Fisheries and Oceans Canada | Pêches et Océans Canada

NOAA FISHERIES
NATIONAL MARINE FISHERIES SERVICE

CERT

Comité d'évaluation des ressources transfrontalières

Document de travail 2014/32

Ne pas citer sans autorisation des auteurs

TRAC

Transboundary Resource Assessment Committee

Working Paper 2014/32

Not to be cited without permission of the authors

Agreement of historical Yellowtail Flounder age estimates: 1963-2007.

Richard S. McBride[1], Sandra J. Sutherland[1], Sarah Merry[1], and Larry Jacobson[2]

[1]Population Biology Branch, [2]Population Dynamics Branch,
Northeast Fisheries Science Center, National Marine Fisheries Service,
166 Water Street, Woods Hole, MA 02543 USA

"This information is distributed solely for the purpose of pre-dissemination peer review. It has not been formally disseminated by NOAA. It does not represent any final agency determination or policy."

Ce document est disponible sur l'Internet à : This document is available on the Internet at :
http://www.mar.dfo-mpo.gc.ca/science/TRAC/trac.html

Canada

ABSTRACT

A comparison of Yellowtail Flounder ages estimated by seven different age experts was performed to determine the consistency of age estimates made at the NEFSC Woods Hole Laboratory from 1963 to 2007. The reader at that time, 2007, used standard procedures to re-age scale impressions that were aged by six previous age readers during 1963-2005, as well as material previously aged by herself (2006-2007; total N = 1,136 fish).

Age-bias plots showed that the four earliest age readers (<1990) had lower precision (inter-reader disagreement range: -3 to +4 years) and tended to estimate higher ages. The magnitude of positive bias was small (mean $\leq$ 1 year) and was restricted to older fish ($\geq$ 4 years). Percent agreement with the recent age reader was low for all previous agers (44-65%) versus 86.2% for the most recent ager against herself. Chang's CV, where lower values indicate higher precision, was higher for all previous agers (6.4-9.5) versus an intra-reader value of 2.3 for the most recent ager. Significant inter-reader asymmetry of paired ages, as measured with the $\chi 2$ statistic, supported higher ages estimated by the earlier age readers, but asymmetry was not evident among the three most recent agers (1990-2009).

Samples were ranked in terms of quality (good, fair, or poor) by the recent age reader, although this criterion had never been used before. When only scales of good quality were included in the comparisons (n = 806), precision improved modestly: percent agreement increased by 1.0 to 9.4% among the seven comparisons, CV typically decreased slightly (<1%), and evidence of asymmetry decreased but still remained common among the early readers.

Two of the oldest Yellowtail Flounder ever aged (14 years) were not included in this specific experiment but they were re-examined in 2014 to support other analyses for the TRAC.

Samples from both fish were aged to be 14 or possibly 15 years old by the new current age expert.

In summary, Yellowtail Flounder age precision is adequate across the entire period measured (1963-2007) and has improved since 1990. Imprecision rises sharply among ages 4-7 but does not continue to increase at older age classes. Ranking and selecting 'good' scale impressions provide only modest improvements in precision. These results: 1) inform discussions about this species' longevity, 2) underscore the need to continue a suite of QA/QC procedures, and 3) contribute a matrix of age uncertainty values that could be incorporated into assessment models.

INTRODUCTION

The purpose of this experiment was to determine the consistency of production ages for Yellowtail Flounder estimated by seven experts at the Northeast Fisheries Science Center since 1963. These ages are used inform estimates of this species' longevity and natural mortality, and they are a stock assessment data source for this economically, culturally important groundfish species. Reasonable attention to validating Yellowtail Flounder aging methods have supported the use of scales or sectioned otoliths (Royce et al. 1959; Lux and Nichy 1969; Walsh and Burnett 2002; Dwyer et al. 2003), and the methods of processing and evaluating these age structures are well documented (Penttila 1988; Pentilla et al. 1988). A recent validation effort, in which scales were taken from tagged fish both during tagging and at recapture, continues to indicate that the scale aging method generates accurate ages in U.S. waters (L. Alade and S. Emery, NEFSC, unpublished data).

In large-scale or "production" aging, cost-benefit analysis supports the NEFSC's method using Yellowtail Flounder fish scales, which is less expensive than sectioning otoliths. This

method has been used continuously since 1963. In 2003, the NEFSC modified their quality assurance and quality controls (QA/QC) procedures to move from a system where a second reader checked the primary reader for consistency of age estimates, to a system where the primary reader is tested against a reference collection and re-ages a subset of samples from the production stream before QA/QC is passed. These QA/QC results are publically available (http://www.nefsc.noaa.gov/fbp/QA-QC/), allowing these measures to be incorporated into population assessments.

A primary reason to move to a one reader system was to reduce costs. Primary readers have performed the task from as few as two to as many as 13 years, so turnover is expected, and actively using two readers is not cost-effective once a suitable reference collection is assembled. This experiment is a formal test to what degree ages estimates might actually agree over the course of several decades and in relation to improvements of these QA/QC procedures. The experiment also allowed us to examine if some of the oldest fish (> 10 years) are reasonable estimates that should inform the parameterization of mortality for this species.

In this study, 1,136 fish scales of Yellowtail Flounder were re-aged by a recent age reader in 2007 (Table 1). Her ages were compared against prior production ages for those fish: ages re-examined from a subsample of her own age estimates (2006-2007) as well as against all six previous Yellowtail Flounder age readers since 1963. Precision levels were measured using standard statistics and graphical approaches, as explained in the Methods and Appendix sections. In addition, a sample including two fish aged originally at 14 y was re-aged by the new current reader (active during 2008-2014) to verify these maximum ages in the database.

