ICES WKSABCAL REPORT 2014

ICES ACOM COMMITTEE

ICES CM 2014/ACOM:35

REF. PGCCDBS

Report of the Workshop on Statistical Analysis of Biological Calibration Studies (WKSABCAL)

13–18 October 2014

Lisbon, Portugal

ICES
International Council for the Exploration of the Sea
CIEM
Conseil International pour l'Exploration de la Mer

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H. C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

Recommended format for purposes of citation:

ICES. 2014. Report of the Workshop on Statistical Analysis of Biological Calibration Studies (WKSABCAL), 13-18 October 2014, Lisbon, Portugal. ICES CM 2014/ACOM:35. 82 pp.

The document is a report of an Expert Group under the auspices of the International Council for the Exploration of the Sea and does not necessarily represent the views of the Council.

Contents

Executive summary

The Workshop on Statistical Analysis of Biological Calibration Studies (WKSABCAL) met in Lisbon 13-18 October 2014 to review applied statistical methods for analysing reader agreement on fish age estimations, in the light of, both, what is warranted from the data suppliers (the readers) and what is required by the data receivers (the stock assessors). The aim of the workshop was to bridge between the age and maturity-calibration workshops and the stock assessment working groups facilitating a full use of the results and considerations from calibration workshops.

The group reviewed a large number of past workshops and exchanges as well as available literature to outline state-of-the-art of statistical methods for analysing reader agreement. Through discussions a selection WKSABCAL recommended the following methods/analysis to be run by age calibration workshops:

- To access bias
    - ABP - Age-bias plot
    - TS - Tests of symmetry
- To access precision
    - APE - Average Percentage Error
    - CV - Coefficient of Variation
- As diagnostics for problems found by the previous analysis
    - Analysis of otolith increments, both through image layers and statistically
- As output to stock assessment groups
    - AREM - Age Readings Error Matrix

All the methods were tested on known-dataset to evaluate their performance. The available software able to perform such analysis was evaluated and suggestions for R-packages were given. In this relation, the WKSABCAL outlined potential additions to the prevailing web-application interface for calibration studies (WebGR), which the group highly recommends implemented to facilitate the operational outcomes of calibration workshops.

The group discussed the range of potential recipients of outcomes from calibration workshops and suggested ways to reach the various groups.

WKSABCAL authored a Chapter for the ICES Cooperative Research Report on State-of-the-art of Age Reading, Calibration and Validation.

1 Introduction

The Workshop on Statistical Analysis of Biological Calibration Studies (WKSABCAL) was proposed by the Planning Group on Commercial Catch, Discards and Biological Sampling (PGCCDBS) with the purpose of reviewing applied statistical methods for analysing reader agreement on fish age estimations, in the light of, both, what is warranted from the data suppliers (the readers) and what is required by the data receivers (the stock assessors). The aim of the workshop was to bridge between the age and maturity-calibration workshops and the stock assessment working groups facilitating a full use of the results and considerations from calibration workshops.

WKSABCAL, chaired by Lotte Worsøe Clausen (DTU AQUA) and Ernesto Jardim (JRC) met in Lisbon 13-17 October 2014 to:

a) compile statistical methods for analysing reader agreement;
b) identify the strengths and weaknesses of each method for fisheries calibration studies;
c) review existing software for analysing calibration workshop data;
d) define data summaries and analysis outputs required by calibration workshop participants and as stock assessment input;
e) draft a review paper based on workshop presentations, discussions and results.

The ToR's aimed at promoting the inclusion into stock assessments of the knowledge of the dynamic biological parameters (e.g. age, growth, maturation) of fish stocks. Inaccurate age determinations are widespread and negatively affect the accuracy of population dynamics studies and stock assessment outcomes. There are numerous cases in which ageing errors contributed to the overexploitation of a population or species (Campana, 2001). Underestimation of age results in overly optimistic estimates of growth and mortality rates and overestimation of age results in underestimation of growth. Thus such errors around the age-estimations need to be accounted for when conducting stock assessment (Punt *et al.*, 2008). Likewise, errors in maturity staging will cause erroneous maturity ogives, which in turn will impair the estimation of spawning-stock biomass and stock–recruitment relationships, and ultimately biomass reference points.

A large proportion of the calibration workshops dealing with age and maturity estimation, are held under the auspices of ICES. Moving beyond precision is increasingly common in these calibration workshops and creating outputs better tailored to the inputs of stock assessment models, would greatly improve the uptake of the results by assessment working groups.

The workshop looked at the output statistics in a functional view; what can outputs from calibration workshops be used for and by which audience? WKSABCAL specified four main purposes that outputs from a calibration workshop should have:

- Test if readers have the same interpretation criterion about the structure;
- Identify if there are differences between readers;
- Ascertain if readers achieve the opinion the same way?
- Determine the right class of the structure;
- Propagate reading uncertainty into stock assessment and management advice.

These outputs range from being visual, descriptive, analytical tests to predictive. The aim of WKSABCAL was then to identify operational outputs from calibration workshops which may fulfil those criteria.

1.1 Report Contents

The report is built following the ToRs merging ToRs a) and b) given their close link. The ToRs are dealt with in separate sections, answering the tasks and is followed by a final section containing conclusions and recommendations for further actions. ToR e) is treated as a chapter for an ICES CRR and is shown in Annex 5.

Table 1.1 ToR's and Section numbers

Term of Reference	Addressed in Section
Compile statistical methods for analysing reader agreement (ToR a)	Section 2
Identify the strengths and weaknesses of each method for fisheries calibration studies (ToR b)	Section 2
Review existing software for analysing calibration workshop data (ToR c)	Section 3
Define data summaries and analysis outputs required by calibration workshop partici-pants and as stock assessment inpu (ToR d)	Section 4
Draft a review paper based on workshop presentations, discussions and results (ToR e)	Annex 5

1.2 General considerations related to output from calibration exercises

The term "precision" is used to describe "agreement" or variability between readings/annotations of the same specimen by the same or different readers. The term "accuracy" is reserved to describe a comparison of ages or stages of maturity generated by readers with the true age or maturity stage for specimens.

When analysing outputs from age-calibration workshops it is imperative to acknowledge that in case there isn't known-age material to calibrate the readers' assessment of ages, then no true bias can be estimated. In the absence of a known-age reference collection, ageing consistency is the best that can be achieved (Campana *et al.*, 1995). Nevertheless, bias is often reported from age-calibration workshops, which shouldn't be interpreted as real bias, when validated ages are not available. Thus bias in this respect is more an expression of the 'skewness' of data around a modal or likely value.

Maturity staging workshops more frequently have true-stages available for calibration, given the option for using histologically validated samples for the calibration exercise. For stock assessment purposes, the main interest lies in the assessment of the maturity status of the individuals (mature/immature), which are used to compute maturity ogives or other maturation models. This situation makes it simpler to address the impact on stock assessment results of errors identifying mature/immature individuals. But it requires development of alternative methods, once that the common methods used for age readings can't be directly applied to maturity calibrations.

The reports from calibration workshops generally give very thorough results, which facilitate getting a common interpretation of the age structures of the otoliths of a given species. The dissemination of the results supports the judgement, by stock assessment scientists, of the quality of the age distributions used in the assessment.

However, considering the very limited inclusion of workshop’s results into stock assessments, the questions are whether the right audience has been reached by these reports, with appropriate data formats, and if the stock assessment models are prepared to include them. The discussions and conclusions in WKSABCAL specifically targeted such intentions.

2 ToR a) and ToR b) Compile statistical methods for analysing reader agreement; identify the strengths and weaknesses of each method for fisheries calibration

2.1 ToR a) Compilation of statistical methods for analysing reader agreement

Prior to the meeting the chairs of ICES age reading workshops were asked to report on the applied methods and their views upon these. The responses from 18 workshops are listed in Table 2.1.1. As suggested by PGCCDBS guidelines for calibration workshops (ICES 2011a), most workshops had used the Guus Eltink calibration sheet (Eltkin, 2000). Only three workshops deviated from the Guus Eltink approach.

WKARMAC (ICES, 2010) also used OMAP v1.3 to assess the level of agreement between readers and labs, analyse differences in age reading interpretation of otolith spatial patterns, and explore the use of metric measurements of otolith structures to minimize divergence in age estimation.

WKARGH (ICES 2011b) used age bias plots combined with the coefficient of variation.

WKSIBCA (ICES 2014) used mixed effects models for pairwise comparisons between readers.

Table 2.1.1. Methods used by ICES age reading workshop between 2005 and 2014. The list is incomplete because not all workshop chairs responded to the request.

	Workshop name	Year	Method	Comments	Software
1	Blue Whiting Otolith Ageing Workshop	2005	Guus Eltink calibration sheet	See comment from WKARMAC	Excel
2	Sandeel Otolith Ageing Workshop	2006	Guus Eltink calibration sheet	See comment from WKARMAC	Excel
3	WKARRG	2007	CV and % agreement	The spreadsheet (Eltink, 2000) was used according to the instructions contained in Guidelines and Tools for Age Reading Comparisons by Eltink *et al.*, (2000). Modal ages were calculated for each otolith, with percentage agreement, mean age and precision coefficient of variation.	Excel
4	WKARFLO	2007, 2008		The results of the age determination were analysed using the spreadsheet of Eltink *et al.*, (2000) modified by Mark Etherton and the SPSS 15.0 statistical package. Comparisons of Modal age between readers to detect bias, comparisons of CV for modal age between different age reading methods.	
5	WKADR	2008	average percent error (APE), CVs, percent agreement	APE (in contrast to percent agreement) makes most sense for long-lived fish such as redfish (*Sebastes* spp.), as it takes into account the longevity in error estimation, i.e. one year deviation in age reading for redfish does not have the same effect on age reading error in stock assessment as it has for e.g. Baltic cod.	Excel (Eltink spreadsheet)
6	WKARNSC	2008	Guus Eltink calibration sheet	See answer for WKARMAC	Excel
7	WKACM	2009	CV and % agreement	The spreadsheet (Eltink, 2000) was used. Modal ages were calculated for each otolith, with percentage agreement, mean age and precision coefficient of variation	Excel

	Workshop name	Year	Method	Comments	Software
8	WKARA	2009	The Guus Eltink spreadsheet'	Eltink spreadsheet has been very valuable for the development of age calibrations. However, being an Excel spreadsheet, the tool is rather sensitive to typological errors and destruction of the formulas in the actual cells and thus the spreadsheet is prone to errors. Additionally, the spreadsheet does require the user to be alert to how the readers are ranked in the columns, as the weighing of the individual readers depends on the column number. This and other inherited features of the spreadsheet make it prone to errors. The copying and pasting of data into the Excel spreadsheet can lead to transcription errors, and the formulas used in the spreadsheet can be accidentally modified, affect-ing the results of the analysis.	Paint Shop Pro for Exchanges of images
9	WKARMAC	2010	Guus Eltink calibration sheet	Not prone to errors/bugs; sensitive to the order of readers; returns only percent agreement and bias estimates - no information of where and why disagreements occur	Excel
10	WKARMAC	2010	OMAP v. 1.3	R-script developed for the workshop by T. Jansen: Assess the level of agreement between readers and labs Analyse differences in age reading interpretation of otolith spatial patterns Explore the usage of metric measurements of otolith structures as a solution to minimize divergence in age estimation	
11	WKARP	2010	CV	Metrics used to compare readers: percentage agreement, bias, CV and APE. Only CV and APE considered to be statistical methods. Two pages filled in for WKARP 2010, one for each of the 2 stats methods (as this is what you instructed?).	Excel (Guus Eltink spreadsheet)
12	WKARP	2010	APE	Metrics used to compare readers: percentage agreement, bias, CV and APE. Only CV and APE considered to be statistical methods. Two pages filled in for WKARP 2010, one for each of the 2 stats methods (as this is what you instructed?).	Excel
13	WKARGH	2011	Coefficient of Variation combined with Age Bias Plots	These methods were not applied at the meeting but in a working paper presented and discussed at the meeting. This is a simple yet effective means of assessing differences between readers and/or age determination methods and was recommended to us by Dr Steven Campana, one of the external advisors invited to our meeting.	

	Workshop name	Year	Method	Comments	Software
14	WKARAS 2011 - Workshop on Age Reading of European Atlantic Sardine	2011	"Workbook Age Reading comparisons" of Eltink (2000) following the recommendations of the "Guidelines and tools for age reading com-parisons" (Eltink *et al.*, 2000)		"Workbook Age Reading comparisons" of Eltink (2000)
15	otolith exchange of European hake	2011	Average percent age error (APE), Beamish and Fournier (1981), Coefficient of Variation (CV), Percentage of Agreement		http://webgr.azti.es/
16	WKACM2	2012	CV and % agreement	The spreadsheet (Eltink, 2000) was used according to the instructions contained in Guidelines and Tools for Age Reading Comparisons by Eltink *et al.*, (2000). Modal ages were calculated for each otolith, with percentage agreement, mean age and precision coefficient of variation.	Excel
17	WKMIAS	2013	CV and APE	As reference age, we used the mean age rather than the modal age, due to the large number of ages obtained in the daily age determination. Although the mean age estimate is not an indicator for the reliability of ageing structure, it may provide useful information regarding over- or underestimation of age by a structure irrespective of fish size class.	
18	otolith exchange of Western and Eastern Baltic cod as preparatory work for WKSIBCA	2014	Pairwise comparisons between readers using Linear Mixed Effects Models with the "alldistances" output by WebGR	Box plot and Excel diagnostics; the analysis examined whether there are consistent differences in growth curves estimated by the different readers.	WebGR; statistical software tool

In addition, 31 ICES reports published between 1992 and 2012 on age reading and calibration exercises were reviewed. Most of them used methods to compare age readings such as Percentage of agreement (PA), average percent error (APE), and Coefficient of variation (CV). After 2000, the majority of reports used the Guus Eltink spreadsheet, i.e. the Workbook Age Reading comparisons of Eltink (2000) and Guidelines and tools for age reading comparisons (Eltink *et al.*, 2000). The Age bias plot was the most commonly used graphical representation for comparisons between age readers. The complete table of ICES age calibration workshops reviewed during the meeting for the use of different analytical methods, software and diagnostics is given in Annex 4.

2.1 ToR b) Identify the strengths and weaknesses of each method for fisheries calibration studies

Of the 17 methods reported in the literature (Table 2.2.1), all can be classified as one of the following: a) identification of bias between age readers or a reference collection; b) estimate of precision among or within age readers; c) diagnostic of age reading differences; and d) preparation of ageing error matrix for use in stock assessment models.

Table 2.2.1. Methods which can be used to evaluate age calibration studies; characteristics of the methods and strengths and weaknesses are given. Colour code highlights methods considered key products of an age calibration study: assess bias (YELLOW), assess precision (ORANGE), diagnostics (RED), output for assessment (GREEN)

	Method	Descriptive statistics	Statistical test	One single number	Visual method	Model–based approach	Precision	Bias	Data requirements	Diagnostics	Strength	Weakness	Observations
1	ABP - Age-bias plot			1	1			1	age readings		Easily interpreted	visual, not a statistical test	
2	TS - Tests of symmetry		1	1				1	age readings			statistical test not picking up non-monotonic aging problems	
3	PTT - Paired t-test		1	1				1	age readings		Easily interpreted		parametric test
4	WPRT - Wilcoxon paired rank test		1	1				1	age readings		Easily interpreted		non-parametric test
5	APE - Average Percentage Error	1		1			1		age readings		Easily interpreted		sensitive to outliers
6	CV - Coefficient of Variation	1		1			1		age readings		Easily interpreted		sensitive to outliers
7	MSD - Mean Square Deviation	1		1			1		age readings		Easily interpreted		
8	CCC - Concordance Correlation Coefficient	1		1			1		age readings		Easily interpreted		

	Method	Descriptive statistics	Statistical test	One single number	Visual method	Model-based approach	Precision	Bias	Data requirements	Diagnostics	Strength	Weakness	Observations
9	TDI - Total Deviation Index	1		1			1		age readings		Easily interpreted		
10	MAD - Modal Age Difference	1		1			1		age readings		Easily interpreted		
11	PA - Percentage Agreement	1		1			1		age readings		Easily interpreted	poor because it is sensitive to range of ages used in the analysis	
12	Rho - Average Spearman's Rho		1	1			1		age readings		Easily interpreted		just a correlation coefficient
13	W - Kendal's Coefficient of Concordance		1	1			1		age readings		Easily interpreted		
14	Tau - Average Tau		1	1			1		age readings		Easily interpreted		
15	MAOI - Model Analysis of Otolith Increments (e.g. with mixed effects models)		1			1		1	each ring of each otoliths marked by readers are required (WebGR)	1		complex output	

	Method	Descriptive statistics	Statistical test	One single number	Visual method	Model-based approach	Precision	Bias	Data requirements	Diagnostics	Strength	Weakness	Observations
16	VAOI - Visual Analysis of Otolith Increments (e.g. using photoshop)	1			1				each ring of each otoliths marked by readers are required (WebGR)	1			use layers in photoshop to visualize age readings and the mixed effects model results
17	AOI - Analysis of Otolith Increments (e.g. using simple statistics)	1		1			1	1	each ring of each otoliths marked by readers are required (WebGR)	1	summary across readers and otoliths		Requires further development
18	AREM - Age Readings Error Matrix					1	1		age readings				bridge towards stock assessment; matrix can use observed or modelled data

Identification of bias is one of the most important products of an age calibration exercise, since it indicates that the age readers are interpreting the growth increments differently. Unless one of the age readings is based on a validated reference collection, it can be difficult to determine which of the age readings (if either) is more accurate. Method (1), the age bias plot, is a widely used method for visually identifying bias. Methods (2) – (4) are statistical counterpoints to the age bias plot. The statistical methods have the advantage of being quantitative. The age bias plot tends to be more sensitive. The WG concluded that both measures of bias are useful in an age calibration or quality control (QC) exercise.

Precision is a measure of consistency or reproducibility within or among age readers. Precision is often characteristic of experienced age readers, since they tend to be very consistent in their interpretations. However, precision is not a measure of accuracy, and precise ageing is not necessarily accurate. Most of the methods reported in Table 2.2.1 were measures of precision (Methods 5-14). Precision estimates can be expressed by a single number (e.g. CV=5%) and thus are easily understood. CV (Method 6) and APE (Method 5) are widely used in the literature, and thus are readily compared among age calibration exercises. They also have the advantage of being relatively insensitive to the age range in the study, unlike the simple percent agreement. The WG concluded that one measure of precision, either CV or APE, was a useful product of an age calibration or QC exercise.

It is important to note that all of the bias tests (both the age bias plot and the statistical tests of symmetry) are limited to pairwise comparisons of age readings; it is not possible to compare more than two readings at a time. In contrast, the precision measures, diagnostics and age error matrices can all be based on as many age readings as are available.

Age readers are often most interested in determining the cause of any systematic differences in age determination (bias) with other age readers, since it may be possible to correct the error if the source is known. These methods may be referred to as diagnostic methods (Methods 15-17). The use of multiple layers in Photoshop (or some other imaging program), with one layer per age reader, allows rapid comparison of growth increment interpretations across multiple age readers. This approach is visual, and not quantitative. Alternatively, simple descriptive statistics can be computed and analysed or else mixed effect models applied to growth increment width data providing a quantitative diagnostic of systematic growth increment interpretation differences. Although these methods likely provide too much detail for stock assessment experts, the WK concluded that they represent useful products of an age calibration exercise.

Age error matrices (Method 18) provide a quantitative summary of age reading error. These matrices, whether empirical or model-based (sensu Punt *et al.*, 2008), report misreading errors as proportion of an age group mis-aged as other ages. Empirical age error matrices are readily calculated from age-frequency tables, while model-based matrices require specialized software. Age error matrices are of particular value to stock assessment clients, since they can be incorporated into age-structured stock assessment models to correct for routine and unavoidable ageing error. Thus the WK concluded that they are an important product of an age calibration exercise.

