sample of fish for which the ages are definitively known; therefore, the NEFSC ageing laboratory has selected samples which have been aged by multiple experienced age readers and for which a consensus age has been agreed upon (Silva *et al.* 2004). For cod and haddock, samples have been assembled from past inter-laboratory exchanges with Canadian age readers; therefore, these reference collections only include fish from the Georges Bank stock. In the case of yellowtail flounder, however, samples from various stocks were chosen, distributed to four age readers experienced in ageing this species, and only fish for which these readers agreed on the age remained in the collection. Reference collections for other species will be assembled in upcoming years.

Calibration

For species without established reference collections, a calibration test may be done. This consists of re-ageing a representative subsample of fish from a previous year, before the current year's production is begun. This determines whether the age reader has a sufficient precision level to generate reliable ages, but does not test the accuracy of these ages.

Intra-reader precision

(listed as Precision tests)

After the completion of production ageing for a given set of samples (i.e., a specific survey, quarter, or year), the age reader conducts a precision test on a representative subsample (usually 50–100 fish) taken from that sample set. Subsamples are randomly selected, but include the range of lengths and sampling locations in the production age sample. Stock management areas are combined together in these tests, except when production ageing for each stock area occurs at different times. Although test ages may differ from production ages, no effort is made to improve results by further examination of samples, nor are production ages revised after tests are conducted.

Inter-reader precision

(listed as Precision or Exchange)

Precision tests between two readers are conducted less frequently, when a change in age reader occurs either due to temporary substitution or when a new age reader has been trained. They are structured similarly to intra-reader precision tests: one reader first ages all the samples (perhaps while doing production ageing of the fish), and a second reader later re-ages a portion (or all) of the sample set. Results of these tests are presented in terms of one of the reader's ages. In cases where one reader has been training the other to age a species, the trainee's ages are presented in terms of the established reader's ages. In other cases, either set of ages may be presented on the x-axis, and no assumption is made as to which person's ages are more reliable.

One specific type of inter-reader test is the interlaboratory exchange. Such tests are annually conducted for cod and haddock in cooperation with Canada's Department of Fisheries and Oceans (DFO). The exchanges consist of each laboratory shipping otolith samples to the other laboratory, and the alternate age reader determining the ages of the samples.

Historically, this ‘two-reader’ approach was used within the Fishery Biology Program to ensure quality control. The primary age reader for a given species would examine all the samples during production ageing, and then the second age reader would review 5–10% of the samples. This tested whether the primary reader had applied ageing criteria in the same way as the second reader.

Data Presentation

For each species, a table is given summarizing the results of all tests which have been conducted. Within this table, the source of samples for each test is listed first, and is linked to more detailed results for each test. If the test samples were from a specific part of the species' range, the stock area is shown; if no stock area is listed, the test was for all management areas combined. Detailed results for each test include an agreement plot, an age-frequency table, and a summary of the test results for each production (or reference) age, in addition to the measures shown in the species table.

Visual inspection of the agreement plots and age-frequency tables will reveal the presence or absence of bias, though these are not quantitative. Tests of symmetry can be useful in quantifying bias when the sample size is large and variability is high.

Variability is measured via both percent agreement and the mean coefficient of variation (CV). These measures are inflated when a bias is present, and thus will not accurately reflect variability if there is a bias. Variability levels are related to various factors inherent in the samples, including the fish species, the age reader’s experience, and the structure used for age determination. Some species/structures are easier to age than others.

Statistical Measures

The following measures are used to characterize the results of tests of ageing consistency at the Fishery Biology Program at the Northeast Fisheries Science Center:

Coefficient of Variation (CV)

The mean coefficient of variation (CV, Campana *et al*. 1995, Chang 1982) is a relatively robust approach to quantifying agreement in fish ages. It yields results which are easier to compare between species and structures. Also, the contribution each fish makes to the CV is relative to the average age assigned to that fish; i.e., a 2-year error in ageing a young fish would increase the measure more than would a 2-year error in an older fish, as the percentage change in age is greater for younger ages.