METHODS

A simulation calculated appropriate samples sizes, and based on practical considerations, 20-40 scales were determined suitable for each age group and each age reader. The seven age readers responsible for aging Yellowtail Flounder were active during 1963-1969, 1970-1982, 1983-1984, 1985-1989, 1990-1991, 1992-2005, and 2006-2007. The age reader at that time, in 2007, re-aged the scales originally aged by herself and all of the historical readers. The experiment was conducted in a "blind" manner so that the recent expert was not aware of the original reader or original age. The recent expert recorded the new age as well as scale quality, in terms of aging, as good, fair, or poor. Although quality was not recorded by historical readers, and has never been used in aging Yellowtail Flounder nor in selecting specimens for this experiment, we added this to indentify the level of quality variations in scales and to select out ages from the higher quality scales for special comparison.

Scales for each reader were selected in a random fashion – without regard to stock area, source (survey vs. commercial sample), location of collection or other factors – and should be representative of age data for Yellowtail Flounder in U.S. waters and transboundary areas with Canada. The original intention was to use survey data only (for convenience) but commercial samples were also used to boost sample sizes where necessary. Database records were extracted for all Yellowtail Flounder aged at NEFSC taken in survey and then split up into groups based on reader and recorded age. A random sample of forty was drawn from each split. Additional commercial samples were obtained haphazardly.

These fish scales had been impressed in a plastic laminate (Penttila et al. 1988) and archived with any additional scale material in a manila envelope, stored in boxes, in a climate controlled warehouse. Once selected, some samples were not found, or particularly for older

ages, the age reader had never estimated the age, so that there were fewer than 40 specimens in many cases (Table 1). In the laboratory, scale impressions were read by the recent age reader following standard methods (Penttila 1988). All identifying information was removed from the sample, except season of collection, which aids in interpretation of the edge and is necessary for final age assignment.

Standard analyses for checking agreement between pairs of ages are not taught in basic statistic courses, so the unsuitability of regression and correlation is explained and the analyses that will be used are outlined here. Familiar regression techniques, where the independent variable is estimated with little or no error, are not appropriate (Ricker 1973). Correlation statistics are also not appropriate because they are conditioned on rejecting a null hypothesis of no relationship, whereas a relevant test should really address by how much the paired ages agree (Bland and Altman 1986).

Instead, a number of approaches have been developed to measure and interpret paired age agreement, including graphical depictions of the data, indices of precision (i.e., repeatability), and tests of symmetry (Lai et al. 1996; Evans and Hoeing 1998; Campana 2001). Graphic analyses include cross-tabulation of ages by each reader, as well as plots of data, summarized with a measure of variance, against a reference line of full agreement (Altman and Bland 1983; Bland and Altman 1986; Campana et al. 1995). Indices include simple percent agreement but also other calculations that incorporate variability present with each age class (i.e., Chang 1982). Tests of symmetry use a χ^2 statistic to demonstrate whether the distribution of ages that disagree are random (null) or asymmetrical (biased) around those ages that do agree. These approaches are commonly used by laboratories aging fish. They have been automated in EXCEL (http://www.nefsc.noaa.gov/fbp/age-prec/) and with R (http://www.r-bloggers.com/age-

precision-and-bias-changes-to-fsa/, http://cran.r-project.org/web/packages/fishmethods/fishmethods.pdf). More details about the display, formulation, complementary nature, and interpretation of these measures are outlined in an Appendix.

In this study, age precision was assessed using three complementary approaches. First, the data were tabulated and plotted to depict agreements and disagreements between age readers. Second, precision was summarized as single indices, calculated as the percent agreement and Chang's coefficient of variation (CV). Third, three forms of a test of symmetry (McNemar's maximally-pooled method, Evans and Hoenig's method which pools along diagonals, and Bowker's unpooled method) were used to examine disagreements for evidence of bias. These three tests, which calculate a $\chi 2$ statistic, were evaluated at a total threshold for statistical significance ($\alpha = 0.05$) as adjusted across all comparisons (7 readers, adjusted $\alpha = 0.007$).

RESULTS

The recent age reader agreed most often with herself, more often with the two most recent readers ($\geq$ 1990), and least often with the four earliest age readers (<1990, Table 2, Fig 1). Percent agreement was low for all inter-reader tests (44-65%) versus 86% for the most recent ager against herself (Table 2). Chang's CV, where lower values indicate higher precision, was higher for inter-reader tests (6.4-9.5%) versus 2.1% for the most recent ager's intra-reader test. The magnitude of these disagreements was small ($\leq$ 1 year) on average, and largely restricted to a few age classes (4 to 7 years; Fig. 1). A positive bias was evident, again most pronounced with the earliest age readers (Table 2). Significant inter-ager asymmetry of paired ages was evident for the first four agers (1963-1989, relative to the recent ager), but no inter- or intra-reader asymmetry was evident among the three most recent agers (1990-2007).

Scale quality had never been used by NEFSC to choose samples or age Yellowtail Flounder. However, when only 'good' scales were included in the comparisons (n = 806, Table 2B), modest improvements in precision were evident. Percent agreement increased by 1.0 to 9.4% among the seven comparisons, CV typically decreased but by less than 1%, and indications of asymmetry decreased but still remained common among inter-reader comparisons with the early agers.

Disagreements were not evident among young age classes (< 4 years), indicating that all readers were recognizing the first few annuli consistently (Fig. 1). Disagreements were most evident between ages 4 and 7. On average, these disagreements were small (i.e., $\leq$ 1 year), and the average disagreement fell within the intra-reader aging error of the most recent reader (i.e., all her second readings were within $\pm$ 1 year of the first readings). Overall, individual inter-reader disagreements ranged from -3 to + 4 years. Few fish > 9 years of age were available, but the disagreements did not appear to increase with increasing age > 7 years.