2.2 Applying the selected output parameters for assessing their suitability

The quantitative parameters were tested on known datasets on two species. A haddock age test data were taken from an international age calibration study involving 7 expert ageing laboratories and age-validated otoliths. Two subsets of data were used in the examples shown below; the first compares two sets of age readings which are very similar, while the second compares age readings where one set of age readings shows substantial bias. The analysis was done using the software tool from NOAA: http://www.nefsc.noaa.gov/fbp/age-prec/Precision.xltx

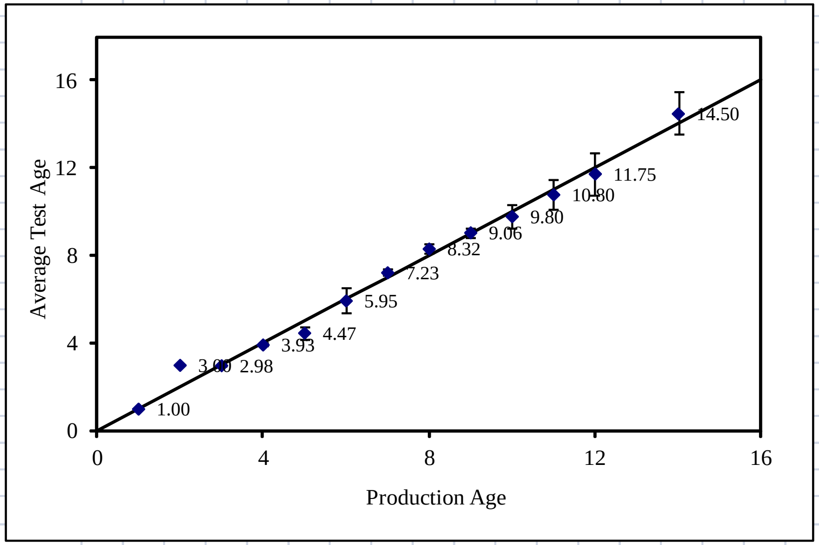
Sample Type/year	Species	Haddock
Aged 2X by ___	Date	10/14/14
	Age Reader	WHvsLH

N Aged	466			Bowker's test	
N Tested	464	Total CV	4.75%	Chi-sq	56.63
N Agreed	292			d.f.	27
Disagreed	172	%Agreement	62.9%	P-value	0.00
					**

Prod Age	N	N Agreed	%Agrmnt	Ave Age	s.d.	C.I.	95% C I	
0			#¡DIV/0!	0.00		####	#¡NUM!	#¡NUM!
1	3	3	100%	1.00		####	#¡NUM!	#¡NUM!
2	1		0%	3.00	#¡DIV/0!	####	######	######
3	62	61	98%	2.98	0.13	0.03	2.95	3.02
4	89	77	87%	3.93	0.36	0.08	3.86	4.01
5	36	14	39%	4.47	0.88	0.29	4.19	4.76
6	19	5	26%	5.95	1.31	0.59	5.36	6.54
7	60	38	63%	7.23	0.65	0.16	7.07	7.40
8	84	43	51%	8.32	0.85	0.18	8.14	8.50
9	50	36	72%	9.06	0.79	0.22	8.84	9.28
10	30	5	17%	9.80	1.52	0.54	9.26	10.34
11	20	6	30%	10.80	1.54	0.68	10.12	11.48
12	8	3	38%	11.75	1.39	0.96	10.79	12.71
13			#¡DIV/0!			####	#¡NUM!	#¡NUM!
14	2	1	50%	14.50	0.71	0.98	13.52	15.48
15			#¡DIV/0!			####	#¡NUM!	#¡NUM!
16			#¡DIV/0!			####	#¡NUM!	#¡NUM!
Total	464	292						

Omitted Samples

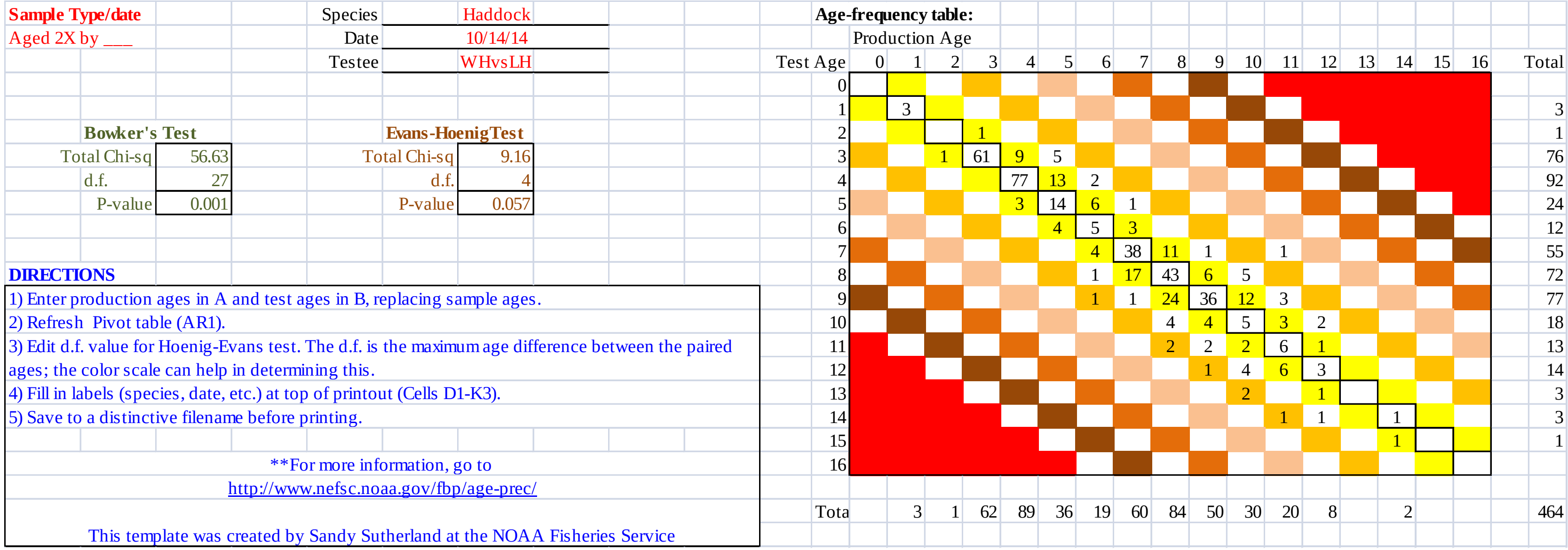
Prod Age	Test Age
	4
	4



Error bars indicate 95% confidence intervals

Test Age	Production Age 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1		3																3
2				1														1
3			1	61	9	5												76
4					77	13	2											92
5					3	14	6	1										24
6						4	5	3										12
7							4	38	11	1		1						55
8							1	17	43	6	5							72
9							1	1	24	36	12	3						77
10									4	4	5	3	2					18
11									2	2	2	6	1					13
12										1	4	6	3					14
13											2		1					3
14												1	1		1			3
15															1			1
16																		
Total		3	1	62	89	36	19	60	84	50	30	20	8		2			464

In the above example, the age bias plot shows no appreciable bias between the two sets of age readings, although there is a very slight tendency for the test age to underage at ages 5-6 and overage at ages 7-8. The Bowker's test of symmetry may detect this difference, since the test result is statistically significant (P=0.00). As such, it indicates that this test is highly sensitive to very small differences. The CV of 4.75% indicates that there is reasonably good precision between the two sets of age readings, since CV values less than 5 are considered relatively precise (Campana, 2001). The APE is not shown, but CV is around 1.41 times the APE (Campana, 2001).



Sample Type/date

Aged 2X by ___

Species	Haddock
Date	10/14/14
Testee	WHvsLH

Bowker's Test

Total Chi-sq	56.63
d.f.	27
P-value	0.001

Evans-HoenigTest

Total Chi-sq	9.16
d.f.	4
P-value	0.057

DIRECTIONS

1) Enter production ages in A and test ages in B, replacing sample ages.

2) Refresh Pivot table (AR1).

3) Edit d.f. value for Hoenig-Evans test. The d.f. is the maximum age difference between the paired ages; the color scale can help in determining this.

4) Fill in labels (species, date, etc.) at top of printout (Cells D1-K3).

5) Save to a distinctive filename before printing.

**For more information, go to
http://www.nefsc.noaa.gov/fbp/age-prec/

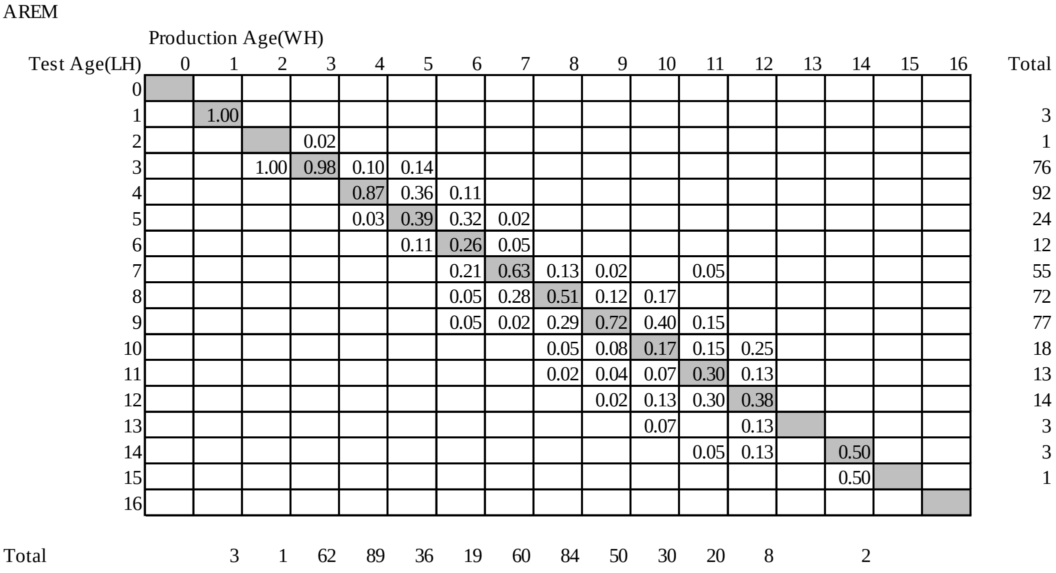
This template was created by Sandy Sutherland at the NOAA Fisheries Service

Age-frequency table:

	Production Age																	
Test Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1		3																3
2				1														1
3			1	61	9	5												76
4					77	13	2											92
5					3	14	6	1										24
6						4	5	3										12
7							4	38	11	1		1						55
8							1	17	43	6	5							72
9							1	1	24	36	12	3						77
10									4	4	5	3	2					18
11									2	2	2	6	1					13
12										1	4	6	3					14
13											2		1					3
14												1	1		1			3
15															1			1
16																		
Tota		3	1	62	89	36	19	60	84	50	30	20	8		2			464

The output above shows the age frequency table that results from this comparison, which is later used to generate the age error matrix. Also shown is an additional test of symmetry (the Hoenig test). This test also shows significant differences between the age readings, but of smaller magnitude than the Bowker's test.

The final product is the ageing error matrix, which is shown below. This matrix shows the proportion of each test age (the reference age) mis-aged as other ages. Therefore, the sum of each row is 1, equal to 100%.



AREM

Production Age(WH)

Test Age(LH)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1		1.00																3
2				0.02														1
3			1.00	0.98	0.10	0.14												76
4					0.87	0.36	0.11											92
5					0.03	0.39	0.32	0.02										24
6						0.11	0.26	0.05										12
7							0.21	0.63	0.13	0.02		0.05						55
8							0.05	0.28	0.51	0.12	0.17							72
9							0.05	0.02	0.29	0.72	0.40	0.15						77
10									0.05	0.08	0.17	0.15	0.25					18
11									0.02	0.04	0.07	0.30	0.13					13
12										0.02	0.13	0.30	0.38					14
13											0.07		0.13					3
14												0.05	0.13		0.50			3
15															0.50			1
16																		
Total		3	1	62	89	36	19	60	84	50	30	20	8		2			

In the second example shown below, the age bias plot indicates substantial underageing of the test age relative to production ages >5 yr. The Bowker's test P value of 0.00 confirms that the underageing is significant. The CV of 25.21% is also very high, but in this case, a measure of precision is much less important given the very high bias between the two sets of age readings.

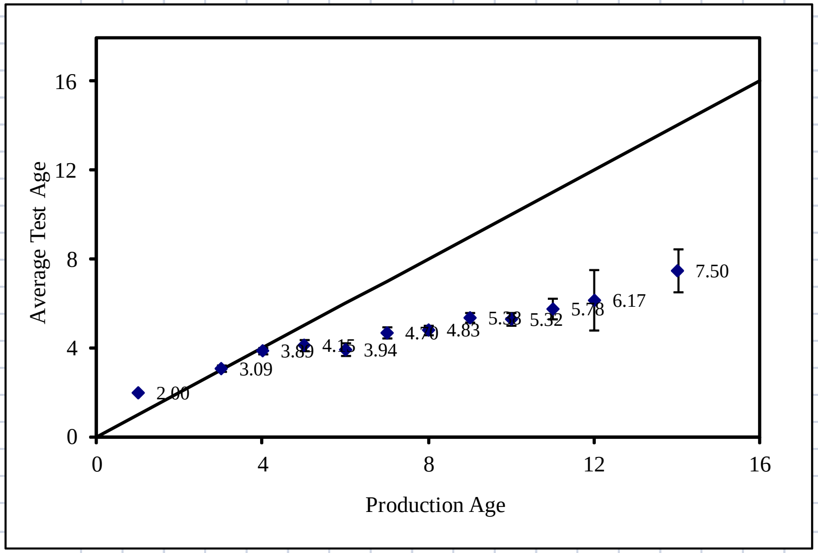
Sample Type/year	Species	Haddock	
Aged 2X by ___	Date	10/14/14	
	Age Reader	WHvsLH	

N Aged	466			Bowker's test	
N Tested	433	Total CV	25.21%	Chi-sq	271.78
N Agreed	93			d.f.	36
Disagreed	340	%Agreement	21.5%	P-value	0.00
					**

Prod Age	N	N Agreed	%Agrmnt	Ave Age	s.d.	C.I.	95% C I	
0			#¡DIV/0!	0.00		####	#¡NUM!	#¡NUM!
1	2		0%	2.00		####	#¡NUM!	#¡NUM!
2			#¡DIV/0!			####	#¡NUM!	#¡NUM!
3	55	37	67%	3.09	0.62	0.16	2.93	3.25
4	85	47	55%	3.89	0.69	0.15	3.75	4.04
5	34	9	26%	4.15	0.74	0.25	3.90	4.40
6	17		0%	3.94	0.56	0.26	3.68	4.21
7	60		0%	4.70	0.93	0.23	4.47	4.93
8	76		0%	4.83	0.93	0.21	4.62	5.04
9	50		0%	5.38	0.78	0.22	5.16	5.60
10	28		0%	5.32	0.82	0.30	5.02	5.62
11	18		0%	5.78	0.94	0.44	5.34	6.21
12	6		0%	6.17	1.72	1.38	4.79	7.54
13			#¡DIV/0!			####	#¡NUM!	#¡NUM!
14	2		0%	7.50	0.71	0.98	6.52	8.48
15			#¡DIV/0!			####	#¡NUM!	#¡NUM!
16			#¡DIV/0!			####	#¡NUM!	#¡NUM!
Total	433	93						

Omitted Samples

Prod Age	Test Age
	4
10	
	3
12	
8	
3	
3	
	4

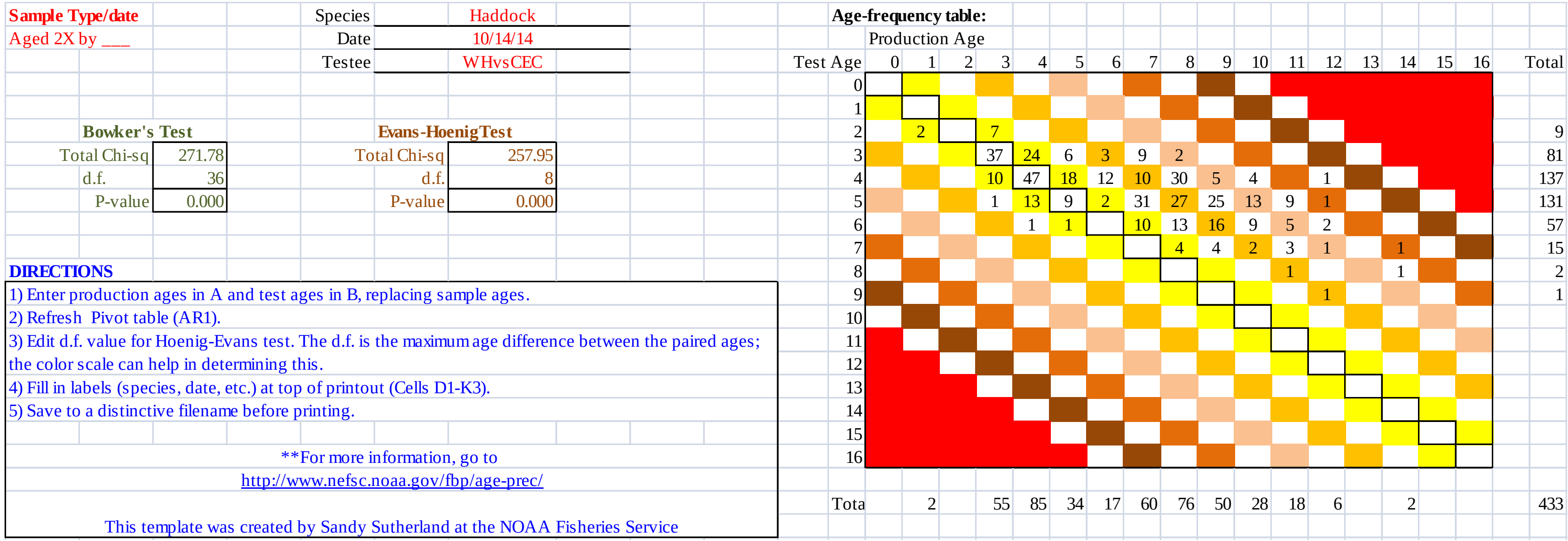


Error bars indicate 95% confidence intervals

Test Age	Production Age 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1																		
2		2		7														9
3				37	24	6	3	9	2									81
4				10	47	18	12	10	30	5	4		1					137
5				1	13	9	2	31	27	25	13	9	1					131
6					1	1		10	13	16	9	5	2					57
7									4	4	2	3	1		1			15
8												1			1			2
9													1					1
10																		
11																		
12																		
13																		
14																		
15																		
16																		
Total		2		55	85	34	17	60	76	50	28	18	6		2			433

The output below presents the age frequency table as well as the highly significant P value of 0.000 from the Hoenig test, confirming that bias is present.

Symmetry2kinds.xltx



Sample Type/date

Aged 2X by ___

Species	Haddock
Date	10/14/14
Testee	WHvsCEC

Bowker's Test	
Total Chi-sq	271.78
d.f.	36
P-value	0.000

Evans-Hoenig Test	
Total Chi-sq	257.95
d.f.	8
P-value	0.000

DIRECTIONS

1) Enter production ages in A and test ages in B, replacing sample ages.
2) Refresh Pivot table (AR1).
3) Edit d.f. value for Hoenig-Evans test. The d.f. is the maximum age difference between the paired ages; the color scale can help in determining this.
4) Fill in labels (species, date, etc.) at top of printout (Cells D1-K3).
5) Save to a distinctive filename before printing.

**For more information, go to
http://www.nefsc.noaa.gov/fbp/age-prec/

This template was created by Sandy Sutherland at the NOAA Fisheries Service

Age-frequency table:

	Production Age																	
Test Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1																		
2		2		7														9
3				37	24	6	3	9	2									81
4				10	47	18	12	10	30	5	4		1					137
5				1	13	9	2	31	27	25	13	9	1					131
6					1	1		10	13	16	9	5	2					57
7									4	4	2	3	1		1			15
8												1			1			2
9													1					1
10																		
11																		
12																		
13																		
14																		
15																		
16																		
Tota		2		55	85	34	17	60	76	50	28	18	6		2			433

Finally, the ageing error matrix from this comparison is shown below.

AREM

	Production Age(WH)																	
Test Age(CEC)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1																		
2		1.00		0.13														9
3				0.67	0.28	0.18	0.18	0.15	0.03									81
4				0.18	0.55	0.53	0.71	0.17	0.39	0.10	0.14		0.17					137
5				0.02	0.15	0.26	0.12	0.52	0.36	0.50	0.46	0.50	0.17					131
6					0.01	0.03		0.17	0.17	0.32	0.32	0.28	0.33					57
7									0.05	0.08	0.07	0.17	0.17		0.50			15
8												0.06			0.50			2
9													0.17					1
10																		
11																		
12																		
13																		
14																		
15																		
16																		
Total		2		55	85	34	17	60	76	50	28	18	6		2			

Although the NOAA software tool was used to generate all of the output shown above, it should be possible to generate all of these products from WebGR or other software after appropriate revision.

2.3 Example of diagnostics based on growth increment width

The analysis of growth increments can be very informative about the reasoning behind the attribution of a specific age. As such, the comparative analysis across readers can be used to evaluate the ageing process and detect severe differences in the interpretation of the otolith. The information generated will be valuable to correct errors or deepening into problems detected in bias and precision analysis.

WebGR stores information about the distance between marks, which constitutes the tool to identify rings. Using such information allows several analyses to be carried out, as mentioned before. Namely, visual analysis of the marks in each otolith, using image editing tools; statistics to extend this analysis and allow a wider assessment of the consistency between readers; and/or mixed effects models to test if readers are consistently interpreting the otolith rings. This section is focused on the second, while an example of using mixed effects models can be seen in WKSIBCA report (ICES, 2014).

Figure 2.4.1 presents the increment widths by mark for 4 readers and 4 otoliths. This analysis is numerically similar to overlaying the images of each reader for a specific otolith to assess their consistency. It allows the analysis of consistency between readers in interpretation of the structures counted.