The CV is based on the differences between the mean age and each given age for each fish, and then these values are averaged over the entire sample set. When two ages are assigned to each fish, the CV is calculated as follows:

$$CV = 100\% \times \frac{1}{N} \sum_{j=1}^{N} \frac{\sqrt{\sum_{i=1}^{2} (X_{ij} - X_j)^2}}{X_j}$$

where X_{ij} is the *i*th age for the *j*th fish, X_j is the mean age of the *j*th fish, and N is the sample size.

Campana (2001) indicates that many ageing laboratories around the world view CVs under 5% to be acceptable among species of moderate longevity and ageing complexity. His description applies to most of the species considered here.

Percent Agreement

The Fishery Biology Program has used this measure since the group's inception, and considers levels of over 80% to be adequate. It is calculated based on the percentage of ages agreed upon relative to the total number aged:

$$\text{Percent Agreement} = 100 \times \frac{\text{Number of agreements}}{N}$$

For this measure, an error in ageing a young fish changes the measure by the same amount as would a similar error for an old fish. Therefore, this statistic is harder to compare between samples sets with different age distributions or across species.

Symmetry Tests

NOTE: At the beginning of 2022, it was decided to switch from the Bowker's test (Bowker, 1948) to the Evans & Hoenig test (Evans & Hoenig, 2015), based on simulation studies by Nesslage et al. (2022) and McBride (2015) indicating that the Evans & Hoenig test is less prone to false positive results than the Bowker's test. All symmetry tests conducted after 1/1/2022 will employ an Evans & Hoenig test; tests conducted prior to that will not be changed.

A symmetry test (Hoenig *et al.* 1995) may be used to test for any departure from symmetry, i. e. bias, within the age-frequency table. However, such a test has low sensitivity when few disagreements exist, so this test was not applied to cases where the percent agreement was 90% or above. Also, tests of symmetry are not conducted for accuracy tests, as the error is assumed to be entirely within the test age and therefore would be visible in the agreement plot and the age-frequency table.

The Evans & Hoenig test is pooled along the diagonal. Where ages differ from one another, it compares values on the age-frequency table which have the same absolute difference in age, such as the paired ages (3,4), (4,3), and (2,1), (1,2). This test statistic is calculated as a chi-square variable, as follows:

$$\chi^2 = \sum_{p=1}^{m-1} \frac{\left(\sum_{j=1}^{m-p} \left(n_{p+j,j} - n_{j,p+j} \right) \right)^2}{\sum_{j=1}^{m-p} \left(n_{p+j,j} + n_{j,p+j} \right)}$$

where m is the maximum age in the data set, p is the difference between age readings, and $n_{p+j,j}$ is the number of fish in row *p+j* and column *j* (Evans & Hoenig, 2015). The value of the degrees of freedom is equal to the maximum difference in the age values.

Bowker's Test of Symmetry

(This section applies only to symmetry tests conducted before 1/1/2022)

The Bowker's test (Hoenig *et al.* 1995, Bowker 1948) does not pool together cells on the age-frequency table. It compares values which represent symmetric errors, such as the paired ages (3,4) and (4,3). This test statistic is calculated as a chi-square variable, as follows:

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

where m is the maximum age in the data set, and n_{ij} is the number of fish in the *i*th row and *j*th column (Hoenig *et al.* 1995, Bowker 1948). The value of the degrees of freedom is equal to the number of non-zero n_{ij}-n_{ji} comparisons in this calculation, to a maximum of m(m-1)/2.

Agreement Plot

The agreement plot graphically shows all age pairs in each test, usually with the production (or reference) ages on the x-axis. Data are jittered so as to improve visibility of overlapping data points. Jittering was accomplished by adding a random number in the range (-0.1, 0.1) to each age within the test. Zero ages were jittered in the range (0.0, 0.1). While not all points may be visible, the exact counts of age pairs may be seen in the age-frequency table below. The diagonal line indicates 1:1 agreement; ideally, all age pairs should fall along this line. This format is similar to that used by Robillard *et al.* (2009).