DISCUSSION

All ages are estimated with some imprecision, which can arise from both inter- and intra-reader sources. Intra-reader precision was estimated only for the recent reader, and it is now impossible to quantify intra-reader precision levels for the historic age readers. Therefore, a full examination of the sources of error among these samples cannot be conducted. Having acknowledged that, it appears that age reader precision has improved over time, and the recent reader who benefited from new QA/QC procedures is the best available for use as a standard. Two tests the recent age reader conducted against the reference collection yielded accuracy levels of 91-93% agreement and CVs of 1.2-1.6% (http://www.nefsc.noaa.gov/fbp/QA-QC/), demonstrating very high levels of reproducibility.

Current standards of intra-reader precision for NEFSC agers are 80% and a Chang's CV below 5% (http://www.nefsc.noaa.gov/fbp/QA-QC/)[1]. These standards were implemented circa 2003, when the NEFSC aging program switched from a two-reader system to a system where one production ager is tested against subsamples of aged fish at regular intervals as well as against a reference collection, if one exists for that species. This shift in QA/QC procedures appears to have followed at least modest increases in precision of Yellowtail Flounder age estimates evident in the 1990s.

Imprecision, resulting in low rates of reproducibility, was most obvious among age classes 4-7. This occurs well after maturation, by age 2 for this species, so the imprecision was not associated with misleading spawning checks as false annuli. However, growth in Yellowtail Flounder slows after maturation, resulting in annuli occurring closer together or compressed along the margin of the scale, which confounds the recognition of distinct annuli. According to the new current age reader, lower rates of precision from age 4 to 7 are associated with this early period of crowding (S. Emery, pers. comm.). In older fish, the pattern of slower growth and closer annuli after age 5 is more evident to the reader, so that additional imprecision with increasing age does not appear to be a problem. Restricting the comparison to only good scales improved measures of agreement but did not eliminate evidence of imprecision and bias.

It is of note that among the complete data set[2], 20 of 23 ages > 10 years were estimated by the 1970-1982 age reader (Table 3). This reader had the highest inter-reader range of disagreement (+4 to -3) and tended to overestimate older age classes of fish, particularly age 7 (average inter-reader bias was $\leq$ 1 year). However, this historic age reader, who wrote the manual on aging Yellowtail Flounder (Penttila 1988), also aged for the longest period and aged

[1] Inter-reader results typically show higher rates of imprecision.
[2] 64,000 Yellowtail Flounder ages collected by spring and autumn resource surveys, 1963-2007 (Table 3).

nearly half (45.2%) of these Yellowtail Flounder. No other reader aged more than 7% of the material. Therefore, older ages reported by this reader are at least partly influenced by the larger sample size aged by this reader. Two of the oldest Yellowtail Flounder ever aged (14 years), sampled in 1975 and 1977, but not included in this specific experiment, were re-examined in 2014 to be included in this report. They were unusually large (55 & 58 cm total length) and determined to be 14 or possibly 15 years old (S. Emery, pers. comm.), which verifies the ages evident from these scales.

Age disagreements are to be expected, both within and between readers. It is the frequency, bias, and magnitude of these disagreements that is of concern. Age imprecision 'smears' year classes together, apparently more by historic agers than by recent agers in this case of Yellowtail Flounder. Bias indicates observation error, such as one reader having differing, criteria for evaluating annuli than another reader. Reducing the magnitude of imprecision and bias have a familiar suite of QA/QC remedies: 1) validation research of the annuli number and spacing, 2) routine monitoring of agreement against a reference collection and a subset of production ages, 3) exchanges between laboratories, and 4) training workshops between agers. These data presented herein may also be used to incorporate age estimate uncertainty into a stock assessment model, and determine the effect of this imprecision and bias that has been revealed by this experiment on the outcome of a stock assessment.

In summary, Yellowtail Flounder age precision is adequate across the entire period (1963-2007) and has improved specifically since 1990. Imprecision rises sharply among ages 4-7 but does not continue to increase at older age classes. Ranking and selecting the 'good' scale impressions provide only modest improvements in precision. These results: 1) inform discussions about natural mortality of this species, 2) underscore the need to continue QA/QC

efforts for production age determination, and 3) contribute a matrix of age uncertainty values that could be incorporated into assessment models.

ACKNOWLEDGMENTS

We thank all the past Yellowtail Flounder age readers: Jeff Darde, Scott Mosely, Fred Nichy, Judy Penttila, Gary Shepherd, and Vaughn Silva. Jay Burnett facilitated many aspects of the original experiment, and Sarah Emery, the newest age reader, provided insight into possible sources of Yellowtail Flounder aging error. We also recognize the countless others who assisted with fish collection and sample processing, or who helped retrieve, organize, and return the archived samples.

LITERATURE CITED

Altman, D. G., and J. M. Bland. 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. Journal of the Royal Statistical Society. Series D (The Statistician) 32(3):307-317.

Bland, J. M., and D. G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i:307-317.

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

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

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

Dwyer, K. S., S. J. Walsh, and S. E. Campana. 2003. Age determination, validation and growth of Grand Bank yellowtail flounder (*Limanda ferruginea*). ICES Journal of Marine Science 60(5):1123-1138.

Evans, G. T., and J. M. Hoenig. 1998. Testing and viewing symmetry in contingency tables, with application to readers of fish ages. Biometrics 54:620-629.