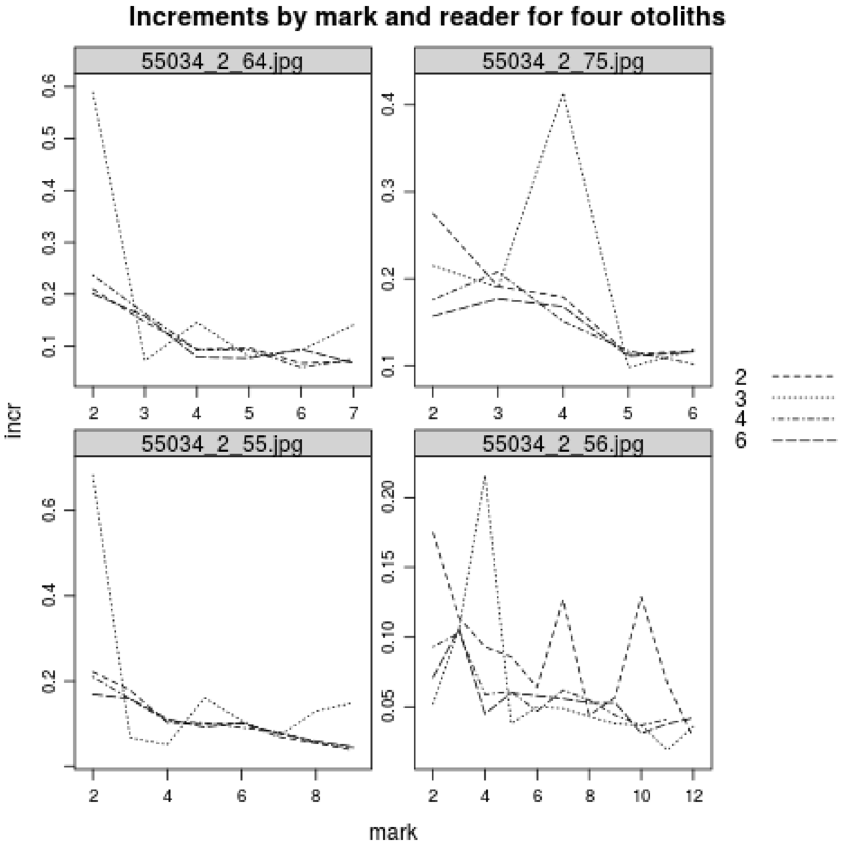


Figure 2.3.1 Increment widths by mark for 4 readers and 4 otoliths.

Figure 2.4.2 shows the coefficient of variation of the increment widths by mark that each reader set across otoliths. This analysis is related with reader's internal consistency. It shows how consistent each reader was in the interpretation of rings. In this case it shows that reader 3 was off in the initial rings/marks, while reader 2 was off on the older ages. The other readers seem consistent and present an increasing precision for older ages.

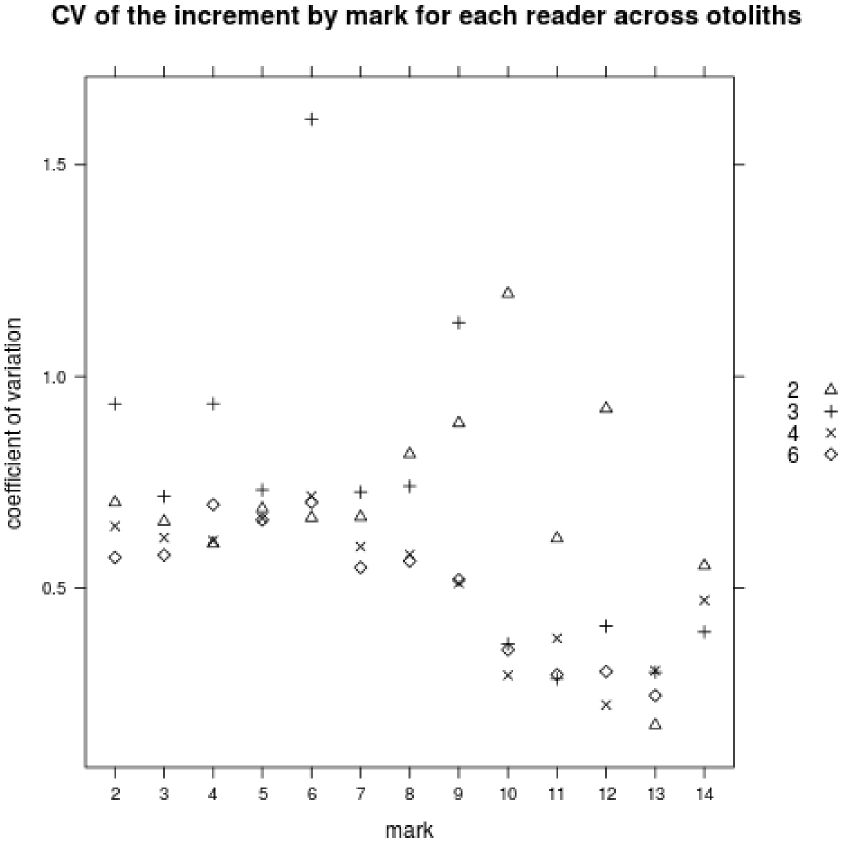


Figure 2.3.2. Coefficient of variation of the increment widths by mark that each reader set across otoliths

Figure 2.4.3 shows the average deviance of the increment width each reader gave to each mark from the median increment for that mark, across readers. It represents consistency between readers once that, in a perfect situation, all readers should identify the rings in the same place in the otolith, which would result in the same distance between marks, and deviance 0.

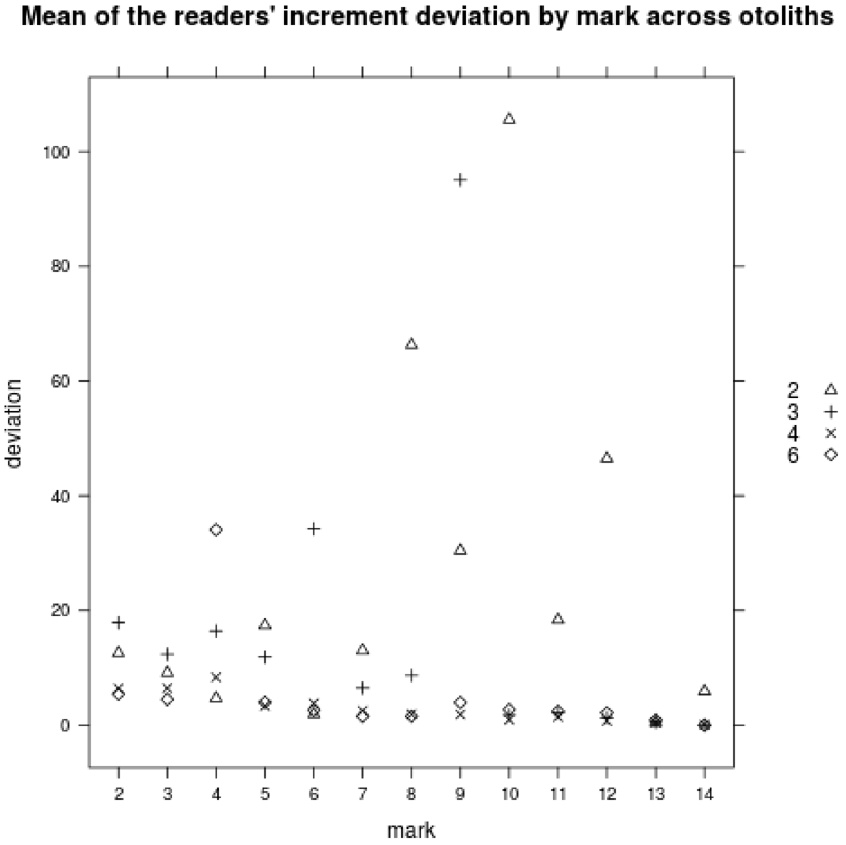


Figure 2.3.3 Average deviance of the increment width each reader gave to each mark from the median increment for that mark, across readers.

In all of these analyses, interpretation errors made at an early stage (i.e. at Mark 2) would be expected to propagate through all subsequent marks. Further work is required to better understand the implications of this error propagation, and how it would affect the interpretation of the analysis.

It's important to note that this analysis is based on a dataset that is not fully appropriate to it. To carry on this type of analysis the distances between rings should be measured along the same axis and, as much as possible, a linear axis. Currently the workshops ran with WebGR does not use this standard, which increases the variability by mark. Additionally, the WebGR design does not use the centre of the *nucleus,* the first mark is counted as one age, which invalidates the analysis of the first increment, a recognized major source of error. All these improvements can be made without having to heavily redesign WebGR.

2.4 Example of age reading error impact on stock assessment results

The ageing error matrix is an output which can readily be used to decrease age-related bias of the stock assessment if applied directly in the stock assessment models. The effect of incorporating ageing error matrices into an age-structured stock assessment model have been demonstrated for several stocks across the world (Reeves, 2003; Catalano and Bence, 2012) and can potentially affect estimates on e.g. mortality and biomass estimates, and the resulting advised TAC as illustrated in the below figures (drawn from Catalano and Bence, 2012).

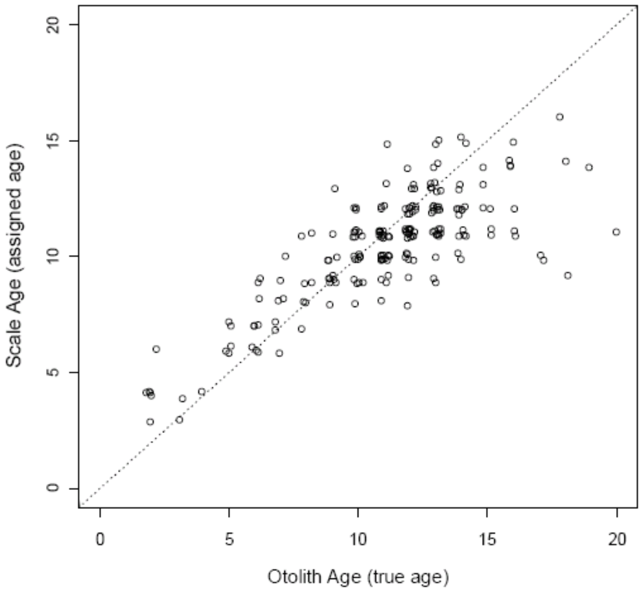


Figure 1. Scale age (y axis) vs. otolith age (x axis) for a lake whitefish from Lake Huron aged by both scales and otoliths. The dots are jittered to reveal the number of samples. The fine dashed line depicts the 1:1 line for reference.

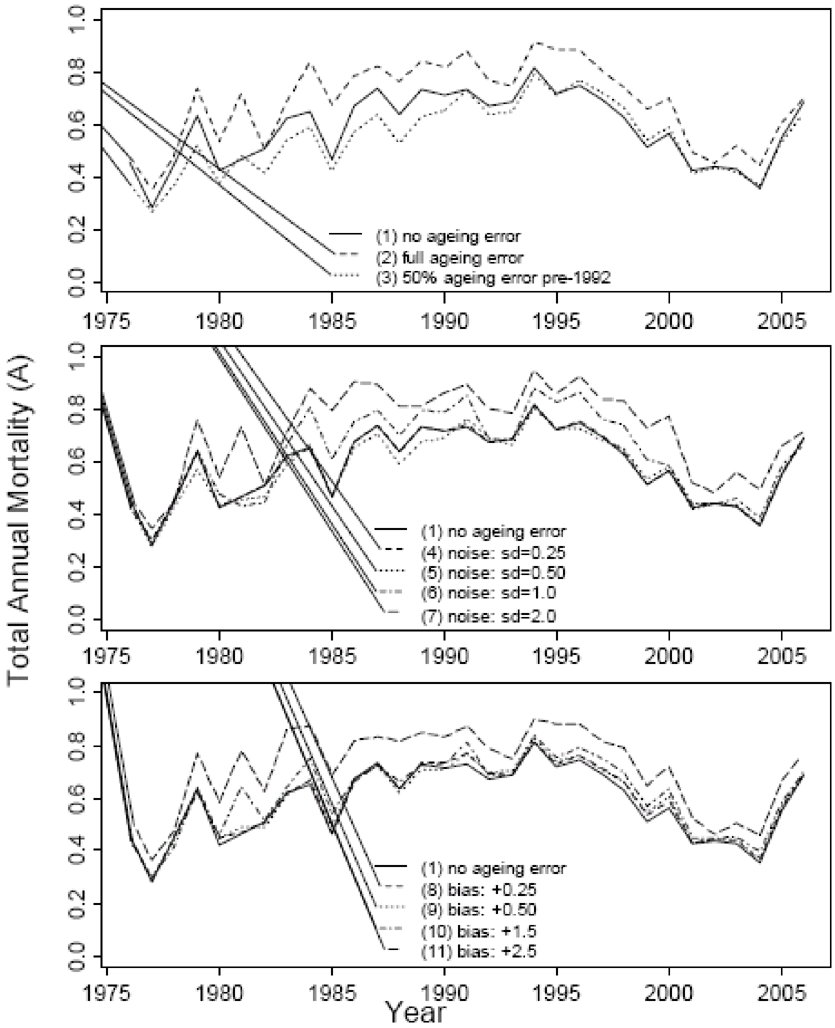


Figure 7. Time series of age-0 lake whitefish total annual mortality estimates from 11 stock assessment models that adjusted for varying types and amounts of ageing error. Each panel contains the estimated time series for the baseline model that did not adjust for ageing error (solid line; model 1). The upper panel shows estimates from two models (models 2 and 3) that adjusted for estimated amounts of ageing error from an analysis of lake whitefish that were double-aged with scales and otoliths. The middle and lower panels show results of hypothetical scenarios representing unbiased ageing with varying levels of noise (middle panel) and nearly noiseless errors with varying degrees of bias (lower panel).

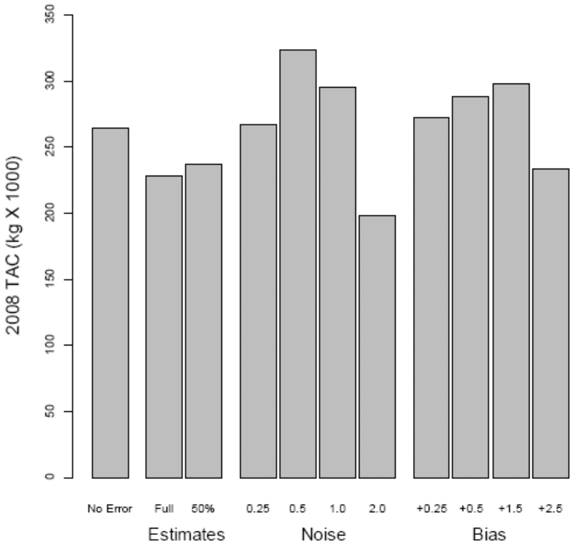


Figure 8. 2008 TACs (projected from the 2006 assessment) from 11 stock assessment models that adjusted for varying types and amounts of ageing error.

2.5 Maturity calibration workshops and suited outputs

The advantage of maturity in contrast to age reading studies is that it is much easier to know the "true" stage of a gonad by means of histological methods than to know a "true" age for an otolith.

However, in case of maturity calibration workshops there are additional difficulties compared to ageing calibration workshops. First and mostly importantly, the staging of gonads is not linked to a discrete or continuous but to a categorical variable since the number associated with each stage is not really a numerical number but a label. This causes that few of the statistical analysis used for ageing workshops can be applied to the staging in maturity calibration workshops.

The second difficulty lays in the different products that data producer (maturity stager) and data receivers (stock assessment scientists) wish to obtain from a maturity calibration exercise.

The maturity stagers participating in a workshop are interested in knowing the difficulties encountered in stage identification and the relative agreement between stagers for each single stage. On the other hand the end-users are mainly interested in the precision/accuracy in estimating the proportion of spawner/non spawner in the population, by length or weight, in order to input it in the used population model to estimate the Spawning-stock biomass (SSB).

Translating the stage categorization into a binomial form, i.e. 0 for immature and 1 for mature, simplifies the outputs and make them more applicable for end-users. Merging groups where the problem is between mature-stage A and mature-stage B returns a less noisy result (i.e. decrease the disagreement) and this conversion allows the application of the FSA package and computation of the statistics selected as output by WKSABCAL – albeit with caution since we are analysing categorical variables with tests designed for discrete variables like age.

The ‘reader-error-plot’ appears useful for the participants as an output from a maturity staging workshop while for the end-user a ‘correction matrix’ can be constructed for aligning the maturity data, showing the misclassification at length.

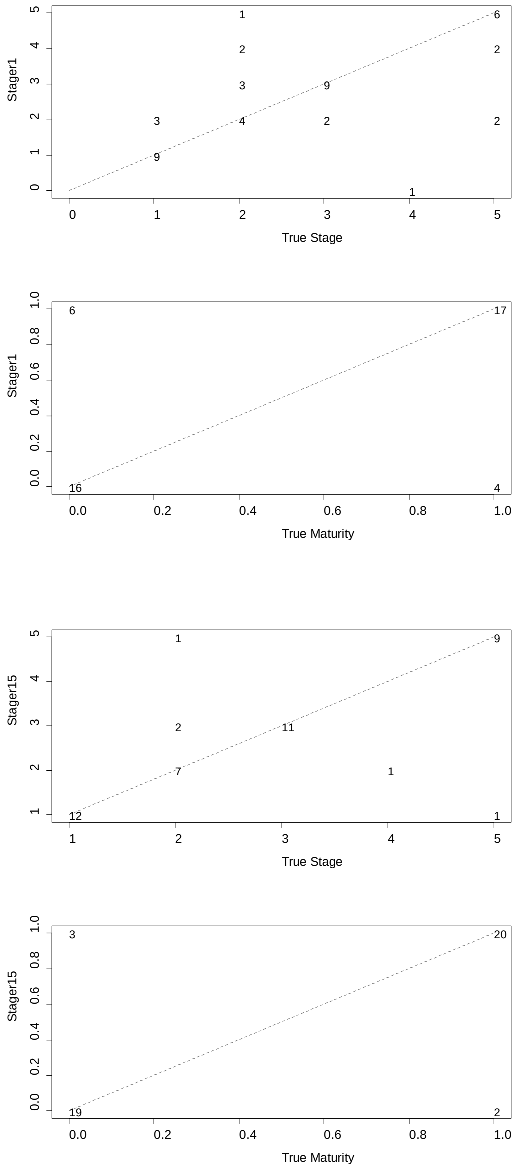


Figure 2.6.1 Stager error plot for all the stages (left) and for maturity (right) for two stagers, expert and new one.

3 ToR c) Review existing software for analysing calibration workshop data

Direct ageing studies require specific software tools to be used in calibration exchanges and workshops. Nowadays most of these studies are based on digital images of calcified structures where readers can annotate their readings. This procedure contributes to the standardization of ageing interpretation criterion among readers (as for example the misinterpretation of false checks or the differences in the position of the first *annulus*), and at the same time these annotated images allow to measure age increments.

Other tools that have been used are electronic forms or integrated databases containing sampling information and where ages or growth increments are stored. When non-integrated software is used, these forms have been developed in spreadsheets that are used for further statistical analysis in order to obtain precision, accuracy (relative or absolute), bias analysis and other outputs. In the case of integrated databases, none of the analysed programs implemented routines for statistical analysis but all of them have extraction routines for exporting the data to statistical programs.

Integrating images and ageing data analysis in the same software may reduce processing time and avoid errors due to data handling; however, available software is currently serving both requirements separately.

Available software was characterized according to two categories:

a) Age or maturity calibration exercises software which are currently based on electronic images of calcified structures or gonads

b) Software for statistical analysis of the results of the calibration exercises.

In the first case existing software known to the WK is described in **Table C1**. All these software allow the calibration of images, the integration of annotated layers from different readers in each of the images and to take real distances between consecutive marks.

The second parameter revised was if the software handles the workshop in web based interface or has to be installed locally and then the results of each reader has to be sent to the WS coordinator. Other parameter that was taken into account was if the software is commercial, not allowing users to modify the software and with an economic cost, or if it is open source, allowing changes to adapt the software to data and without cost.

The integration of a database able to handle the associated data to the images was analysed too, resulting that only scientific/industrial image analysis software and WebGR have integrated databases that can store that kind of data. This **Table C1** finally details how easy is to measure age increments and if the software is compatible with the multilayer-TIF format, which is the most commonly employed when working with images that need to handle various annotated layers.

Software for age data analysis.

Table C2 shows the available tools for statistical computing and graphics that were available and known by WKSABCAL, with a description of the statistical methods computed by software.

The strengths and weaknesses of the analysed statistical methods are discussed in section 2.2 of the report.

Table C1.

Software for age and maturity calibration exercises	**Characteristics**					
All these programs allow to calibrate images, annotate, measure	Web based	Type of license	Main purpose	Integrated Data Base (sample data, Images, annotated layers, ages)	Multilayer TIF Format	Easiness for measuring age increments
WebGR	Yes	Open source	Designed for Calibration Exercises	Integrated	No	Automatic
Adobe Photohop	No	Commercial	Photo editor	No	Yes	Time consuming
PaintShop Pro	No	Commercial	Photo editor	No	Yes	Time consuming
Gimp	No	Open source	Photo editor	No	No	Time consuming
ImageJ (TreeRings)	No	Open source	Image analysis	Integrated	Yes	Automatic
Visilog (TNPC)	No	Commercial	Image analysis	Integrated	No	Automatic
Nis-Elements D	No	Commercial	Image analysis	Integrated (only marked layers)	No	Automatic
Image-Pro (Otolith fish ageing)	No	Commercial	Image analysis	Integrated	Yes	Automatic
Image-Pro (Age & Shape)	No	Commercial	Image analysis	Integrated	Yes	Automatic

Table C2.