A common assumption in statistical presentation is that the x-axis portrays 'better' data than the y-axis. This is a disadvantage of the age-bias plot, and why we have opted to use agreement plots rather than the more typical age-bias plot. We aim to portray paired ages as equally likely within most tests, with neither set of ages expected to be more reliable. While the agreement plot is not perfect, it should be less prone to misinterpretation than the age-bias plot.

The only tests in which one set of ages is expected to be more reliable are (a) accuracy tests, where the reference age has been reviewed & agreed upon by multiple age readers, and (b) training situations, where one person is being trained by a more experienced person and inter-reader precision tests are used to measure the trainee's progress.

Age-Frequency Table

The age-frequency matrix shows the numbers of samples at each age for both the production (or reference) age across the top, and the test age on the left. The grey boxes along the main diagonal of the matrix indicate the number of samples for which both ages are in agreement; fewer samples falling outside these boxes indicate better consistency. Numbers above this diagonal indicate fish which were given a lower age during the test, while numbers below this were given a higher test age; greater distance from the main diagonal indicates a greater difference between the two ages. Totals (at the right & bottom) indicate the age distribution within the test for both set of ages.

When a test compares ages between two readers, one reader's ages are listed across the top; the other is on the left. No assumption is made in these tests as to which reader is expected to be more accurate or precise, except when one reader is listed as a trainee.

Results Summary

This table shows a breakdown of the test results for each production (or reference) age. It gives the total number at each age, the number agreed upon during the test, the percentage of agreements at that age, and the average test age. The number of samples agreed upon is the same as in the main diagonal of the age-frequency table. Again, for inter-reader precision tests, one person's age is chosen to be the basis for the other's results; aside from training exercises, this is not intended to indicate that either set of ages is expected to be more reliable.

References

Bowker AH. 1948. A test for symmetry in contingency tables. J. Am. Statistical Assoc. 43:572-574.

Campana SE. 2001. Accuracy, precision, and quality control in age determination, including a review of the use and abuse of age validation methods. J. Fish Bio. 59:197-242.

Campana SE, Annand MC, McMillan JI. 1995. Graphical and statistical methods for determining the consistency of age determinations. Trans. Am. Fish. Soc. 124:131-138.

Chang WYB. 1982. A statistical method for evaluating the reproducibility of age determination. Can. J. Fish. Aquat. Sci. 39:1208-1210. (Available at **https://doi.org/10.1139/f82-158**)

Evans GT, Hoenig JM. 1998. Testing and viewing symmetry in contingency tables, with application to readers of fish ages. Biometrics 54: 620-629. (Available at **https://doi.org/10.2307/3109768**)

McBride RS. 2015. Diagnosis of paired age agreement: a simulation of accuracy and precision effects. ICES J. Mar. Sci. 72: 2149–2167. (Available at **http://refhub.elsevier.com/S0165-7836(22)00032-7/sbref18**)

Nesslage G, Schueller AM, Rezek AR, Mroch RM. 2022. Influence of sample size and number of age classes on characterization of ageing error in paired-age comparisons. Fish. Res. 249 Article 106236. (Available at **https://doi.org/10.1016/j.fishres.2022.106255**)

Penttila J, Dery LM. 1988. Age determination methods for northwest Atlantic species. NOAA Tech. Rep. NMFS-72; 135 p. (Available at **https://www.fisheries.noaa.gov/resource/document/age-determination-methods-northwest-atlantic-species**)

Robillard E, Reiss CS, Jones CM. 2009. Age-validation and growth of bluefish (*Pomatomus saltatrix*) along the East Coast of the United States. Fish. Res. 95:65-75. (Available at **http://dx.doi.org/10.1016/j.fishres.2008.07.012**)

Silva V, Munroe N, Pregracke SE, Burnett J. 2004. Age structure reference collections: the importance of being earnest. *In* Johnson DL, Finneran TW, Phelan BA, Deshpande AD, Noonan CL, Fromm S, Dowds DM, compilers. Current fisheries research and future ecosystems science in the Northeast Center: collected abstracts of the Northeast Fisheries Science Center's Eighth Science Symposium, Atlantic City, New Jersey, February 3-5, 2004. Northeast Fish. Sci. Cent. Ref. Doc. 04-01; p. 60.