Fishery Biology Program (website) Quality Assurance and Quality Control Estimates for the

Production Ageing of Northwest Atlantic Species (www.nefsc.noaa.gov/fbp/QA-QC/)

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

Lai, H. L., V. F. Gallucci, D. R. Gunderson, and R. F. Donnelly. 1996. Age determination in fisheries: methods and applications to stock assessment. Pages 82-178 *in* V. F. Ballucci, S. B. Saila, D. J. Gustafson, and B. J. Rothschild, editors. Stock assessment: Quantitative methods and applications for small-scales fisheries. CRC Press, Lewis Publishers, Boca Raton, FL.

Lux, F. E., and F. E. Nichy. 1969. Growth of yellowtail flounder, *Limanda ferruginea* (Storer), on three New England fishing grounds. International Commission for the North Atlantic Fisheries, Research Bulletin 6:5-25.

NEFSC (Northeast Fisheries Science Center). 1998. Report of the 28th Northeast Regional Stock Assessment Workshop (28th SAW): Public Review Workshop. Northeast Fish. Sci. Cent. Ref Doc. 99-07; 49 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026.

NEFSC (Northeast Fisheries Science Center). 2012. 54th Northeast Regional Stock Assessment Workshop (54th SAW) Assessment Report. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 12-18; 600 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/publications/

Penttila, J. 1988. Yellowtail Flounder. Pages 119-124 *In*: Penttila, J., and L. M. Dery. 1988. Age determination methods for northwest Atlantic species. NOAA Technical Report NMFS 72(25 September 2009):135 pp. (http://www.nefsc.noaa.gov/fbp/age-man/yt/yt.htm)

Penttila, J., F. Nichy, J. Ropes, L. Dery, and A. Jerald. 1988. Methods and equipment. Pages 7-16 *In*: Penttila, J., and L. M. Dery. 1988. Age determination methods for northwest Atlantic species. NOAA Technical Report NMFS 72(25 September 2009):135 pp. (http://www.nefsc.noaa.gov/fbp/age-man/yt/yt.htm)

Ricker, W. E. 1973. Linear regressions in fishery research. Journal of the Fisheries Research Board of Canada 30:409-434.

Royce, W. F., R. F. Buller, and E. D. Premetz. 1959. Decline of the yellowtail flounder (*Limanda ferruginea*) off New England. Fishery Bulletin, U. S. 146:1-267.

Walsh, S. J., and J. Burnett. 2002. The Canada-United States yellowtail flounder age reading workshop, 28-30 November 2000, St. John's Newfoundland. Northwest Atlantic Fisheries Organization Scientific Council Studies 35:1-59.

Table 1. Yellowtail Flounder scales aged for each age reader and original age group. Age readers are numbered in chronological order and years indicate when they were responsible for aging Yellowtail Flounder at the NEFSC.

Original age	Age reader						
	Reader 1 1963-1969	Reader 2 1970-1982	Reader 3 1983-1984	Reader 4 1985-1989	Reader 5 1990-1991	Reader 6 1992-2005	"Recent" Reader 7 2006-2007
0	0	7	0	0	9	16	8
1	7	14	15	10	19	17	20
2	13	18	16	8	20	19	20
3	12	17	18	7	20	20	20
4	9	18	19	10	20	20	20
5	12	18	19	10	18	20	20
6	27	33	36	14	21	30	20
7	30	32	22	16	7	26	7
8	22	31	20	7	0	18	3
9	17	26	1	0	0	7	0
10	4	31	0	0	0	0	0
11	2	16	0	0	0	0	0
12	0	7	0	0	0	0	0
Total	155	268	166	82	134	193	138

Table 2. Comparisons of (A) all scales (*n* = 1,136) and (B) 'good' quality scales only (*n* = 806). Indices of precision (PA, CV) and tests of symmetry (McNemar, Evans & Hoenig, Bowker) for each comparison are tabulated by the time period of aging by the original reader, 1963-2007. Significance of each test of symmetry is evaluated at 0.007 (**bold**), which is an adjustment of α = 0.05 among seven comparisons. (See Appendix for formulation of the indices and calculation and evaluation of the χ2 statistics used in the tests of symmetry.)

A) All scales							
Sample period	1963-69	1970-82	1983-84	1985-89	1990-91	1992-05	2006-07
Number of ages	155	268	166	82	134	193	138
Indices							
PA	51.0	44.4	53.6	56.1	63.4	65.3	86.2
CV	7.30	8.64	9.49	8.23	8.39	6.40	2.12
McNemar							
χ2	15.2	41.9	33.7	18.8	0.51	0.015	2.58
df	1	1	1	1	1	1	1
P	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	0.470	0.903	0.108
Evans and Hoenig							
χ2	17.1	44.6	35.3	19.7	3.67	0.418	2.58
df	3	4	4	3	3	3	1
P	**0.0007**	**< 0.0001**	**< 0.0001**	**0.0002**	0.300	0.936	0.108
Bowker							
χ2	31.0	74.4	49.4	24.7	22.6	5.2	9
df	15	25	16	12	11	12	6
P	0.0088	**< 0.0001**	**< 0.0001**	0.016	0.020	0.950	0.173

Table 2 (continued)

B) Good scales only							
Historic Reader	1963-69	1970-82	1983-84	1985-89	1990-91	1992-05	2006-07
Number of ages	79	137	141	73	118	150	108
Indices							
PA	59.5	48.9	58.9	58.9	64.4	74.7	92.6
CV	6.41	7.70	8.94	7.94	8.58	5.61	1.26
McNemar							
χ2	15.1	6.91	24.9	19.2	0.381	0.421	4.5
df	1	1	1	1	1	1	1
P	**0.0001**	0.0085	**< 0.0001**	**< 0.0001**	0.537	0.516	0.034
Evans and Hoenig							
χ2	16.5	9.85	25.8	19.6	5.03	1.32	4.5
df	3	4	4	3	3	3	1
P	**0.0009**	0.043	**< 0.0001**	**0.0002**	0.170	0.723	0.034
Bowker							
χ2	21.1	40.8	41.3	24.7	27.8	12.8	6
df	14	20	16	12	11	12	6
P	0.099	**0.004**	**< 0.001**	0.016	0.004	0.381	0.423

Table 3. Number of Yellowtail Flounder aged, by year and age class, from the NEFSC spring and autumn resource surveys.