Software for calibration results analysis	Framework	Source	Data handling	Computed statistics and graphics														
				MSD	CCC	TDI	CP	MAD	PA	APE	CV	AREM	Age Bias plot	Text of simetry	Rho	W	Tau	Mixed effects models
Age Reading Comparisons (G. Etkin, 2000, excel spreadsheet)	MS-Excel		Easy to use. Prone to errors	Yes				Yes	Yes		Yes		Yes	Wilcoxon signed-rank test				
NOAA –NEFSC excel workbooks (Templates for Calculating Ageing Precision)	MS-Excel	http://www.nefsc.noaa.gov/fbp/age-prec/index.html	Easy to use. Prone to errors						Yes		Yes		Yes	McNemar's, Evans&Hoening's, Bowke's				
Agreement	R-package	http://cran.r-project.org/	Knowledge of R required	Yes	Yes	Yes	Yes											
agRee	R-package	http://cran.r-project.org/	Knowledge of R required		Yes													
irr	R-package	http://cran.r-project.org/	Knowledge of R required		Yes											Yes		
KappaGUI	R-package	http://cran.r-project.org/	Knowledge of R required		Yes											Yes		
FSA	R-package	http://fishr.wordpress.com/fsa/	Knowledge of R required					Yes	Yes	Yes	Yes		Yes	McNemar's, Evans&Hoening's, Bowke's				
nwfscAgeingError (AGEMAT.exe) (Punt et al., 2008)	R-package	https://r-forge.r-project.org/R/?group_id=1316	Knowledge of R required									Yes						
lme4	R-package	http://cran.r-project.org/	Knowledge of R required															Yes

MSD= Mean Square Deviation, CCC= Concordance Correlation Coefficient, TDI= Total Deviation Index, CP= Coverage Probability, MAD= Modal Age Difference, PA= Percentage Agreement
APE= Average Percentage Error, CV= Coefficient of Variation, AREM= Age Readings Error Matrix, Rho= Average Spearman's Rho, W= Kendal's Coefficient of Concordance, Tau= Average Tau

4 ToR d) Define data summaries and analysis outputs required by calibration workshop participants and as stock assessment input

4.1 Feedback from Calibration workshops, Benchmarks and Working Groups chairs

A calibration workshop has the basic purpose, as part of the quality assurance procedure, to identify sources of errors and inconsistencies among laboratories in stock-specific biological measurements, quantify these errors and ultimately include them in stock assessment. In this view the results of these exercises, published in extensive ICES reports, have the purpose of reaching both the personnel observing and classifying the biological structures and the scientists involved in the estimation of stock biological parameters. The initial idea of this ToR was to investigate how the calibration workshops outcomes could be improved in order to be useful and exhaustive for both recipients.

Several surveys were conducted among calibration workshops, benchmarks and working group chairs. The chairs were inquired to answer some questions in order to identify weaknesses in the currently used data analysis and possible enhancements.

The response by those groups, although not wide as expected, highlighted that so far the final report from these calibration workshops has been hardly interpreted by the participants, if not explained to them by national coordinators, and at the same time overlooked during the process of benchmarking and stock assessment.

For what it concerns age calibration workshops, readers found it very difficult to interpret some of the tables included into the report, arguing that a visual output and an exhaustive text explaining for instance the observed CV is more useful than a number in a table. What a reader is more interested in, is his/her own performance compared with the others.

To address the requirements of the users of age related data, a questionnaire was sent to the recent chairs of stock assessment working groups, benchmark and review groups. Most answers only required a yes or a no, but respondents were encouraged to elaborate their answers wherever they found it useful. The following 14 questions were sent to persons involved in 40 assessments/benchmarks during the recent three years.

Please give the name of the WG/Benchmark + stock, and for each stock in question answer the following questions:

Has age estimation and corresponding errors been explicitly addressed in the WG?

Has ageing error been considered explicitly during the benchmark?

Was there an assumption of ageing accuracy (lack of bias in relation true age)?

Was there an assumption of ageing precision (reproducibility)?

What was the chosen/preferred stock assessment model?

Did ageing bias or uncertainty influence the choice of stock assessment model?

Has the selected stock assessment model possibility to handle estimated age uncertainty?

Has the selected stock assessment model identified specific ageing errors in the samples?

Would you like the stock assessment model to handle age uncertainty?

Has any stock assessment model been run with input of age error?

Has the influence of error on stock perception been analysed?

Has there been a process of feedback to age readers?

In what format should age estimation error be input to stock assessment?

The questionnaire was answered by 6 chairs responsible for work in 4 assessment working groups and 5 benchmarks including results concerning 34 stocks analysed with more than 10 different models.

The general pattern was that WG and Benchmark exercises were little concerned with explicit age reading errors however assuming that data had reading errors but were without bias. Only in a few cases the choice of assessment model was influenced by problems of age reading errors. For Greenland halibut a production model has been used and for the anglerfish stocks (4) and hake stocks (2) length based methods were chosen due to inability to age the fish. Among the surveyed methods stock synthesis (SS3) was identified as having the option of applying input of ageing error, however this option was not used in any of the cases. There were no reports where the stock assessment method directly had pointed out flaws in the age estimation process; however ageing errors were occasionally identified during the initial data screening during the stock assessment WG.

An example illustrating the process is the Icelandic Greenland halibut stock. An age based model was used until around 2004 but abandoned because of assumed poor quality of the age reading confirmed in a workshop in 2006. A stock production model has been used since. A workshop was held in 2011 where different age reading techniques were used. No agreed procedure is, however, agreed upon yet. A benchmark was held in November 2013, which continued to use the stock production model, but it was also decided to work more on a gadget model (where growth information is used - from otolith reading and growth estimates from tagging studies). The gadget model is supposed to replace the stock production model when it is worked through properly.

The opinions were divided as to whether stock assessment models should be able to handle age uncertainty. In some cases the assumption of uncertainty in catch-at-age in some models (e.g. SAM, North Sea turbot assessment) appears to satisfy the stock assessors. With regard to input of age reading error into existing models only “stock synthesis” was identified as having this capability (although SAM is being developed to do so too) however there was very limited experience among the responders with this option. Similarly other approaches to analyse influence of ageing error on stock perception were quite limited and the feedback to age readers even more so.

Defining the best format for age reading errors as input to stock assessment and benchmarks was indicated to be a topic for further research. An immediate useful input would be an estimate of which ages may be confounded. The age from which strong confusion occurs could be the basis for defining the +group. Further the possibility to separate ageing errors from sampling errors in both catches and surveys was seen as a modelling advantage. However this separation would also have potential but relatively minor effects on estimates of weights at age and maturity-at-age in the stock assessment model fits.

4.1.1 The Age Reading Error Matrix

Different output formats from age reading workshops was discussed at WKSABCAL from submitting tables of reader raw data over ageing error matrices to summary statistics of variation and bias.

Age reading error matrices were found to provide an intermediate level of detail useful at routine stock assessment exercises, however disaggregated data at individual fish and reader level with additional spatially and temporally resolved covariates would probably be preferable in many benchmark situations where modelling of data quality is an issue.

For data limited approaches an age reading error matrix may not fill the needs for modelling growth, due to a potential correlation between observed age and size.

Exploring the age reading error matrix through simulation

McBride´s simulation data were analysed to compare the empirical age reading error matrix with the Bias plot for 5 readers. One observation for each true age was selected and one simulated age was computed for each of 5 readers with different performance parameters.

Precision was calculated for each simulated reader and all readers together (Table 4.1.1). Symmetry tests and bias plots were carried out for each reader vs. true age (Table 4.1.2 and Figure 4.1.1). RPackage FSA (Version: 0.4.28. Derek H. Ogle. http://fishr.wordpress.com/) was used.

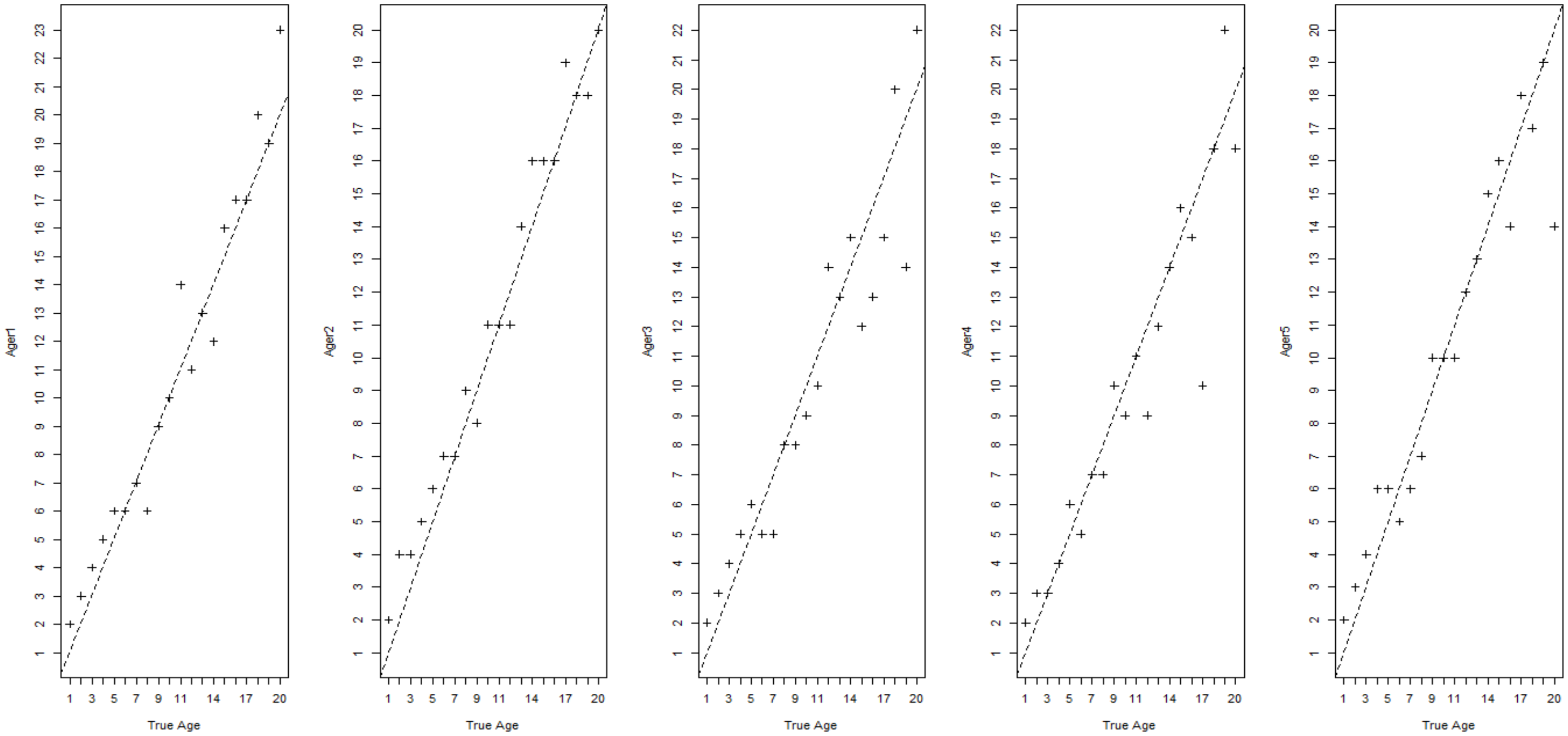
Table 4.1.1. Precision of the 5 simulated readers and all readers together vs. true age.

	CV	APE	PercAgree
AGER 1	10.23	7.232	35
AGER 2	10.72	7.579	25
AGER 3	13.68	9.675	10
AGER 4	10.74	7.598	30
AGER 5	11.8	8.344	20
ALL	11.88	8.676	10

Table 4.1.2. Tests of symmetry for the 5 simulated readers vs. true age.

	SymmertyTest	McNemars	EvansHoenig	Bowkers
AGER 1	df	1	3	13
	chi.sq	3.77	6.83	13.00
	p	0.05	0.08	0.45
AGER 2	df	1	2	14
	chi.sq	5.40	6.00	13.00
	p	0.02	0.05	0.53
AGER 3	df	1	4	17
	chi.sq	0.00	3.60	16.00
	p	1.00	0.46	0.52
AGER 4	df	1	4	11
	chi.sq	0.29	2.00	8.00
	p	0.59	0.74	0.71
AGER 5	df	1	3	14
	chi.sq	0.25	1.69	12.00
	p	0.62	0.64	0.61

Figure 4.1.1. Bias plots for the simulated age readings (y-axis) vs. true age (x-axis) dotted line indicates identity (1:1).



All readings were combined and the proportion of the simulated ages according to each true age was recorded in the age reading error matrix.

Table 4.1.3. Empirical ageing reader error matrix. Proportion of each true age (rows) assigned to each read age for all readers together (columns). Grey cells indicate identity (1:1).

AGE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	22	23
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0.8	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0.2	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0.2	0.6	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0.6	0.2	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0.2	0.2	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0.2	0.4	0.2	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0.4	0.2	0.4	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0.4	0.4	0.2	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0.4	0.4	0	0	0.2	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0.2	0	0.4	0.2	0	0.2	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0.2	0.6	0.2	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0.2	0.4	0.2	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0.8	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0.2	0	0.2	0.2	0.2	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.4	0	0.4	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0.2	0.4	0	0.2	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0.2	0	0.2	0.2	0.2

In Figure 4.1.1 the individual errors of the different readers are shown, whereas in Table 4.1.3 the total age error matrix for all readers together is tabulated. For age readers the graphical illustration in plots may readily indicate areas of difference in age interpretation that could be further investigated. On the other hand in the tabulated output the overall error structure is easily interpreted and contains the necessary information for stock assessment working groups.

4.2 Computing ageing errors to include in stock assessments

Age structured stock assessments fit population dynamics models to fishery dependent information (e.g. landings and discards) and fishery-independent information (e.g. scientific survey abundance estimates). This information includes age composition data for landings, discards, and abundance indices. Central to the use of age composition data in stock assessments is the relationship between ages obtained by reading hard structures from animals and their true age (Punt *et al.*, 2008). In this relationship bias and imprecision may occur: Ageing bias occurs when there are systematic differences between the ages assigned to animals and their true age, ageing imprecision occurs when age reading errors occur randomly (Punt *et al.*, 2008).

Errors in the age composition data may smooth out estimates of recruitment (Reeves, 2003) if not accounted for. Consequently, ageing errors can mask the existence of stock–recruit relationships (Dorval *et al.*, 2013; Richards *et al.*, 1992). Modern statistical models can take account of bias and precision in estimating ageing errors (Richards *et al.*, 1992; Punt *et al.*, 2008; Dorval *et al.*, 2013).

Following Punt *et al.*, (2008) and Dorval *et al.*, (2013), age-reading error can be represented using a matrix that specifies the probability of an animal of true age a being aged to be age a', $P(a'|a)$. To construct the function $P(a'|a)$ we need to estimate parameters of functions that could be used to model the relationship between true and estimated age. Richards *et al.*, (1992) developed a method for this based on maximum likelihood, allowing for considerable flexibility in the relationship between (true) age and the expectation and imprecision of the estimate of this age. Dorval *et al.*, (2013) coined the implementation of this model by Punt *et al.*, (2008) the "Ageing Error Matrix (Agemat)" model.

The Ageing Error Matrix (Agemat) model computes ageing-error matrices based on otoliths that are aged multiple times by one or more readers. When assuming that (1) ageing bias depends on reader and the true age of fish; (2) the age-reading error standard deviation depends on reader and true age; and (3) age-reading error is normally distributed around the expected age then the probability to assign an age *a*' to a fish of true age *a* is (Punt *et al.*, 2008):

$$P^i(a'|a,\phi) = \int_{a'}^{a'+1} \frac{1}{\sigma_a^i(\phi)\sqrt{2\pi}} \mathrm{e}^{\frac{-\left(a'-b_a^i(\phi)\right)^2}{2\left(\sigma_a^i(\phi)\right)^2}} da'. \text{ Eq. [1]}$$

Here b_a^i is the expected age when reader i determines the age of a fish of true age a, and similarly σ_a^i is the standard deviation. Both are affected by ϕ, a vector of parameters ultimately determining the ageing error matrices. Following Punt *et al.*, (2008) the probabilities obtained from Eq. [1] can be set to zero for values of a'< 0 and larger than a prespecified maximum age.

The age-reading error standard deviation and ageing bias are subsequently described using functional forms. The functional form of the age-reading error standard deviation and ageing bias frequently used in literature (see Richards *et al.*, 1992, Heifetz *et al.*, 1998, Punt *et al.*, 2008, Dorval *et al.*, 2013) uses three parameters and allowing for linear or non-linear functions.

The values for the parameter vector ϕ can be estimated by maximizing the likelihood function for the complete set of age readings *A*, using a set of *J* otoliths, read by all readers

$$L(A|\beta,\phi) = \prod_{j=i}^{J} \sum_{a=L}^{H} \beta_a \prod_{i=1}^{I} P^i\left(a_{i,j}\middle|a,\phi\right), \text{ Eq. [2]}$$

where $a_{i,j}$ is the age assigned by reader i to the j^{th} otolith (Punt *et al.*, 2008, Dorval *et al.*, 2013). The minimum and maximum ages are given by *L* and *H*, respectively.

4.2.1 Use of the Ageing Error Matrix in stock assessments

The model predictions upon which the likelihood component in the stock assessment for the age composition data is based are then a function of the model estimate for the observed catch of animals of age a after accounting for ageing error. Given $P(a'|a).$, this prediction is

$$C_{a'} = \sum_a C_a P(a'|a).$$

Here, C_a is the model estimate of the catch of animals of age a, and $C_{a'}$ is the model estimate of the catch of animals of perceived age a' after accounting for age-reading error.

The ability to account for age-reading error is included in several stock assessment programs, such as stock synthesis (Methot, 2000, 2007), Coleraine (Hilborn *et al.*, 2003), and CASAL (Bull *et al.*, 2003). However, although these assessment programs include the ability to account for age-reading error given an age-reading error matrix, they do not include the facility to internally estimate age-reading error matrices (Punt *et al.*, 2008). Also, assessment models are not uniformly structured. For example, assessment programs used in southern Australia allow including ageing errors per individual reader (Punt *et al.*, 2008), whereas Stock Synthesis (SS, Methot, 2000) allows only a single vector of ageing error as input in the model (Dorval *et al.* 2013).

Implementation

An implementation of the methodology in R as a package named "nwfscAgeingError" has been developed by Thorson *et al.*, (2013). The main function accepts data with rows being unique reading records and columns corresponding to readers or labs with a unique reading error and bias. The model allows for approximately 15 unique columns. An additional column on the left-hand side of the data matrix indicates the number of otoliths with that unique read record.

The main function named 'FnRun()' writes data in the necessary format and then calls an executable created in ADMB language. The model requires several additional inputs on the function form of the bias and the imprecision. Once the model is run and parameters are estimated, plotting routines are available.

Additional considerations

Use of the ageing error matrix in age structured model has been well described. Methods for using and interpreting ageing error matrices outside age structured assessments should be explored. For instance, how ageing error affects methods for assessing data limited stocks does not appear to be studied in equal detail. However, methods using the von Bertalanffy growth curve and its parameters L_∞ and K estimated from reading hard structures from animals probably suffer from ageing error.

In addition, transforming age-read information outside assessments should also be explored.

4.3 Data summaries and analysis outputs from calibration workshops and their dissemination

There is an obvious 'basic' audience for the outputs from a calibration workshop. Outcomes in fact ought to be directly used by both the calibration workshop participants (data producers) for improving the precision of age/maturity determinations and the stock assessment groups (data receivers) for integration in assessment models. WKSABCAL ascertained that there is however a wide range of additional potential recipients all with various requirements for the output from a calibration workshop report. The discussion led to the identification of seven potential data receivers.

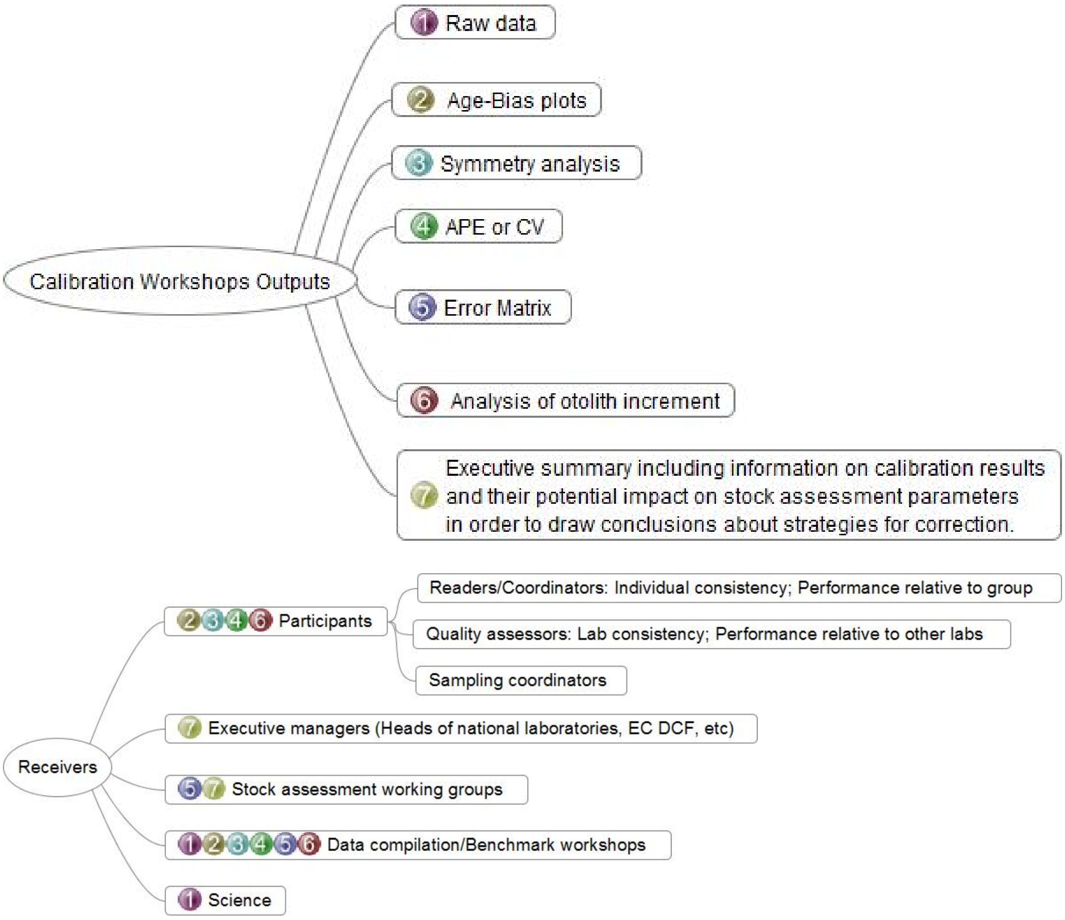

Calibration Workshops Outputs
1 Raw data
2 Age-Bias plots
3 Symmetry analysis
4 APE or CV
5 Error Matrix
6 Analysis of otolith increment
7 Executive summary including information on calibration results and their potential impact on stock assessment parameters in order to draw conclusions about strategies for correction.
Receivers
2 3 4 6 Participants
Readers/Coordinators: Individual consistency; Performance relative to group
Quality assessors: Lab consistency; Performance relative to other labs
Sampling coordinators
7 Executive managers (Heads of national laboratories, EC DCF, etc)
5 7 Stock assessment working groups
1 2 3 4 5 6 Data compilation/Benchmark workshops
1 Science

Figure 4.3.1 on flow of information from WKs to specified receivers, illustrating the various needs

As already stressed the participants' main interest is simply evaluating their consistency compared to the others while an age reader coordinator as well as a sampling coordinator will desire to also receive from calibration workshops' results a perception of the difficulties of a particular stock age reading, ascertaining the inter-readers variation (considering also other institutes). Results will be used as a guide for deciding whether actions for potential improvement or correction are needed. Age bias plots, symmetry analysis, APE/CV and increment analysis fulfil all those requirements. The same outputs will be also a tool for Quality assessors (in case of accredited laboratories) to rate the accuracy of an institute and individual readers in order for them to assess compliance with the norm.

Head of national laboratories and EU DCF, on the other hand, will only wish for a comprehensive and conclusive explanation of results and their impact on stock assessment parameters in order to draw conclusion about strategies for correction.

Stock assessment working groups will need a more quantitative analysis of the results for obtaining a better understanding on the perceived variance around the estimated age proportions for their particular stocks. In this case age error matrices (see section 4.1.1) represent a more operational output that could directly be applied in stock assessment age based models.

Raw data need to be always available in an easily accessible format, possibly directly downloadable form WebGR, for data compilation and benchmark workshops or for any other potential research investigation.

5 Conclusions

After analysing a list of methods available for the analysis of agreement between different analysts, WKSABCAL recommended the following methods/analysis to be run by age calibration workshops:

- To access bias
  - ABP - Age-bias plot
  - TS - Tests of symmetry
- To access precision
  - APE - Average Percentage Error
  - CV - Coefficient of Variation
- As diagnostics for problems found by the previous analysis
  - Analysis of otolith increments, both through image layers and statistically
- As output to stock assessment groups
  - AREM - Age Readings Error Matrix

For maturity staging workshops, the group explored different outputs and concluded that maturity staging bias plots would be the best outputs for participants, while for assessment working groups it should be focus on the maturity model correction or errors. The workshop did not explore how staging errors could be taken by Probabilistic Maturation Reaction Norms, but recommend this method to be explored.

It is important to note that if validated material is unavailable, bias cannot be computed and the analysis is limited regarding its assessment. It is thus recommended that regardless of the scope of the calibration workshop, an effort should be made to validate the age reading of the species/stock under consideration.

Additionally, the group concluded that the raw data from a calibration workshop should be documented by ICES and made available for those interested in running alternative estimation procedure or for more thorough analyses that may be available in future; this would assure that historical datasets can be analysed with the same method and past and present age reading performance can be compared and corrections could be integrated in long-term assessment datasets. The portfolio of analysis/methods identified by WKSABCAL allows a thorough perspective of the precision and bias/deviation associated with the ageing process of the stock, which should constitute an important knowledge set for those interested, like age or maturity analysts, stock assessment working groups etc.

The survey carried out by WKSABCAL showed that stock assessment working groups are aware of the importance of age reading errors may have on their results, but none of them is currently integrating these errors. The link between calibration workshops and stock assessment is very weak and not operational, which makes it very difficult to integrate these error sources. WKSABCAL explored how such links can be made more operational and concluded that the usage of Age Reading Error Matrices could be the right output to be provided by age calibration workshops to stock assessment working groups.

Nevertheless, an operational integration of age reading errors and/or maturity staging errors into stock assessment requires methodological developments that go beyond the scope of this workshop. WKSABCAL explored possible ways to facilitate this integration through the development of likelihood components in stock assessment models, MCMC approaches, or the like.

Taking into account the current advisory framework in ICES, WKSABCAL consider that benchmark workshops may be the right place to explore and decide, if and how these uncertainties should be integrated in to stock assessment. While stock assessment working groups should be provided with an operational methodology, that does not require long thorough analysis of the calibration workshops' results.

With regards to the software packages available, WebGR is the most suitable for running workshops, while FSA is the most complete for the analysis of the age reading results. None of these packages can run all the methods recommended by the WKSABCAL, although they can be further developed to accommodate most of them. In particular FSA is focused on age readings and do not have routines for maturity staging. WebGR has the potentiality to develop and implement the methods recommended by WKSABCAL.

Finally, the group concluded that there are other actors that may take advantage of calibration workshops' results and developed a framework to disseminate the information in a tailored way, to improve the uptake of the results by the different stakeholders.

6 References

Bull B, Francis RICC, Dunn A, McKenzie A, Gilbert DJ, and Smith MH (2003) CASAL (C++ algorithmic stock assessment laboratory): CASAL user manual v2.01-2003/8/01. NIWA Tech. Rep. No. 124.

Campana, S.E., Annand, M.C., McMillan, J.I. 1995. Graphical and Statistical Methods for Determining the Consistency of Age Determinations. Transactions of the American Fisheries Society Vol. 124, 1: 131-138

Campana, S.E. 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.

Catalano, M.J. and Bence, J.R. 2012. The sensitivity to assumed ageing error of the stock assessment used to recommend lake whitefish yield for 2008 from management unit WFH01 of Lake Huron. QFC Tech. Rep. 2011-01

Dorval E, McDaniel JD, Porzio DL, Felix-Uraga R, Hodes V, Rosenfield S (2013) Computing and selecting ageing errors to include in stock assessment models of Pacific sardine (Sardinops saga) CalCOFI Rep., Vol. 54, 2013

Eltink, A.T.G.W. 2000. Age reading comparisons. (MS Excel workbook version 1.0 October 2000) Internet: http://efan.no.

Heifetz J, Anderi D, Maloney NE, and Rutecki TL (1998) Age validation and analysis of ageing error from marked and recaptured sablefish, Anoplopama fimbria. Fish. Bull. (Washington, D.C.), 97: 256–263.

Hilborn R, Maunder, M, Parma A, Ernst B, Payne J, and Starr P (2003) COLERAINE: a generalized age-structured stock assessment manual. User’s manual version 2.0. SAFS-UW-0116

ICES 2010. Report on the Workshop on Age Reading of Mackerel (WKARMAC). Available at http://www.ices.dk/community/Pages/PGCCDBS-doc-repository.aspx#gui

ICES 2011a. PGCCDBS Guidelines for Workshops on Age Calibration (update). Available at http://www.ices.dk/community/Pages/PGCCDBS-doc-repository.aspx#gui

ICES 2011b. Report on the Workshop on Age Reading of Greenland halibut (WKARGH). Available at http://www.ices.dk/community/Pages/PGCCDBS-doc-repository.aspx#gui

ICES 2014. Report of the Workshop on Scoping for Integrated Baltic Cod Assessment. 1–3 October 2014. Gdynia, Poland. ICES CM 2014/ACOM:62

Methot R. (2000) Technical description of the Stock Synthesis Assessment Program. NOAA Technical Memo. NMFS-NWFSC-43, 46p.

Methot R (2007) User manual for the integrated analysis program Stock Synthesis 2.

Punt, A.E., Smith, D.C., KrusicGolub, K. and 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. Canadian Journal of Fisheries and Aquatic Sciences, 65: 1991-2005.

Reeves, S. A. 2003. A simulation study of the implications of age-reading errors for stock assessment and management advice. – ICES Journal of Marine Science, 60: 314–328.

Richards LJ, Schnute JT, Kronlund AR, and Beamish RJ (1992) Statistical models for the analysis of ageing error. Can. J. Fish. and Aquat. Sci. 49:1801–1815.

Thorson JT, Stewart IJ, Taylor IG, Punt AE (2013) Using a recruitment-linked multispecies stock assessment model to estimate common trends in recruitment for US West Coast groundfishes. Mar Ecol Prog Ser 483: 245–256

7 Annexes

Annex 1: List of participants

NAME	INSTITUTE	EXPERIENCE	TOR AND TASK	E-MAIL ADDRESS
Ernesto Jardim	JRC	Calcified structures, statistical analysis, stock assessment	Co-Chair	ernesto.jardim@jrc.ec.europa.eu
Lotte Worsøe Clausen	DTU Aqua	Calcified structures, stock assessment	Co-Chair	law@aqua.dtu.dk
Cristina Morgado	ICES			cristina@ices.dk
Steven Campana	Canada	Calcified structures, statistical analysis	Invited external expert	steven.campana@dfo-mpo.gc.ca
Uwe Krumme	Thünen Institute of Baltic Sea Fisheries, Rostock	Biol.sampling	ToR a)	uwe.krumme@ti.bund.de
Julia Wischnewski	Thünen Institute of Baltic Sea Fisheries, Rostock	Statistical analysis	ToR b)	julia.wischnewski@ti.bund.de
Iñaki Quincoces	Azti	WebGR	ToR c)	iquincoces@azti.es
Henrik Mosegaard	DTU Aqua	Calcified structures, stock assessment	ToR d)	hm@aqua.dtu.dk
Pablo Quelle	IEO Santander	Calcified structures, statistical analysis	ToR a)	pablo.quelle@st.ieo.es
Richard McBride	NOAA	Calcified structures, statistical analysis	ToR b)	Richard.McBride@noaa.gov (by correspondence)
Enrique Rodriguez-Marin	IEO Santander	Calcified structures, statistical analysis	ToR c)	rodriguez.marin@st.ieo.es
Francesca Vitale	SLU	Maturity staging, stock assessment	ToR d)	francesca.vitale@slu.se
Jan Jaap Poos	IMARES	statistical analysis, assessments and R-coding	ToR b)	janjaap.Poos@wur.nl (from Wednesday)
Kestas Plauska	Fisheries Service, Lithuania	-	ToR a)	kestas.plauska@gmail.com
Deividas Norkus	Fisheries Service, Lithuania	-	ToR a)	deividas.norkus36@gmail.com

Name	Institute	Experience	ToR and Task	e-mail address
Ana Costa	IPMA, Portugal	Maturity staging, stock assessment	Organiser	amcosta@ipma.pt
Cristina Nunes	IPMA, Portugal	Histological (Gonad) structures, maturity staging	ToR d)	cnunes@ipma.pt
Andreia Silva	IPMA, Portugal		ToR d)	avsilva@ipma.pt
Catarina Maia	IPMA, Portugal		ToR d)	cmaia@ipma.pt
Inês Farias	IPMA, Portugal	Calcified structures, statistical analysis	ToR d)	ifarias@ipma.pt

Annex 2: WKSABCAL 2014 Term of Reference

WKSABCAL - Workshop on Statistical Analysis of Biological Calibration Studies

2013/2/ACOM35 A **Workshop on Statistical Analysis of Biological Calibration Studies [WKSABCAL]** will be established Lotte Worsøe Clausen, Denmark and Ernesto Jardim, Portugal, and will meet in Lisbon 13–17 October 2014 to:

a) Compile statistical methods for analysing reader agreement;

b) Identify the strengths and weaknesses of each method for fisheries calibration studies;

c) Review existing software for analysing calibration workshop data;

d) Define data summaries and analysis outputs required by calibration workshop participants and as stock assessment input;

e) Draft a review paper based on workshop presentations, discussions and results.

WKSABCAL will report by 3 November 2014 for the attention of ACOM and PGCCDBS.

Supporting Information

Priority:	High. Age and maturity data are fundamental parts of the stock assessment process and a great deal of effort is put into ensuring the data are of high quality. Therefore it is important that the analytical tools used at age, maturity and other calibration workshops are fit for purpose, delivering informative outputs for the workshop participants and the stock assessment process.
Scientific justification and relation to action plan:	This work relates to quality assurance of biological measurements as part of ICES' goal to advise on the sustainable use of living marine resources. Calibration workshops dealing with age and maturity estimation are funded and held under the auspices of the PGCCDBS. The main objectives of these important workshops are to decrease bias and improve the precision of age/maturity determinations between scientists from different laboratories. The end results are published in extensive ICES reports. However, there is a question of whether the right audience is reached by these reports. Moving beyond precision is increasingly common in calibration workshops and creating outputs better tailored to input for stock assessment models would greatly improve the application of the results. PGCCBDS (2010) also recognized that there is a need to review current methods of analysing data from calibration studies and consider issues such as agreement measures for the age of long-lived species and the best way to incorporate histologically validated samples for maturity staging comparisons. Finally, at a broader level, there is a large body of research on agreement statistics and methodology available from the field of medical statistics so it would be beneficial to transfer this knowledge into the fisheries arena.
Resource requirements:	No specific resource requirements beyond the need for members to prepare for and participate in the meeting.
Participants:	Participants should include a mixture of scientists with expertise in statistical methods, stock assessment, age reading and maturity staging.
Secretariat facilities:	None.

Financial:	Funding for external experts on the statistical methods may be required. The chairs seek to collaborate with NAFO to ease the invitation of experts outside the ICES system.
Linkages to advisory committees:	The workshop will link to ACOM through PGCCDBS.
Linkages to other committees or groups:	The outputs will be directly relevant to all age reading and maturity staging workshops. PGMed
Linkages to other organizations:	This topic links to the EU DCF, the COST (European Cooperation in the field of Scientific and Technical Research) Action FA0601 "Fish Reproduction and Fisheries" (FRESH) and the WebGR project (http://webgr.azti.es).

Annex 3: Agenda

Monday 13 October 2014

13.00-13.30 Introduction and welcome (Chairs)

13.30-14.15 Presentation of tor's and subgroups aims revisited (Chairs)

14.15-14.45 Any other business which participants feel should be discussed at WKSABCAL (Plenary session)

14.45-15.00 Agreeing on agenda (Plenary session)

15.00-15.30 Break

15.30-17.45 Subgroup presentations of results with respect to tor's (Subgroup Representatives)

17.45-18.00 Agreeing on tomorrows agenda (All)

Tuesday 14 October 2014

09.00-10.45 Work in subgroups on respective tasks (Subgroups)

10.45-11.15 Break

11.15-13.00 Work in subgroups on respective tasks (Subgroups)

13.00-14.00 Lunch

14.00-16.00 Work in subgroups on respective tasks (Subgroups)

16.00-16.30 Break

16.30-18.30 Status update from each subgroup (Plenary session)

Wednesday 15 October 2014

09.00-10.45 The CRR-Review paper; introduction and draft outline (Lotte Worsøe Clausen; All)

10.45-11.15 Break

11.15-13.00 Testing methods on known data (Subgroups)

13.00-14.00 Lunch

14.00-16.00 Richard McBride presentation and discussion of various methods for evaluation of calibration results (Plenary session)

16.00-16.30 Break

16.30-18.00 Testing methods on known data; visualizing outcomes (Plenary session)

Thursday 16 October 2014

09.00-10.45 Testing methods on known data; visualizing outcomes and evaluate application; write-up section for report (Subgroups)

10.45-11.15 Break

11.15-13.00 Work in subgroups to finalize drafts (Subgroups)

13.00-14.00 Lunch

14.00-16.00 Work in subgroups to finalize drafts (Subgroups)

16.00-16.30 Break

16.30-18.00 Overview of report sections; what is missing; assign authors to sections (Plenary session)

Friday 17 October 2014

09.00-10.45 Cleaning up, continue finalizing report and CRR chapter (Plenary session)

10.45-11.15 Break

11.15-12.00 Recommendations (Plenary session)

12.00-13.00 AOB (Plenary session)

Annex 4: List of reports reviewed for ToR a)

List of ICES age calibration workshops reviewed during the meeting for the use of different analytical methods, software and diagnostics

Workshop	Year	Analysis	Methods, software	Age parameters	Presentation tape
Report of the second Workshop on Age Reading of Red Mullet and Striped Red Mullet	2012	CV, MAD, PA, Relative bias, precision coefficient, mean age, first/second reading	TNPC software, Excel spreadsheet	Marginal–increment analysis, Otolith distances between nucleus to the edge and each ring	Histograms, bar plot, linear, box plot; tables; Otolith photos; smart arts
Report of the Roundnose grenadier (Coryphaenoides rupestris) Otolith Exchange Scheme 2011	2011	PA, MAD, CV, Relative bias, standart deviation, precision coefficient	*Not mentioned*	distance between nucleus and first translucent ring	Bar plots, bias plots, linear graphs; otolith photos; tables
Bay of Biscay sole otoliths exchange 2011	2011	PA, MAD, CV, Relative bias,	*Not mentioned*	Fish Length, male/female	Bar plots, bias plots, linear graphs; tables; otolith photos,
Report on otolith exchange of European hake	2011	Percentage of agreement, CV, APE,	WebGR, Excel spreadsheet (guus)	age	Fish length Histogram, box-whisker plots, otoliths photos, tables
Report of the Workshop on Age Reading of Dab (WKARDAB)	2010	Percentage of agreement, CV,	Excel spreadsheet (guus), TNPC software	Otolith radius	Fish length Histogram, CV%, % agreement and STDEV were plotted against MODAL age, otoliths photos, reference collection images, tables
Report of the Workshop on Age Reading of North Sea (IV) and Skagerrak-Kattegat (IIIa) Plaice (WKARP	2010	Percentage of agreement, CV, age bias, APE,	WebGR, TNPC software	Age	reference collection images, age composition Histogram, otoliths photos, bar plots (type of edge), bias plots, tables
Report of the Workshop on Age Reading of Red mullet Mullus barbatus and Striped mullet Mullus surmuletus (WKACM)	2009	Agreement; CV; percent agreement; standard deviation by modal age; bias; back-calculation of lengths; Kruskal–Wallis; marginal increment analysis	TNPC; Eltink *et al.*, (2000) spreadsheet	Age reading; ring radius; length frequency	Tables; linear plots; histograms; box-and-whisker plots

Workshop	Year	Analysis	Methods, software	Age parameters	Presentation tape
Workshop on Age Reading of European and American Eel (WKAREA)	2009	back-calculated growth rates	Excel spreadsheet	Age reading; Validation of age determination; Fish length	Not available
Report of the Workshop on Age Reading of Greenland Cod (WKARGC)	2009	modal length progression; Deviations from modal age estimates; CV; standard deviation; percent agreement	EFAN methodology (Eltink *et al.*, 2000, Eltink 2000); model was coded in Proc NLIN in SAS	Fish size; age readings; age group is described by three parameters: the mean length (m), the standard deviation (s) and the abundance (a); Age–length-Key	histograms; age bias plots, length distribution
Report of the Workshop on Age estimation of European hake (WKAEH)	2009	accuracy and precision; quality control and quality assurance (??); Percentage of readings agreement (PA); average percent error (APE); Coefficient of variation (CV); Wilcoxon signed rank tests; transition matrix	GIMP2.6 software; TNPC V4.1; Excel ad-hoc Workbook, "AGE COMPARATIONS.XLS" (Eltink *et al.*, 2000)	ALKs; age reading (blind and supervised); OTC validation; assumed a reference age for comparison	histograms (length frequency distribution); scatterplot; Box-whisker plots (for age estimate distribution analysis and ring-to-nucleus distances distribution analysis)
Workshop on Age reading of European anchovy (WKARA)	2009	von Bertalanffy growth equation	Workbook Age Reading comparisons of Eltink (2000) and Guidelines and tools for age read-ing comparisons (Eltink *et al.*, 2000).	microincrement daily growth as validation for first annual ring; age reading; MIA;	tables; age bias plots
Report of the Workshop on Age Reading of Mackerel (WKARMAC)	2009	CV, % Agreement, Bias	Age Comparison Worksheet (Eltink *et al.*, 2000).	age reading; otolith weight; length distribution	tables; age bias plots