YEAR	AGE															All
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1963	.	496	348	415	157	35	9	3	5	1	.	.	.	.	.	1469
1964	.	126	264	93	92	45	11	3	.	.	.	.	.	.	.	634
1965	.	55	188	127	53	31	2	2	1	.	.	.	.	.	.	459
1966	.	49	112	51	18	5	5	1	.	.	.	.	.	.	.	241
1967	.	139	383	213	53	14	4	5	.	.	.	.	.	.	.	811
1968	1	42	330	489	168	34	9	7	2	.	.	.	.	.	.	1082
1969	.	32	275	552	248	41	16	8	2	.	2	.	.	1	.	1177
1970	.	177	421	659	591	182	50	29	10	6	4	1	.	.	.	2130
1971	.	253	883	875	1246	271	58	19	4	11	3	1	1	.	.	3625
1972	152	182	1250	1236	1003	612	117	18	4	4	1	.	.	.	.	4579
1973	.	230	479	893	482	319	321	41	10	9	3	.	1	.	.	2788
1974	60	230	374	341	425	211	85	75	8	3	.	3	.	.	.	1815
1975	.	160	378	134	109	104	33	16	7	.	1	2	2	.	1	947
1976	.	90	589	189	65	57	53	24	15	6	5	.	.	.	.	1093
1977	5	186	336	366	99	34	30	17	10	4	4	3	1	.	1	1096
1978	2	418	886	382	213	72	25	27	16	2	1	.	.	.	.	2044
1979	1	518	916	529	164	92	43	33	14	5	1	1	.	.	.	2317
1980	5	347	1098	976	465	89	67	18	10	7	1	1	.	.	.	3084
1981	.	264	782	437	230	90	46	14	6	1	1	.	.	.	.	1871
1982	.	107	817	482	177	86	30	14	2	1	2	.	1	.	.	1719
1983	.	80	491	818	146	27	7	4	4	1	.	.	.	.	.	1578
1984	1	91	243	230	228	82	29	7	4	2	.	.	.	.	.	917
1985	.	169	316	158	93	80	18	5	3	.	.	.	.	.	.	842
1986	.	52	564	173	67	41	6	1	.	.	.	.	.	.	.	904
1987	.	98	167	258	47	21	18	9	2	.	.	1	.	.	.	621
1988	.	284	262	88	70	47	11	6	2	.	.	.	.	.	.	770
1989	1	65	677	217	54	12	2	2	.	.	.	.	.	.	.	1030
1990	9	49	318	1036	136	21	6	1	1	.	.	.	.	.	.	1577
1991	1	110	293	445	442	76	11	4	.	.	.	.	.	.	.	1382
1992	.	59	190	243	174	49	7	2	.	.	.	.	.	.	.	724
1993	1	101	191	168	115	12	2	3	.	.	.	.	.	.	.	593
1994	1	150	311	201	105	60	21	4	.	.	.	.	.	.	.	853
1995	3	98	232	236	123	35	7	6	2	2	.	.	.	.	.	744
1996	.	86	222	337	205	62	5	1	.	.	.	.	.	.	.	918
1997	.	149	434	492	244	87	8	1	.	.	.	.	.	.	.	1415
1998	3	274	589	500	174	73	19	2	.	.	.	.	.	.	.	1634
1999	1	242	605	560	161	72	15	7	1	1	.	.	.	.	.	1665
2000	3	88	462	435	160	34	20	4	.	.	.	.	.	.	.	1206
2001	2	82	323	504	155	51	17	15	.	.	.	.	.	.	.	1149
2002	13	127	319	738	330	84	19	13	4	1	.	.	.	.	.	1648
2003	21	119	472	337	271	87	23	10	5	1	.	1	.	.	.	1347
2004	11	154	399	361	283	273	126	36	6	1	.	.	.	.	.	1650
2005	29	525	487	515	344	497	325	67	15	.	1	.	.	.	.	2805
2006	8	266	470	447	177	29	8	2	2	.	.	.	.	.	.	1409
2007	2	100	610	551	237	42	9	3	.	.	.	.	.	.	.	1554

Figure legends

Figure 1. Graphical depiction of age agreements and error between agers for scales collected during the seven periods, 1963-2007. For each age reader, two sections are shown. First is an age-frequency table, showing the distribution of age estimates (years) by the recent age reader (newage) versus the historical reader (age). Secondly, an age-bias plot shows the mean difference (cross) between the paired ages plotted against the age by the recent reader. The dashed line indicates perfect agreement. The colored vertical bars represent 95% confidence intervals of the age difference between readers; they are show in red if this range is significantly different from 0, and are blue otherwise. Sample size is depicted along the top of each graph.

Figure 1A.