WORKSHOP	YEAR	ANALYSIS	METHODS, SOFTWARE	AGE PARAMETERS	PRESENTATION TAPE
Report of the Workshop on Age Reading of Turbot (WKART)	2008	Percentage of agreement	ORACLE which is an improved version of the Guus spreadsheet		Tables, otolith photos,
Report of the Workshop on Age Determination of Redfish (WKADR)	2008	Percentage of agreement, CV, APE, von Bertalanffy	QCapture software		age bias plot, otolith Schematic drawing, otoliths drawings showing growth timing
Report of the Workshop on Age Reading of Baltic Herring (WKARBH)	2008	Percentage of agreement, CV,	Guus spreadsheet , validation (Increment width)		The grey tone profile (for validation), otolith photos, age bias plot
Report of the 2nd Workshop of aage reading of Flounder (WKARFLO)	2008	Percent of Agreement, standard deviation, MAD, CV, Average Percentage Error, relative bias, one-sample Wilcoxon rank sum test, t-test	SPSS 15.0	Fish length/age, staned otoliths, male/female	Scatterplots, age bias plot, otolith photos, tables
Report of the Workshop on Age Reading of North Sea Cod (WKARNSC)	2008	Percent of Agreement, standard deviation, CV, Wilcoxon's test, relative bias, MAD	Spreadsheet software	Marking the identified age structures on an agreed axis on digital images, otolith weight, fish length/ age	otolith photos, linear plots, histogram, gantt chart, tables
Report of the Workshop on Age Reading of Baltic Sprat (WKARBS)	2008	Percent of Agreement, CV, Wilcoxon signed ranks test, MAD, iter-reader bias	*Not mentioned*	age reading	tables, age bias plot
Report of the Work Shop on age estimation of sprat.	2004	Percentage of agreement, CV, Relative bias, age bias	*Not mentioned*	Validation (MIA and otolith weight frequency))	age bias plots, tables, otolith photos (daily and annual rings), growth plot, histogram
Workshop on Megrim Otolith Age Readings	2004	Percentage of agreement, CV, APE, Exploratory data analysis (EDA)	Excel spreadsheet (guus)	Fish length frequency	box-whisker plots, age bias plots, histogram, otoliths photos, tables

Workshop	Year	Analysis	Methods, software	Age parameters	Presentation tape
PLAICE AGE DETERMINATION EXCHANGE AND WORKSHOP 13-14 MAY 2003 OSTEND PRELIMINARY REPORT	2003	Percentage of agreement, CV, Relative bias,	Excel spreadsheet (guus)	Age	none
BLACK SCABBARD FISH (Aphanopus carbo) OTOLITH EXCHANGE (1998-1999)	1999	Age bias plot, standart deviation, PA, MAD, Bertalanffy curve	Spreadsheet developed by Guus Eltink	Fish length vs. age	Linear, box plot, scaterplot, histogram graphs; tables
Horse mackerel otolith workshop	1999	APE, standart deviation, CV, PA	Excel spreadsheet	Fish length vs. age	Histogram, age bias plot, linear, bar plot, notched box plot, graphs, Otoliths photos
Report of the Workshop on Mackerel Otolith Reading	1995	Percent of agreement, CV, Age bias plot, standard deviation, Wilcoxson's test, Modal Age Difference	*Not mentioned*	Marked-recaptured fish length/age, age from the reader/modal age	Regression lines, whisker plots, notched box plot, age bias plots tables,
Report of the Workshop on age reading of *Sebastes* spp.	1995	Age readers comparison, scale/otolith comparison	*Not mentioned*	Age of differently prepared otoliths	Scatterplots, tables
Final Results of the Mackerel (*Scomber scombrus*, L.) Exchange Programme in Otolith 1994	1994	PA, CV, APE, standard deviation, Age bias plot, Wilcoxson's test	Excel spreadsheet	Fish length vs. age	Linear, box plot whisker plots, bar plots, histogram graphs; tables
Final Results of the Mackerel (*Scomber* Exchange *scombrus*, L.) Programme in Otolith 1994.	1994	average length by age; percentage of agreement betwren readers; standard deviations; Wilcoxson's test	Excel spreadsheet	Age reading;	Notched Box and Whisker plot; age bias plots of each age reader against the modal age; histograms
Final Results of the Mackerel (*Scomber scombrus*, L.) Exchange Programme in Otolith 1994	1994	Percent of Agreement, CV, APE, standard deviation, Age bias plot, Wilcoxson's test	Excel spreadsheet	Fish length/age	Linear, box plot whisker plots, bar plots, histogram graphs, Age bias plot tables

Workshop	Year	Analysis	Methods, software	Age parameters	Presentation tape
Report of the blue whiting otolith reading Workshop	1992	PA, standard deviation, CV	*Not mentioned*	Fish length/weight; Otoliths length/height/ weight/diameter, Ring diameter	Linear, progression plot, box plot graphs; tables
Report of the blue whiting otolith reading Workshop	1992	Agreement (%);Standard deviation and coefficient of variation; ANOVA for differences in the mean ring diameter	*Not mentioned*	Age reading; Ring diameter; Otoliths diameter	Tables and graphs (scatterplots, regression line, boxplots, linear plot).
Report of the blue whiting otolith reading Workshop	1992	Percent of agreement, standard deviation, CV of mean age	*Not mentioned*	Fish length/weight; Otolith length/height/ weight/ring diameter	Linear, progression plot, box plot graphs; tables

Annex 5: CRR chapter 7

ICES Cooperative Research Report on Age Reading. Chapter 7: Statistical handling of uncertainty in age estimations

By WKSABCAL 2014; Editors Lotte Worsøe Clausen and Ernesto Jardim

Introduction

Inaccurate age determinations are widespread and negatively affect the accuracy of population dynamics studies and stock assessment outcomes. There are numerous cases in which ageing errors contributed to the overexploitation of a population or species (Campana, 2001). Underestimation of age results in overly optimistic estimates of growth and mortality rates and overestimation of age results in underestimation of growth. Thus such errors around the age-estimations need to be accounted for when conducting stock assessment (Punt *et al.*, 2008). Likewise, errors in maturity staging will cause erroneous maturity ogives, which in turn will impair the estimation of spawning-stock biomass and stock–recruitment relationships, and ultimately biomass reference points.

A large proportion of the calibration workshops dealing with age and maturity estimation, are held under the auspices of ICES. Moving beyond precision is increasingly common in these calibration workshops and creating outputs better tailored to the inputs of stock assessment models, would greatly improve the uptake of the results by assessment working groups.

The term “precision” is used to describe “agreement” or variability between readings/annotations of the same specimen by the same or different readers. The term “accuracy” is reserved to describe a comparison of ages or stages of maturity generated by readers with the true age or maturity stage for specimens.

When analysing outputs from age-calibration workshops it is imperative to acknowledge that in case there isn’t known-age material to calibrate the readers’ assessment of ages, then no true bias can be estimated. In the absence of a known-age reference collection, ageing consistency is the best that can be achieved (Campana *et al.*, 1995). Nevertheless, bias is often reported from age-calibration workshops, which shouldn’t be interpreted as real bias, when validated ages are not available. Thus bias in this respect is more an expression of the ‘skewness’ of data around a modal or likely value.

The reports from calibration workshops generally give very thorough results, which facilitate getting a common interpretation of the age structures of the otoliths of a given species. The dissemination of the results supports the judgement, by stock assessment scientists, of the quality of the age distributions used in the assessment. However, considering the very limited inclusion of workshop’s results into stock assessments, the questions are whether the right audience has been reached by these reports, with appropriate data formats, and if the stock assessment models are prepared to include them. The discussions and conclusions in this Chapter specifically targeted such intentions.

Statistical methods for analysing output from calibrations

31 ICES reports published between 1992 and 2012 on age reading and calibration exercises were reviewed with the purpose of identifying the applied methods for analyses of calibration data. The majority used methods to compare age readings

such as Percentage of agreement (PA), average percentage error (APE), and Coefficient of variation (CV). After 2000, the majority of reports used the Guus Eltink spreadsheet, i.e. the Workbook Age Reading comparisons of Eltink (2000) and Guidelines and tools for age reading comparisons (Eltink *et al.*, 2000). These are useful tools for a general understanding of the uncertainty in age determinations for a stock or for an age reader. However, these estimates are not suitable for incorporation of age reading uncertainty in stock assessments. Different approaches have been taken to account for age reading error in stock assessments. Several studies compare stock assessment results based on different age scenarios. These scenarios can reflect inter-reader variability or differences between observed and true ages. They can be calculated based on alternative ALK's, CAA's or growth models.

Most of the recent studies addressing this issue quantify an ageing error matrix (AEM) to use in the stock assessment model. The elements of the AEM are the probabilities that a sampled fish of true age class *a* is assigned to one of the observed age classes. Although the basic concept is the same, the approaches taken to estimate the probabilities differ. First, the statistical models differ, both in functional forms and distributional assumptions. Second, what is taken to be "true age" is usually not really the true age. Examples include simulated true age, known age (based on mark-recapture studies), nearest integer to mean age across readers, modal age, otolith age (with observed age based on other CS), or one preparation method (with observed age based on other preparation methods). A readability score has been seen used as a factor in their statistical model to estimate the probabilities of the AEM (Candy *et al.*, 2012). This is of particular interest with regard to the 3pt grading system recommended by PGCCDBS and WKNARC 2011 (ICES 2011). Readability score, such as the 3pt grading system recommended by PGCCDBS and WKNARC 2011 or the 5 class system used in Australia is a (subjective) variable. It is not a probability or error estimate and therefore not directly applicable in a stochastic assessment model. Thus a readability score should not be applied as a selection criterion for age data included in the stock assessment, because this may cause bias as readability score can be correlated with age (Candy *et al.*, 2012) and it is expected to be correlated with growth rate. Candy *et al.*, (2012) included readability score in their statistical model to estimate the probabilities for an AEM. This approach requires the selection of a readability score to produce the predicted probabilities for the AEM to be used in the assessment model. There are two concerns about this approach which should be taken into consideration when using this method. First, age and readability score are both included in the statistical model, but the results suggest collinearity between these 2 variables. Second, the selection of a readability score to produce an AEM for the assessment model implies that the AEM is not representative of the observed (variability of) readability scores.

Of the 17 methods reported in the literature (Table 7.1), all can be classified as one of the following: a) identification of bias between age readers or a reference collection; b) estimate of precision among or within age readers; c) diagnostic of age reading differences; and d) preparation of ageing error matrix for use in stock assessment models.

Table 7.1. Methods which can be used to evaluate age calibration studies; characteristics of the methods and strengths and weaknesses are given. Colour code highlights methods considered key products of an age calibration study.

	Method	Descriptive statistics	Statistical test	One single number	Visual method	Model-based approach	Precision	Bias	Data requirements	Diagnostics	Strength	Weakness	Observations
1	ABP - Age-bias plot			1	1			1	age readings		Easily interpreted	visual, not a statistical test	
2	TS - Tests of symmetry		1	1				1	age readings			statistical test not picking up non-monotonic aging problems	
3	PTT - Paired t-test		1	1				1	age readings		Easily interpreted		parametric test
4	WPRT - Wilcoxon paired rank test		1	1				1	age readings		Easily interpreted		non-parametric test
5	APE - Average Percentage Error	1		1			1		age readings		Easily interpreted		sensitive to outliers
6	CV - Coefficient of Variation	1		1			1		age readings		Easily interpreted		sensitive to outliers
7	MSD - Mean Square Deviation	1		1			1		age readings		Easily interpreted		
8	CCC - Concordance Correlation Coefficient	1		1			1		age readings		Easily interpreted		
9	TDI - Total Deviation Index	1		1			1		age readings		Easily interpreted		
10	MAD - Modal Age Difference	1		1			1		age readings		Easily interpreted		
11	PA - Percentage Agreement	1		1			1		age readings		Easily interpreted	poor because it is sensitive to range of ages used in the analysis	
12	Rho - Average Spearman's Rho		1	1			1		age readings		Easily interpreted		just a correlation coefficient
13	W - Kendal's Coefficient of Concordance		1	1			1		age readings		Easily interpreted		
14	Tau - Average Tau		1	1			1		age readings		Easily interpreted		
15	MAOI - Model Analysis of Otolith Increments (e.g. with mixed effects models)		1			1		1	each ring of each otoliths marked by readers are required (WebGR)	1		complex output	
16	VAOI - Visual Analysis of Otolith Increments (e.g. using photoshop)	1			1				each ring of each otoliths marked by readers are required (WebGR)	1			use layers in photoshop to visualize age readings and the mixed effects model results
17	AOI - Analysis of Otolith Increments (e.g. using simple statistics)	1		1			1	1	each ring of each otoliths marked by readers are required (WebGR)	1	summary across readers and otoliths		Requires further development
18	AREM - Age Readings Error Matrix					1	1		age readings				bridge towards stock assessment; matrix can use observed or modelled data

Strengths and weaknesses of each listed method

Identification of bias is one of the most important products of an age calibration exercise, since it indicates that the age readers are interpreting the growth increments differently. Unless one of the age readings is based on a validated reference collection, it can be difficult to determine which of the age readings (if either) is more accurate. Method (1), the age bias plot, is a widely used method for visually identifying bias. Methods (2)-(4) are statistical counterpoints to the age bias plot. The statistical methods have the advantage of being quantitative. The age bias plot tends to be more sensitive. Both measures of bias are useful in an age calibration or quality control (QC) exercise.

Precision is a measure of consistency or reproducibility within or among age readers. Precision is often characteristic of experienced age readers, since they tend to be very consistent in their interpretations. However, precision is not a measure of accuracy, and precise ageing is not necessarily accurate. Most of the methods reported in Table 7.1 are measures of precision (Methods 5-14). Precision estimates can be expressed by a single number (e.g. CV=5%) and thus are easily understood. CV (Method 6) and APE (Method 5) are widely used in the literature, and thus are readily compared among age calibration exercises. They also have the advantage of being relatively insensitive to the age range in the study, unlike the simple percentage agreement. It is recommended to include at least one measure of precision, either CV or APE, as a useful product of an age calibration or QC exercise.

It is important to note that all of the bias tests (both the age bias plot and the statistical tests of symmetry) are limited to pairwise comparisons of age readings; it is not possible to compare more than two readings at a time. In contrast, the precision measures, diagnostics and age error matrices can all be based on as many age readings as are available.

Age readers are often most interested in determining the cause of any systematic differences in age determination (bias) with other age readers, since it may be possible to correct the error if the source is known. These methods may be referred to as diagnostic methods (Methods 15-17). The use of multiple layers in Photoshop (or some other imaging program), with one layer per age reader, allows rapid comparison of growth increment interpretations across multiple age readers. This approach is visual, and not quantitative. Alternatively, simple descriptive statistics can be computed and analysed or else mixed effect models applied to growth increment width data providing a quantitative diagnostic of systematic growth increment interpretation differences. Although these methods likely provide too much detail for stock assessment experts, they represent useful products of an age calibration exercise.

Age error matrices (Method 18) provide a quantitative summary of age reading error. These matrices, whether empirical or model-based (sensu Punt *et al.*, 2008), report misreading errors as proportion of an age group mis-aged as other ages. Empirical age error matrices are readily calculated from age-frequency tables, while model-based matrices require specialized software. Age error matrices are of particular value to stock assessment clients, since they can be incorporated into age-structured stock assessment models to correct for routine and unavoidable ageing error. Thus they are an important product of an age calibration exercise.

Test of methods on a known dataset

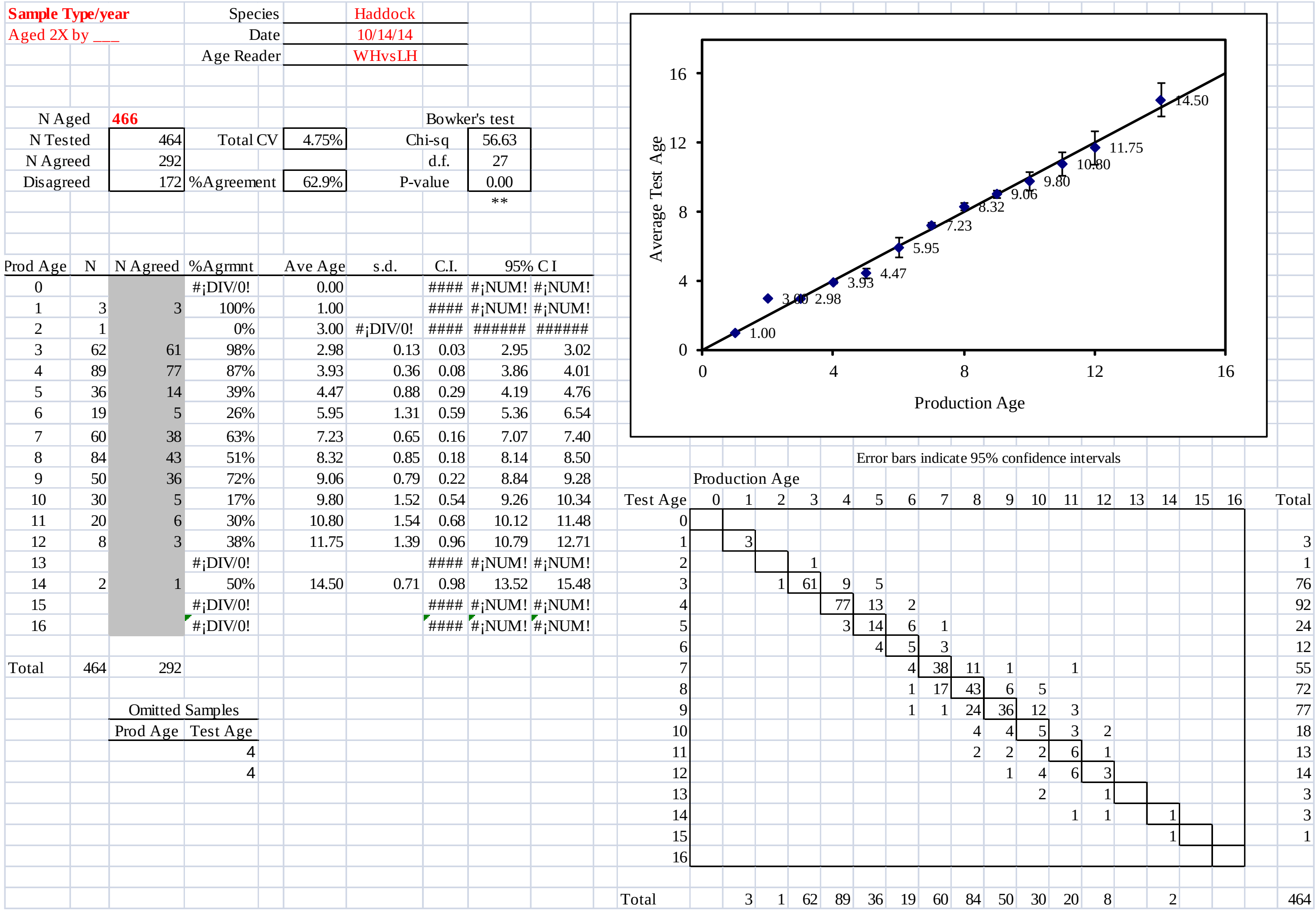
The quantitative parameters were tested on known datasets on two species. A haddock age test data were taken from an international age calibration study involving 7 expert ageing laboratories and age-validated otoliths. Two subsets of data were used in the examples shown below; the first compares two sets of age readings which are very similar, while the second compares age readings where one set of age readings shows substantial bias. The analysis was done using the software tool from NOAA:

http://www.nefsc.noaa.gov/fbp/age-prec/

In Figure 7.1, the age bias plot shows no appreciable bias between the two sets of age readings, although there is a very slight tendency for the test age to underage at ages 5-6 and overage at ages 7-8. The Bowker's test of symmetry may detect this difference, since the test result is statistically significant (P=0.00). As such, it indicates that this test is highly sensitive to very small differences. The CV of 4.75% indicates that there is reasonably good precision between the two sets of age readings, since CV values less than 5 are considered relatively precise (Campana, 2001). The APE is not shown, but CV is mathematically equal to 1.41 times the APE. Figure 7.2 then shows the age frequency table that results from this comparison, which is later used to generate the age error matrix. Also shown is an additional test of symmetry (the Hoenig test). This test also shows significant differences between the age readings, but of smaller magnitude than the Bowker's test. The final product is the ageing error matrix, which is shown in Figure 7.3. This matrix shows the proportion of each test age (the reference age) mis-aged as other ages. Therefore, the sum of each row is 1, equal to 100%.

The second test set used data with considerable bias. Figure 7.4 illustrates how the age bias plot indicates substantial underageing of the test age relative to production ages >5 yr. The Bowker's test P value of 0.00 confirms that the underageing is significant. The CV of 25.21% is also very high, but in this case, a measure of precision is much less important given the very high bias between the two sets of age readings. In Figure 7.5, it is evident that the age frequency table as well as the highly significant P value of 0.000 from the Hoenig test, confirms such bias and finally Figure 7.6 shows how the ageing error matrix from this comparison appear.

Precision.xltx



Sample Type/year	Species	Haddock	
Aged 2X by ___	Date	10/14/14	
	Age Reader	WHvsLH	

N Aged	466			Bowker's test	
N Tested	464	Total CV	4.75%	Chi-sq	56.63
N Agreed	292			d.f.	27
Disagreed	172	%Agreement	62.9%	P-value	0.00
					**

Prod Age	N	N Agreed	%Agrmnt	Ave Age	s.d.	C.I.	95% C I	
0			#¡DIV/0!	0.00		####	#¡NUM!	#¡NUM!
1	3	3	100%	1.00		####	#¡NUM!	#¡NUM!
2	1		0%	3.00	#¡DIV/0!	####	######	######
3	62	61	98%	2.98	0.13	0.03	2.95	3.02
4	89	77	87%	3.93	0.36	0.08	3.86	4.01
5	36	14	39%	4.47	0.88	0.29	4.19	4.76
6	19	5	26%	5.95	1.31	0.59	5.36	6.54
7	60	38	63%	7.23	0.65	0.16	7.07	7.40
8	84	43	51%	8.32	0.85	0.18	8.14	8.50
9	50	36	72%	9.06	0.79	0.22	8.84	9.28
10	30	5	17%	9.80	1.52	0.54	9.26	10.34
11	20	6	30%	10.80	1.54	0.68	10.12	11.48
12	8	3	38%	11.75	1.39	0.96	10.79	12.71
13			#¡DIV/0!			####	#¡NUM!	#¡NUM!
14	2	1	50%	14.50	0.71	0.98	13.52	15.48
15			#¡DIV/0!			####	#¡NUM!	#¡NUM!
16			#¡DIV/0!			####	#¡NUM!	#¡NUM!
Total	464	292						

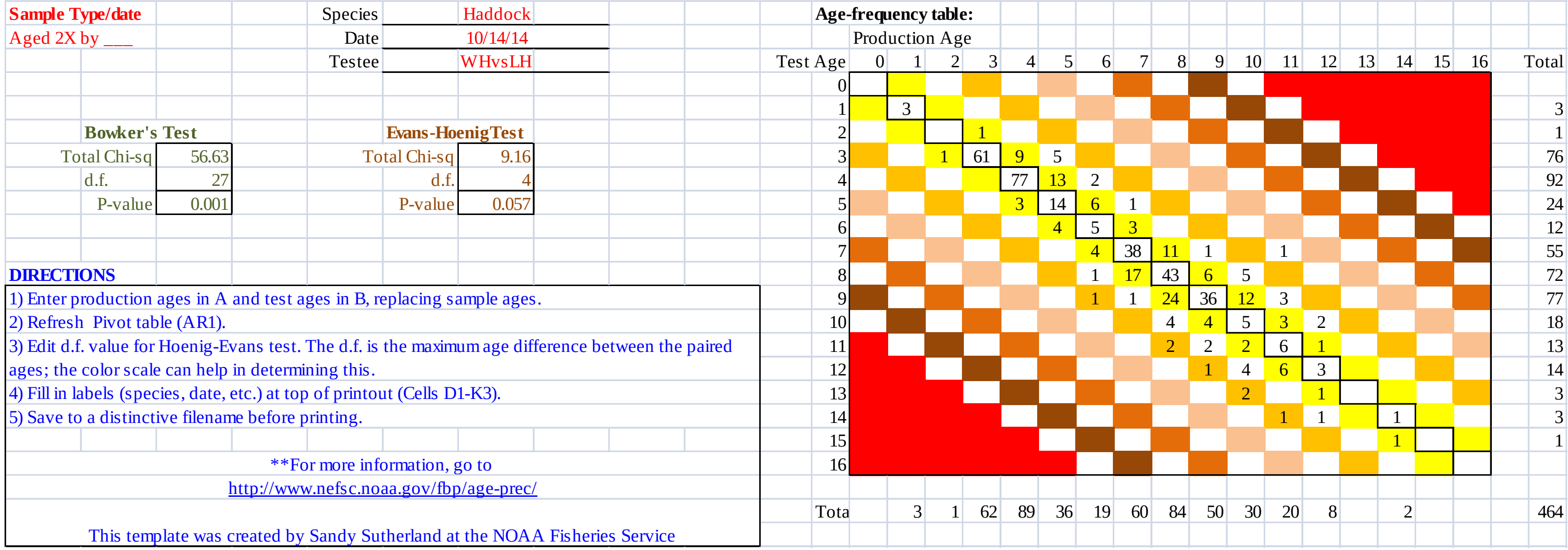
Omitted Samples

Prod Age	Test Age
	4
	4

Error bars indicate 95% confidence intervals

	Production Age																	
Test Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1		3																3
2				1														1
3			1	61	9	5												76
4					77	13	2											92
5					3	14	6	1										24
6						4	5	3										12
7							4	38	11	1		1						55
8							1	17	43	6	5							72
9							1	1	24	36	12	3						77
10									4	4	5	3	2					18
11									2	2	2	6	1					13
12										1	4	6	3					14
13											2		1					3
14												1	1		1			3
15															1			1
16																		
Total		3	1	62	89	36	19	60	84	50	30	20	8		2			464

Figure 7.1. Age bias plot for Dataset 1 without detectable bias



Sample Type/date
Aged 2X by ___

Species	Haddock
Date	10/14/14
Testee	WHvsLH

Bowker's Test

Total Chi-sq	56.63
d.f.	27
P-value	0.001

Evans-Hoenig Test

Total Chi-sq	9.16
d.f.	4
P-value	0.057

DIRECTIONS

1) Enter production ages in A and test ages in B, replacing sample ages.
2) Refresh Pivot table (AR1).
3) Edit d.f. value for Hoenig-Evans test. The d.f. is the maximum age difference between the paired ages; the color scale can help in determining this.
4) Fill in labels (species, date, etc.) at top of printout (Cells D1-K3).
5) Save to a distinctive filename before printing.

**For more information, go to
http://www.nefsc.noaa.gov/fbp/age-prec/

This template was created by Sandy Sutherland at the NOAA Fisheries Service

Age-frequency table:

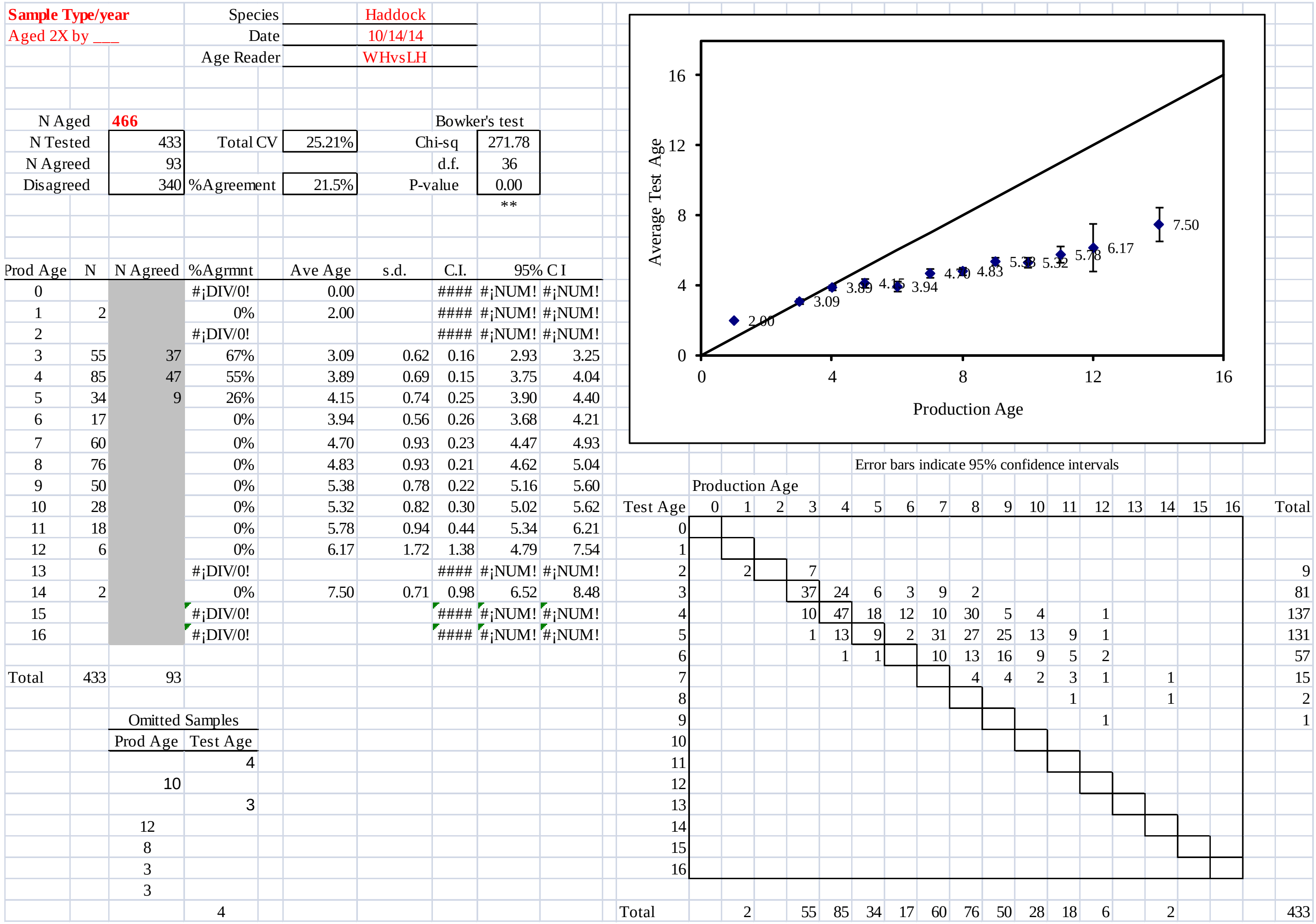
	Production Age																	
Test Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1		3																3
2				1														1
3			1	61	9	5												76
4					77	13	2											92
5					3	14	6	1										24
6						4	5	3										12
7							4	38	11	1		1						55
8							1	17	43	6	5							72
9							1	1	24	36	12	3						77
10									4	4	5	3	2					18
11									2	2	2	6	1					13
12										1	4	6	3					14
13											2		1					3
14												1	1		1			3
15															1			1
16																		
Tota		3	1	62	89	36	19	60	84	50	30	20	8		2			464

Figure 7.2. Age frequency table from the comparison and an additional test of symmetry (Hoenig test) on the data without detectable bias.

AREM

	Production Age(WH)																	
Test Age(LH)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1		1.00																3
2				0.02														1
3			1.00	0.98	0.10	0.14												76
4					0.87	0.36	0.11											92
5					0.03	0.39	0.32	0.02										24
6						0.11	0.26	0.05										12
7							0.21	0.63	0.13	0.02		0.05						55
8							0.05	0.28	0.51	0.12	0.17							72
9							0.05	0.02	0.29	0.72	0.40	0.15						77
10									0.05	0.08	0.17	0.15	0.25					18
11									0.02	0.04	0.07	0.30	0.13					13
12										0.02	0.13	0.30	0.38					14
13											0.07		0.13					3
14												0.05	0.13		0.50			3
15															0.50			1
16																		
Total		3	1	62	89	36	19	60	84	50	30	20	8		2			

Figure 7.3. The final outcome: the Ageing Error Matrix



Sample Type/year	Species	Haddock	
Aged 2X by ___	Date	10/14/14	
	Age Reader	WHvsLH	

N Aged	466			Bowker's test	
N Tested	433	Total CV	25.21%	Chi-sq	271.78
N Agreed	93			d.f.	36
Disagreed	340	%Agreement	21.5%	P-value	0.00
					**

Prod Age	N	N Agreed	%Agrmnt	Ave Age	s.d.	C.I.	95% C I	
0			#¡DIV/0!	0.00		####	#¡NUM!	#¡NUM!
1	2		0%	2.00		####	#¡NUM!	#¡NUM!
2			#¡DIV/0!			####	#¡NUM!	#¡NUM!
3	55	37	67%	3.09	0.62	0.16	2.93	3.25
4	85	47	55%	3.89	0.69	0.15	3.75	4.04
5	34	9	26%	4.15	0.74	0.25	3.90	4.40
6	17		0%	3.94	0.56	0.26	3.68	4.21
7	60		0%	4.70	0.93	0.23	4.47	4.93
8	76		0%	4.83	0.93	0.21	4.62	5.04
9	50		0%	5.38	0.78	0.22	5.16	5.60
10	28		0%	5.32	0.82	0.30	5.02	5.62
11	18		0%	5.78	0.94	0.44	5.34	6.21
12	6		0%	6.17	1.72	1.38	4.79	7.54
13			#¡DIV/0!			####	#¡NUM!	#¡NUM!
14	2		0%	7.50	0.71	0.98	6.52	8.48
15			#¡DIV/0!			####	#¡NUM!	#¡NUM!
16			#¡DIV/0!			####	#¡NUM!	#¡NUM!
Total	433	93						

Omitted Samples

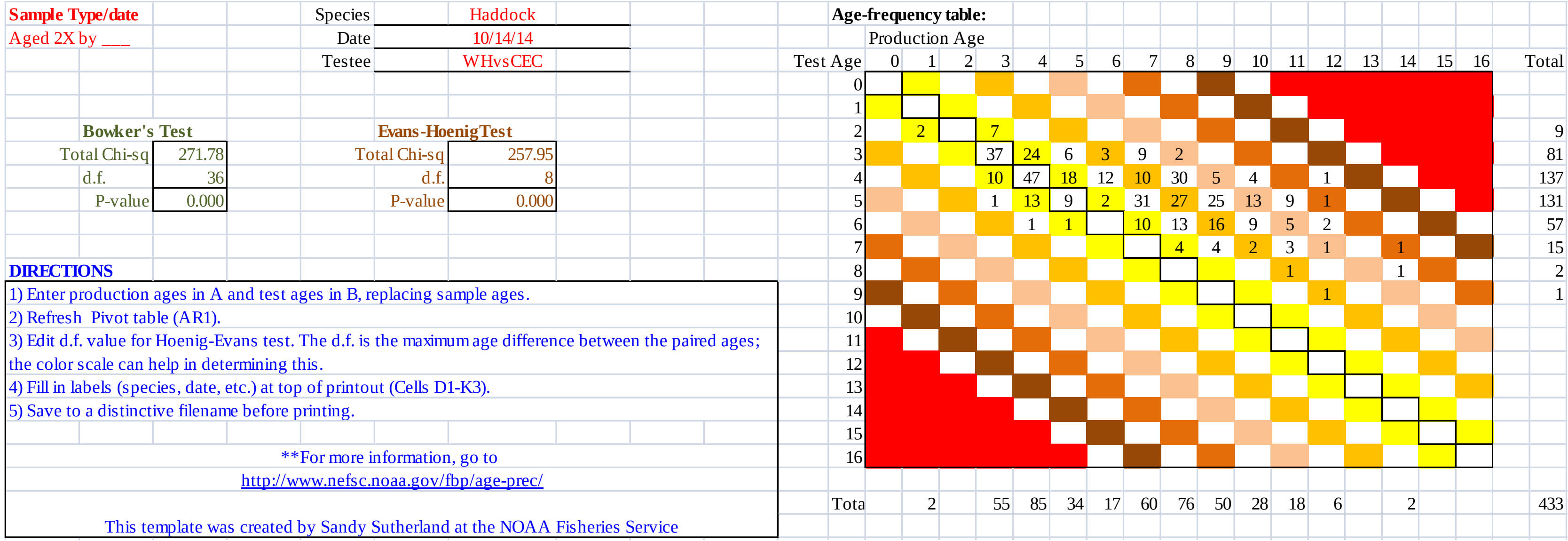
Prod Age	Test Age
	4
10	
	3
12	
8	
3	
3	
	4

Error bars indicate 95% confidence intervals

	Production Age																	
Test Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1																		
2		2		7														9
3				37	24	6	3	9	2									81
4				10	47	18	12	10	30	5	4		1					137
5				1	13	9	2	31	27	25	13	9	1					131
6					1	1		10	13	16	9	5	2					57
7									4	4	2	3	1		1			15
8												1			1			2
9													1					1
10																		
11																		
12																		
13																		
14																		
15																		
16																		
Total		2		55	85	34	17	60	76	50	28	18	6		2			433

Figure 7.4. Analysis of the dataset with substantial bias.

Symmetry2kinds.xltx



Sample Type/date	Species	Haddock
Aged 2X by ___	Date	10/14/14
	Testee	WHvsCEC

Bowker's Test		**Evans-Hoenig Test**	
Total Chi-sq	271.78	Total Chi-sq	257.95
d.f.	36	d.f.	8
P-value	0.000	P-value	0.000

DIRECTIONS

1) Enter production ages in A and test ages in B, replacing sample ages.

2) Refresh Pivot table (AR1).

3) Edit d.f. value for Hoenig-Evans test. The d.f. is the maximum age difference between the paired ages; the color scale can help in determining this.

4) Fill in labels (species, date, etc.) at top of printout (Cells D1-K3).

5) Save to a distinctive filename before printing.

**For more information, go to
http://www.nefsc.noaa.gov/fbp/age-prec/

This template was created by Sandy Sutherland at the NOAA Fisheries Service

Age-frequency table:

	Production Age																	
Test Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1																		
2		2		7														9
3				37	24	6	3	9	2									81
4				10	47	18	12	10	30	5	4		1					137
5				1	13	9	2	31	27	25	13	9	1					131
6					1	1		10	13	16	9	5	2					57
7									4	4	2	3	1		1			15
8												1			1			2
9													1					1
10																		
11																		
12																		
13																		
14																		
15																		
16																		
Tota		2		55	85	34	17	60	76	50	28	18	6		2			433

Figure 7.5. Symmetry test on the data with substantial bias.

AREM

Production Age(WH)

Test Age(CEC)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
0																		
1																		
2		1.00		0.13														9
3				0.67	0.28	0.18	0.18	0.15	0.03									81
4				0.18	0.55	0.53	0.71	0.17	0.39	0.10	0.14		0.17					137
5				0.02	0.15	0.26	0.12	0.52	0.36	0.50	0.46	0.50	0.17					131
6					0.01	0.03		0.17	0.17	0.32	0.32	0.28	0.33					57
7									0.05	0.08	0.07	0.17	0.17		0.50			15
8												0.06			0.50			2
9													0.17					1
10																		
11																		
12																		
13																		
14																		
15																		
16																		
Total		2		55	85	34	17	60	76	50	28	18	6		2			

Figure 7.6. The Ageing Error Matrix on the second dataset.

Although the NOAA software tool was used to generate all of the output shown above, it should be possible to generate all of these products from WebGR or other software after appropriate revision.

Additional analysis based on distances between identified otolith structures

The analysis of growth increments can be very informative about the reasoning behind the attribution of a specific age. As such, the comparative analysis across readers can be used to evaluate the ageing process and detect severe differences in the interpretation of the otolith. The information generated will be valuable to correct errors or deepening into problems detected in bias and precision analysis.

WebGR stores information about the distance between marks, which constitutes the tool to identify rings, by themselves the basis to attribute an age to an individual. Using such information allows several analyses to be carried out, as mentioned before. Namely, visual analysis of the marks in each otolith, using image editing tools; statistics to extend this analysis and allow a wider assessment of the consistency between readers; and/or mixed effects models to test if readers are consistently interpreting the otolith rings. This section is focused on the second, while an example of using mixed effects models can be seen in WKSIBCA report (ICES, 2014).

Figure 7.7 presents the increment widths by mark for 4 readers and 4 otoliths. This analysis is numerically similar to overlaying the images of each reader for a specific otolith to assess their consistency. It allows the analysis of consistency between readers in interpretation of the structures counted.