Ages from fish collected 1963-69

	newage									
age	1	2	3	4	5	6	7	8	9	10
1	7	0	0	0	0	0	0	0	0	0
2	0	13	0	0	0	0	0	0	0	0
3	0	1	7	4	0	0	0	0	0	0
4	0	0	1	6	2	0	0	0	0	0
5	0	0	0	6	5	0	1	0	0	0
6	0	0	0	1	7	15	2	2	0	0
7	0	0	0	1	2	7	14	6	0	0
8	0	0	0	0	0	5	8	5	3	1
9	0	0	0	0	0	0	8	5	4	0
10	0	0	0	0	0	0	1	0	0	3
11	0	0	0	0	0	0	0	0	2	0

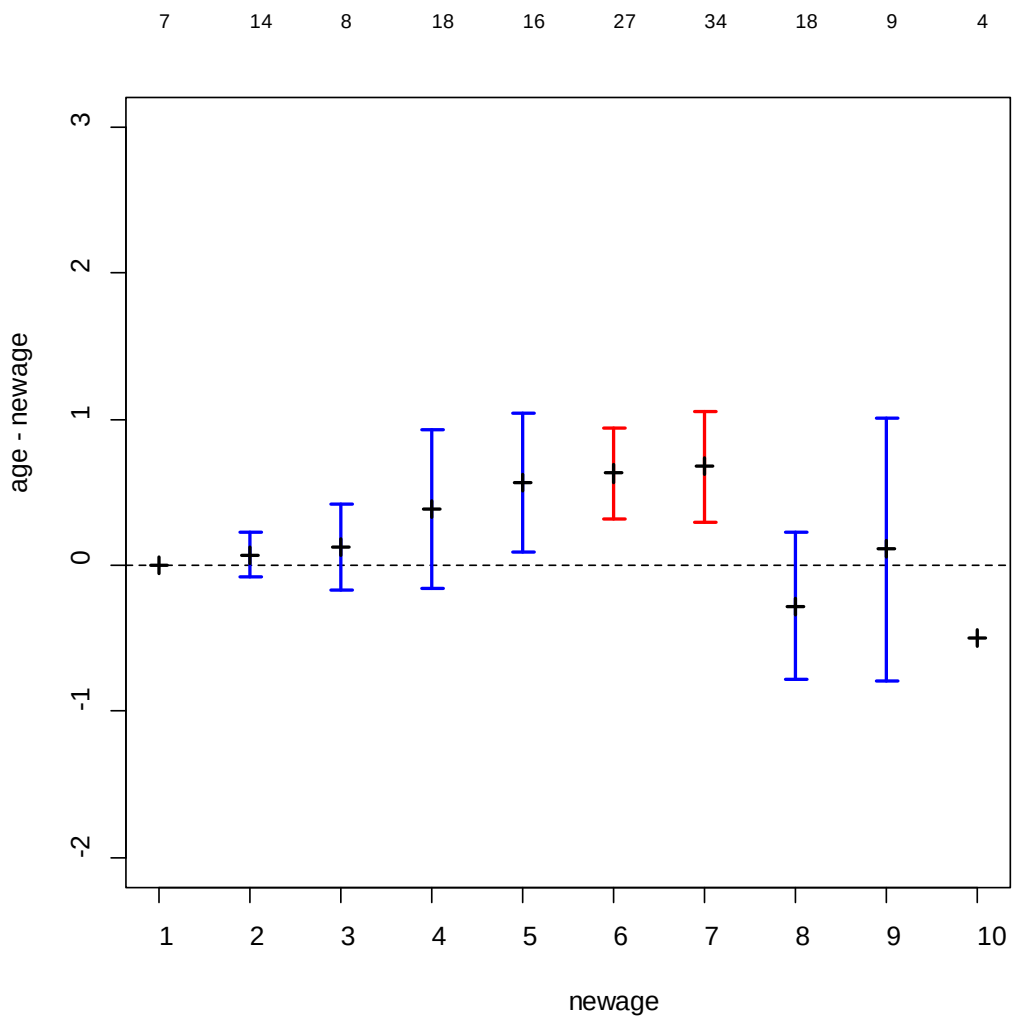


Figure 1B (cont.).

Ages from fish collected 1970-1982

	newage												
age	0	1	2	3	4	5	6	7	8	9	10	11	12
0	7	0	0	0	0	0	0	0	0	0	0	0	0
1	0	14	0	0	0	0	0	0	0	0	0	0	0
2	0	0	15	3	0	0	0	0	0	0	0	0	0
3	0	0	1	14	2	0	0	0	0	0	0	0	0
4	0	0	0	2	9	7	0	0	0	0	0	0	0
5	0	0	0	1	3	10	2	0	2	0	0	0	0
6	0	0	0	0	3	11	14	3	2	0	0	0	0
7	0	0	0	0	1	6	6	17	0	2	0	0	0
8	0	0	0	0	1	1	2	15	5	4	2	1	0
9	0	0	0	0	0	0	2	7	7	7	2	0	1
10	0	0	0	0	0	0	0	7	7	8	7	2	0
11	0	0	0	0	0	0	0	0	3	8	5	0	0
12	0	0	0	0	0	0	0	0	0	0	3	4	0

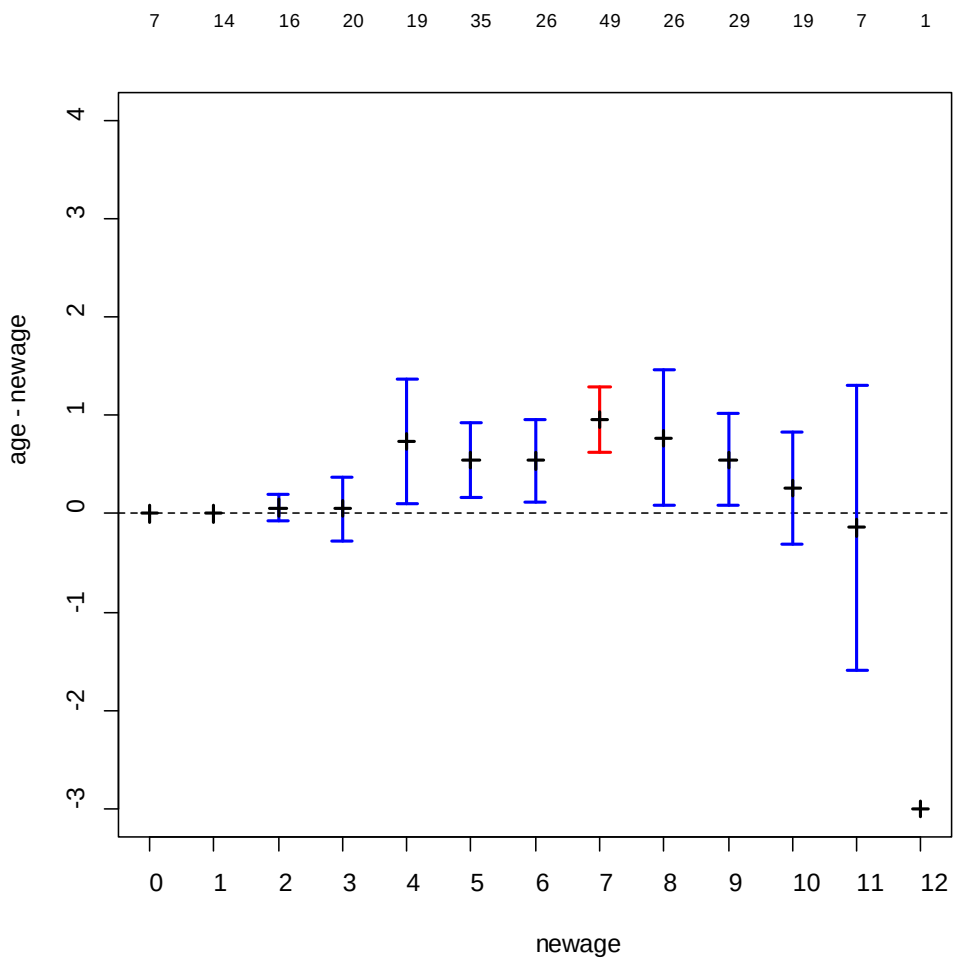


Figure 1C (cont.).
Ages from fish collected 1983-1984

	newage									
age	0	1	2	3	4	5	6	7	8	9
1	1	13	1	0	0	0	0	0	0	0
2	0	0	13	3	0	0	0	0	0	0
3	0	0	1	15	2	0	0	0	0	0
4	0	0	0	7	10	2	0	0	0	0
5	0	0	0	3	5	9	1	1	0	0
6	0	0	0	0	7	14	15	0	0	0
7	0	0	0	0	2	2	8	9	1	0
8	0	0	0	0	1	1	3	8	5	2
9	0	0	0	0	0	0	0	0	1	0