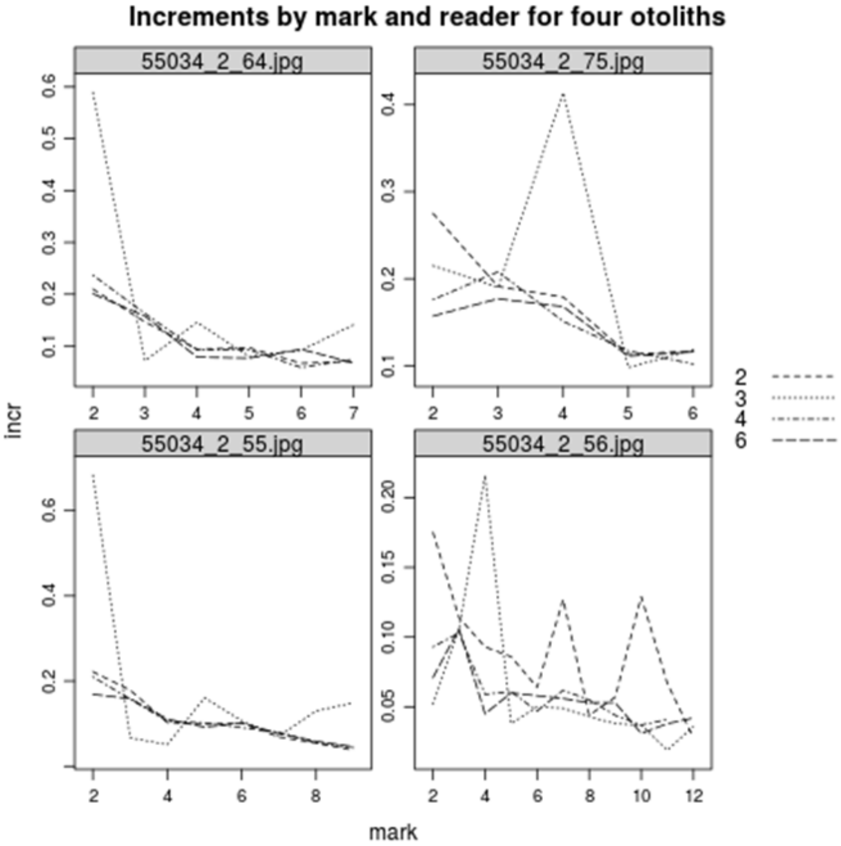


Figure 7.7. Increments by mark and reader for 4 selected otoliths

Figure 7.8 shows the coefficient of variation of the increment widths by mark that each reader set across otoliths. This analysis is related with reader's internal consistency. It shows how con-sistent each reader was in the interpretation of rings. In this case it shows that reader 3 was off in the initial rings/marks, while reader 2 was off on the older ages. The other readers seem consistent and present an increasing precision for older ages.

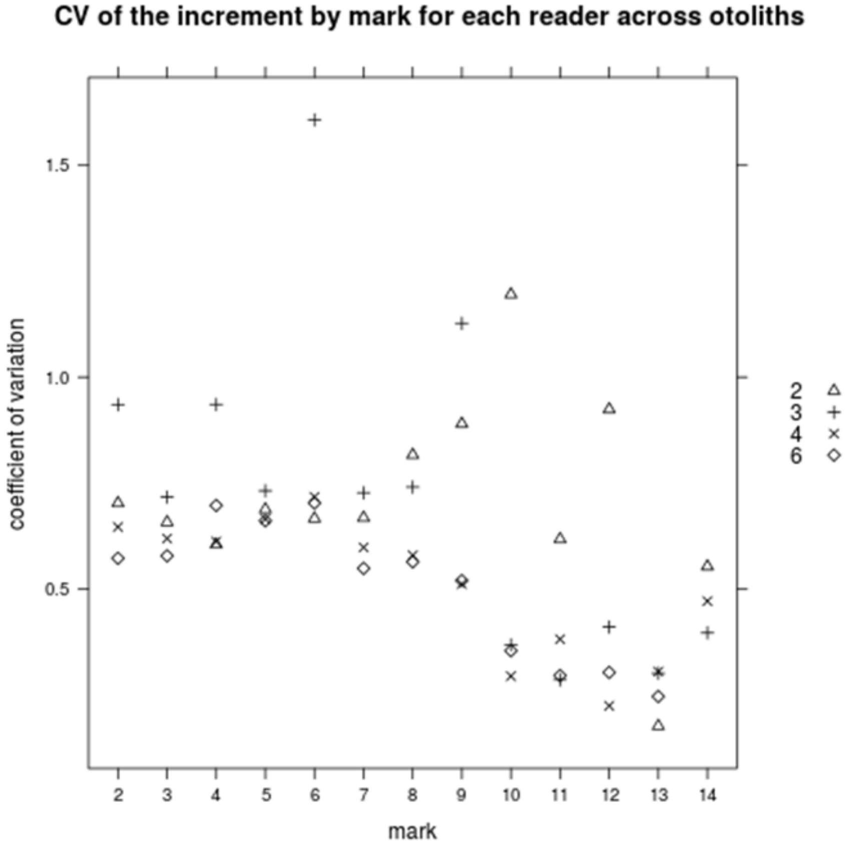


Figure 7.8. Coefficient of variation of the increment widths by mark that each reader set across otoliths

Figure 7.9 shows the average deviance of the increment width each reader gave to each mark from the median increment for that mark, across readers. It represents consistency between readers once that, in a perfect situation, all readers should identify the rings in the same place in the otolith, which would result in the same distance between marks, and deviance 0.

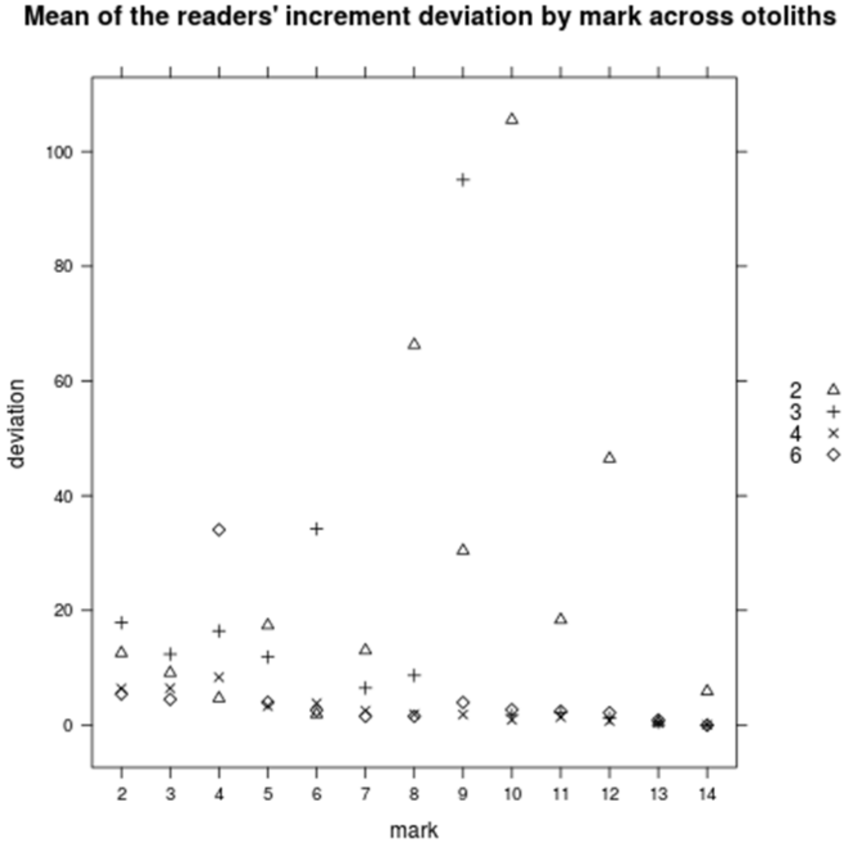


Figure 7.9. Average deviance of the increment width each reader gave to each mark from the median increment for that mark, across readers.

In all of these analyses, interpretation errors made at an early stage (i.e. at Mark 2) would be expected to propagate through all subsequent marks. Further work is required to better under-stand the implications of this error propagation, and how it would affect the interpretation of the analysis.

It is important to note that this analysis is based on a dataset that is not fully appropriate to it. To carry on this type of analysis the distances between rings should be measured along the same axis and, as much as possible, a linear axis. Currently the workshops ran with WebGR does not use this standard, which increases the variability by mark. Additionally, the WebGR design does not use the centre of the nucleus, the first mark is counted as one age, which invalidates the analysis of the first increment, a recognized major source of error. All these improvements can be made without having to heavily redesign WebGR.

Existing software

Direct ageing studies require specific software tools to be used in calibration exchanges and workshops. Nowadays most of these studies are based on digital images of calcified structures where readers can annotate their readings. This procedure contributes to the standardization of ageing interpretation criterion among readers (as for example the misinterpretation of false checks or the differences in the position of the first annulus), and at the same time these annotated images allow to measure age increments.

Other tools that have been used are electronic forms or integrated databases containing sampling information and where ages or growth increments are stored. When non-integrated software is used, these forms have been developed in spreadsheets that are used for further statistical analysis in order to obtain precision, accuracy (relative or absolute), bias analysis and other outputs. In the case of integrated databases, none of the analysed programs implemented routines for statistical analysis but all of them have extraction routines for exporting the data to statistical programs.

Integrating images and ageing data analysis in the same software may reduce processing time and avoid errors due to data handling; however, available software is currently serving both requirements separately.

Available software was characterized according to two categories:

1) Age or maturity calibration exercises software which are currently based on electronic images of calcified structures or gonads

2) Software for statistical analysis of the results of the calibration exercises.

In the first case existing software known to the WK is described in Table 7.2. All these software allow the calibration of images, the integration of annotated layers from different readers in each of the images and to take real distances between consecutive marks.

The second parameter revised was if the software handles the workshop in web based interface or has to be installed locally and then the results of each reader has to be sent to the WS coordinator. Other parameter that was taken into account was if the software is commercial, not allowing users to modify the software and with an economic cost, or if it is open source, allowing changes to adapt the software to data and without cost.

The integration of a database able to handle the associated data to the images was analysed too, resulting that only scientific/industrial image analysis software and WebGR have integrated databases that can store that kind of data. This Table 7.2 finally details how easy is to measure age increments and if the software is compatible with the multilayer-TIF format, which is the most commonly employed when working with images that need to handle various annotated layers.

Table 7.3 shows the available tools for statistical computing and graphics that were available and known by at the time of writing this Chapter, with a description of the statistical methods computed by software.

Software for age and maturity calibration exercises	**Characteristics**					
All these programs allow to calibrate images, annotate, measure	Web based	Type of license	Main purpose	Integrated Data Base (sample data, Images, annotated layers, ages)	Multilayer TIF Format	Easiness for measuring age increments
WebGR	Yes	Open source	Designed for Calibration Exercises	Integrated	No	Automatic
Adobe Photohop	No	Commercial	Photo editor	No	Yes	Time consuming
PaintShop Pro	No	Commercial	Photo editor	No	Yes	Time consuming
Gimp	No	Open source	Photo editor	No	No	Time consuming
ImageJ (TreeRings)	No	Open source	Image analysis	Integrated	Yes	Automatic
Visilog (TNPC)	No	Commercial	Image analysis	Integrated	No	Automatic
Nis-Elements D	No	Commercial	Image analysis	Integrated (only marked layers)	No	Automatic
Image-Pro (Otolith fish ageing)	No	Commercial	Image analysis	Integrated	Yes	Automatic
Image-Pro (Age & Shape)	No	Commercial	Image analysis	Integrated	Yes	Automatic

Table 7.2

Software for calibration results analysis	Framework	Source	Data handling	Computed statistics and graphics														
				MSD	CCC	TDI	CP	MAD	PA	APE	CV	AREM	Age Bias plot	Text of simetry	Rho	W	Tau	Mixed effects models
Age Reading Comparisons (G. Etkin, 2000, excel spreadsheet)	MS-Excel		Easy to use. Prone to errors	Yes				Yes	Yes		Yes		Yes	Wilcoxon signed-rank test				
NOAA –NEFSC excel workbooks (Templates for Calculating Ageing Precision)	MS-Excel	http://www.nefsc.noaa.gov/fbp/age-prec/index.html	Easy to use. Prone to errors						Yes		Yes		Yes	McNemar's, Evans&Hoening's, Bowke's				
Agreement	R-package	http://cran.r-project.org/	Knowledge of R required	Yes	Yes	Yes	Yes											
agRee	R-package	http://cran.r-project.org/	Knowledge of R required		Yes													
irr	R-package	http://cran.r-project.org/	Knowledge of R required		Yes											Yes		
KappaGUI	R-package	http://cran.r-project.org/	Knowledge of R required		Yes											Yes		
FSA	R-package	http://fishr.wordpress.com/fsa/	Knowledge of R required					Yes	Yes	Yes	Yes		Yes	McNemar's, Evans&Hoening's, Bowke's				
nwfscAgeingError (AGEMAT.exe) (Punt et al., 2008)	R-package	https://r-forge.r-project.org/R/?group_id=1316	Knowledge of R required									Yes						
lme4	R-package	http://cran.r-project.org/	Knowledge of R required															Yes

MSD= Mean Square Deviation, CCC= Concordance Correlation Coefficient, TDI= Total Deviation Index, CP= Coverage Probability, MAD= Modal Age Difference, PA= Percentage Agreement APE= Average Percentage Error, CV= Coefficient of Variation, AREM= Age Readings Error Matrix, Rho= Average Spearman's Rho, W= Kendal's Coefficient of Concordance, Tau= Average Tau

Table 7.3.

Guidelines for data summaries and analysis outputs from calibration workshops

A calibration workshop has the basic purpose, as part of the quality assurance procedure, to identify sources of errors and inconsistencies among laboratories in stock-specific biological measurements, quantify these errors and ultimately include them in stock assessment. In this view the results of these exercises, published in extensive ICES reports, have the purpose of reaching both the personnel observing and classifying the biological structures and the scientists involved in the estimation of stock biological parameters.

Different output formats from age reading workshops from submitting tables of reader raw data over ageing error matrices to summary statistics of variation and bias are appropriate depending on the audience of the report and results.

Age reading error matrices provide an intermediate level of detail useful at routine stock assessment exercises, however disaggregated data at individual fish and reader level with additional spatially and temporally resolved covariates would probably be preferable in many benchmark situations where modelling of data quality is an issue.

For data limited approaches an age reading error matrix may not fill the needs for modelling growth, due to a potential correlation between observed age and size.

The following methods/analysis are recommended to be run by age calibration workshops:

- To access bias
  - ABP - Age-bias plot
  - TS - Tests of symmetry
- To access precision
  - APE - Average Percentage Error
  - CV - Coefficient of Variation
- As diagnostics for problems found by the previous analysis
  - Analysis of otolith increments, both through image layers and statistically
- As output to stock assessment groups
  - AREM - Age Readings Error Matrix

It is important to note that if validated material is unavailable, bias cannot be computed and the analysis is limited regarding its assessment. It is thus recommended that regardless of the scope of the calibration workshop, an effort should be made to validate the age reading of the species/stock under consideration.

Additionally, the raw data from a calibration workshop should be documented by ICES and made available for those interested in running alternative estimation procedure or for more thorough analyses that may be available in future; this would assure that historical datasets can be analysed with the same method and past and present age reading performance can be compared and corrections could be integrated in long-term assessment datasets. The portfolio of analysis/methods identified in the present Chapter allows a thorough perspective of the precision and bias/deviation associated with the ageing process of the stock, which should constitute an important knowledge set for those interested, like age or maturity analysts, stock assessment working groups, etc.

With regards to the software packages available, WebGR is the most suitable for running workshops, while FSA is the most complete for the analysis of the age reading results. None of these packages can run all the methods recommended here although they can be further developed to accommodate most of them, where in particular WebGR has the potential to develop and implement the methods recommended.

General conclusions and future perspectives

Within ICES, the Assessment and Benchmark WG have in general hitherto been less concerned with explicit age reading errors however assuming that data had reading errors but were without bias. Defining the best format for age reading errors as input to stock assessment and benchmarks has been indicated to be a topic for further research. An immediate useful input would be an estimate of which ages may be confounded. The age from which strong confusion occurs could be the basis for defining the +group. Further the possibility to separate ageing errors from sampling errors in both catches and surveys was seen as a modelling advantage. However this separation would also have potential but relatively minor effects on estimates of weights at age and maturity-at-age in the stock assessment model fits.

The ability to account for age-reading error is included in several stock assessment programs, such as stock synthesis (Methot, 2000, 2007), Coleraine (Hilborn *et al.*, 2003), and CASAL (Bull *et al.*, 2003). However, although these assessment programs include the ability to account for age-reading error given an age-reading error matrix, they do not include the facility to internally estimate age-reading error matrices (Punt *et al.*, 2008). Also, assessment models are not uniformly structured. For example, assessment programs used in southern Australia allow including ageing errors per individual reader (Punt *et al.*, 2008), whereas Stock Synthesis (SS, Methot, 2000) allows only a single vector of ageing error as input in the model (Dorval at al. 2013).

An implementation of the methodology in R as a package named "nwfscAgeingError" has been developed by Thorson *et al.*, (2013). The main function accepts data with rows being unique reading records and columns corresponding to readers or labs with a unique reading error and bias. The model allows for approximately 15 unique columns. An additional column on the left-hand side of the data matrix indicates the number of otoliths with that unique read record.

The main function named 'FnRun()' writes data in the necessary format and then calls an executable created in ADMB. The model requires several additional inputs on the function form of the bias and the imprecision. Once the model is run and parameters are estimated, plotting routines are available

Use of the ageing error matrix in age structured model has been well described. Methods for using and interpreting ageing error matrices outside age structured assessments should be explored. For instance, how ageing error affects methods for assessing data limited stocks does not appear to be studied in equal detail. However, methods using the von Bertalanffy growth curve and its parameters L_∞ and K estimated from reading hard structures from animals probably suffer from ageing error.

In addition, transforming age-read information outside assessments should also be explored.

The link between calibration workshops and stock assessment is very weak and not operational, which makes it very difficult to integrate these error sources. The usage of Age Reading Error Matrices could be the right output to be provided by age calibration workshops to stock assessment working groups to facilitate such a link.

However, an operational integration of age reading errors and/or maturity staging errors into stock assessment requires methodological developments of available assessment models and calibration analysis tools.

References

Bull B, Francis RICC, Dunn A, McKenzie A, Gilbert DJ, and Smith MH (2003) CASAL (C++ algorithmic stock assessment laboratory): CASAL user manual v2.01-2003/8/01. NIWA Tech. Rep. No. 124.

Campana, S.E., Annand, M.C., McMillan, J.I. 1995. Graphical and Statistical Methods for Determining the Consistency of Age Determinations. Transactions of the American Fisheries Society Vol. 124, 1: 131-138

Campana, S.E. 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.

Catalano, M.J. and Bence, J.R. 2012. The sensitivity to assumed ageing error of the stock assessment used to recommend lake whitefish yield for 2008 from management unit WFH01 of Lake Huron. QFC Tech. Rep. 2011-01

Dorval E, McDaniel JD, Porzio DL, Felix-Uraga R, Hodes V, Rosenfield S (2013) Computing and selecting ageing errors to include in stock assessment models of Pacific sardine (Sardinops saga) CalCOFI Rep., Vol. 54, 2013

Eltink, A.T.G.W. 2000. Age reading comparisons. (MS Excel workbook version 1.0 October 2000) Internet: http://efan.no.

Heifetz J, Anderi D, Maloney NE, and Rutecki TL (1998) Age validation and analysis of ageing error from marked and recaptured sablefish, Anoplopama fimbria. Fish. Bull. (Washington, D.C.), 97: 256–263.

Hilborn R, Maunder, M, Parma A, Ernst B, Payne J, and Starr P (2003) COLERAINE: a generalized age-structured stock assessment manual. User's manual version 2.0. SAFS-UW-0116

ICES 2010. Report on the Workshop on Age Reading of Mackerel (WKARMAC). Available at http://www.ices.dk/community/Pages/PGCCDBS-doc-repository.aspx#gui

ICES 2011a. PGCCDBS Guidelines for Workshops on Age Calibration (update). Available at http://www.ices.dk/community/Pages/PGCCDBS-doc-repository.aspx#gui

ICES 2011b. Report on the Workshop on Age Reading of Greenland halibut (WKARGH). Available at http://www.ices.dk/community/Pages/PGCCDBS-doc-repository.aspx#gui

ICES 2014. Report of the Workshop on Scoping for Integrated Baltic Cod Assessment. 1–3 October 2014. Gdynia, Poland. ICES CM 2014/ACOM:62

Methot R. (2000) Technical description of the Stock Synthesis Assessment Program. NOAA Technical Memo. NMFS-NWFSC-43, 46p.

Methot R (2007) User manual for the integrated analysis program Stock Synthesis 2.

Punt, A.E., Smith, D.C., KrusicGolub, K. and 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. Canadian Journal of Fisheries and Aquatic Sciences, 65: 1991-2005.

Reeves, S. A. 2003. A simulation study of the implications of age-reading errors for stock assessment and management advice. – ICES Journal of Marine Science, 60: 314–328.

Richards LJ, Schnute JT, Kronlund AR, and Beamish RJ (1992) Statistical models for the analysis of ageing error. Can. J. Fish. and Aquat. Sci. 49:1801–1815.

Thorson JT, Stewart IJ, Taylor IG, Punt AE (2013) Using a recruitment-linked multispecies stock assessment model to estimate common trends in recruitment for US West Coast groundfishes. Mar Ecol Prog Ser 483: 245–256

Annex 6: Recommendations

RECOMMENDATION	ADRESSED TO
Update workshops guidelines to include recommended outputs and the dissemination framework.	WGBIOP
Explore the solutions identified by the WKSABCAL, or alternatives, for the integration of error on age readings and maturity staging into stock assessment models.	Working Group on Methods for Fish Stock Assessments (MGWG) Also an option as an ICES JMS theme; to be evaluated by PubCom
Update WebGR and FSA to integrate the methods and outputs identified by the WK.	Maintainers of WebGR and FSA (to be addressed by the ICES secretariate)
Analysis of the effects of error/bias in ageing and staging in stock assessment models.	Future Benchmark Workshops (to be addressed by the ICES secretariate)