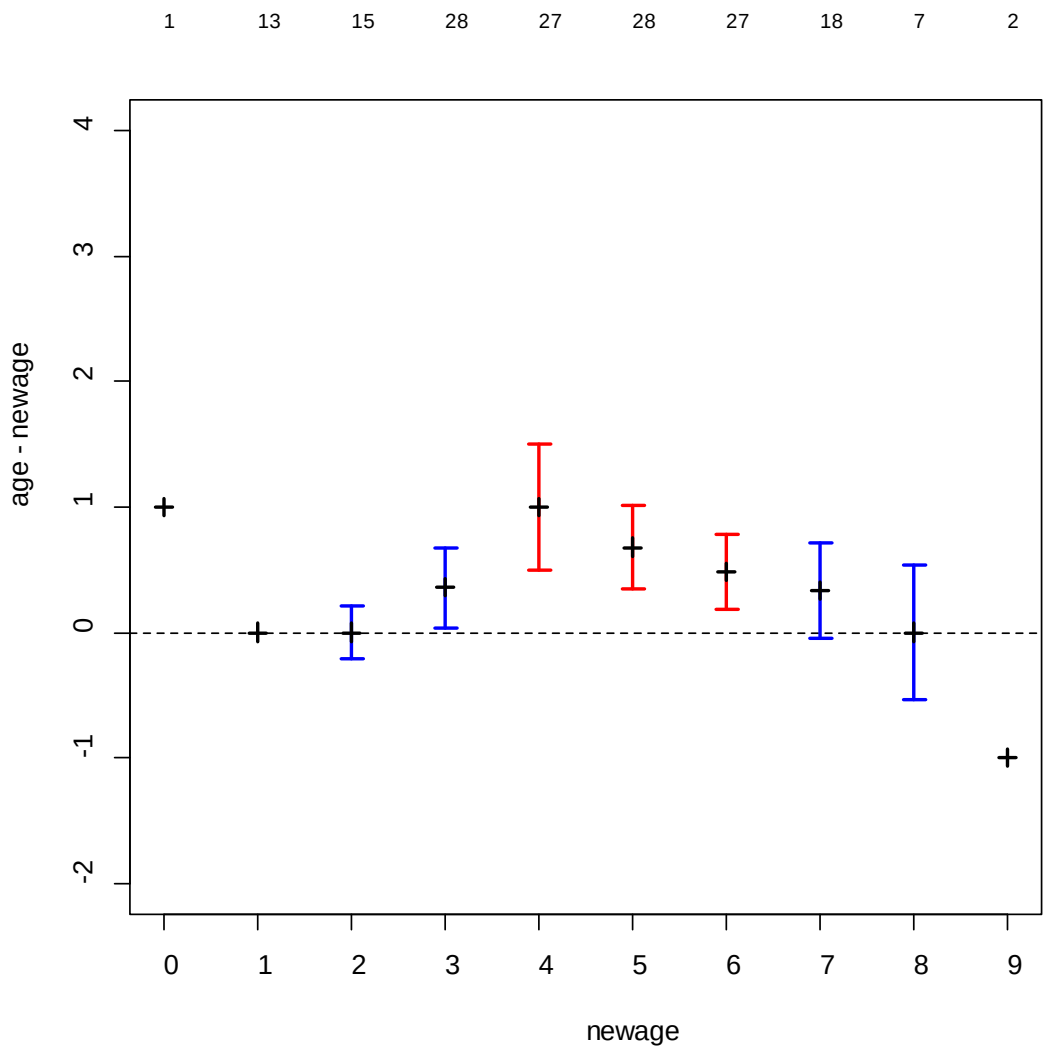


Figure 1D (cont.).

Ages from fish collected 1985-1989

	newage							
age	1	2	3	4	5	6	7	8
1	9	1	0	0	0	0	0	0
2	0	7	1	0	0	0	0	0
3	0	2	5	0	0	0	0	0
4	0	0	3	7	0	0	0	0
5	0	0	2	3	5	0	0	0
6	0	0	0	1	6	5	2	0
7	0	0	0	0	3	6	6	1
8	0	0	0	0	1	2	2	2

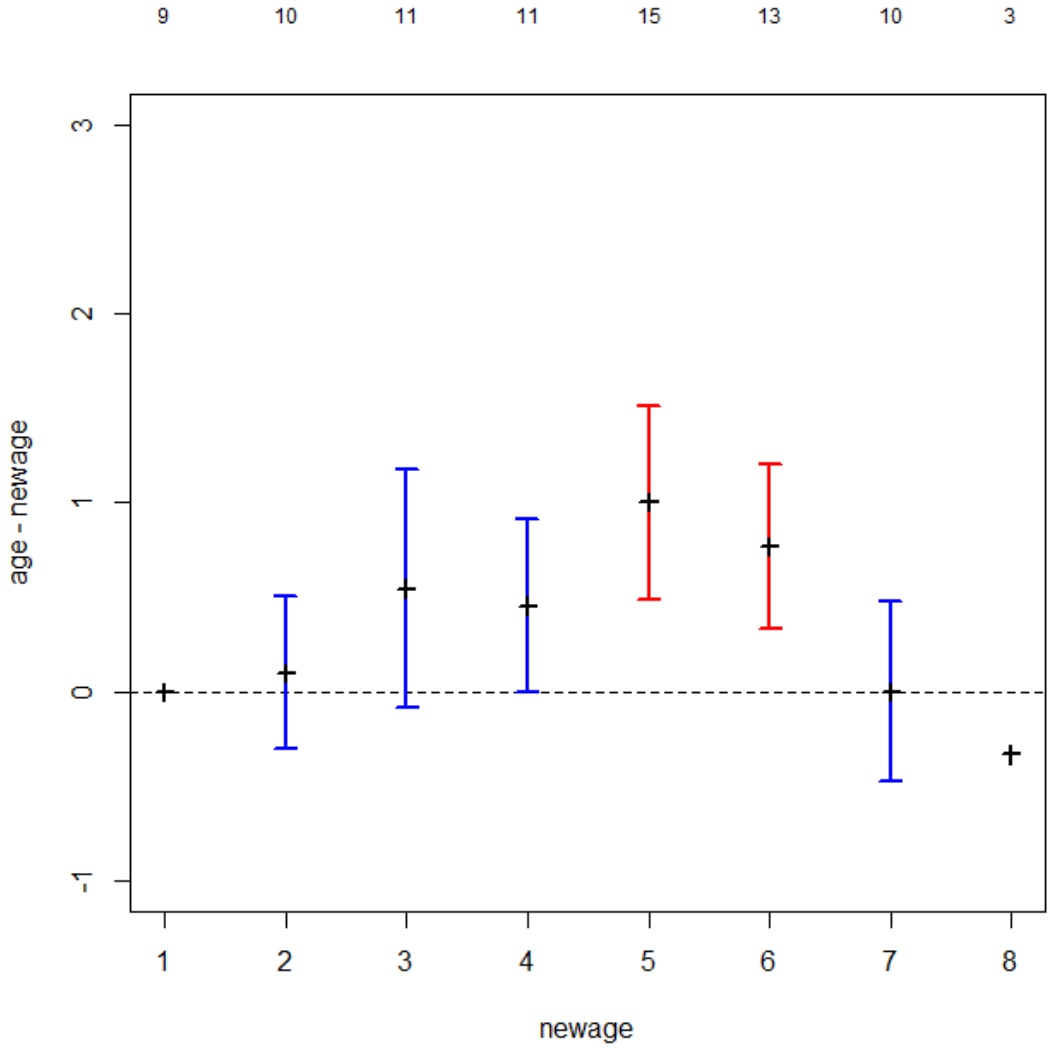


Figure 1E (cont.).

Ages from fish collected 1990-1991

	newage									
age	0	1	2	3	4	5	6	7	8	9
0	9	0	0	0	0	0	0	0	0	0
1	1	15	3	0	0	0	0	0	0	0
2	0	1	17	2	0	0	0	0	0	0
3	0	0	2	9	9	0	0	0	0	0
4	0	0	0	2	17	1	0	0	0	0
5	0	0	0	0	10	5	1	2	0	0
6	0	0	0	0	0	4	9	5	2	1
7	0	0	0	0	0	1	1	4	0	1

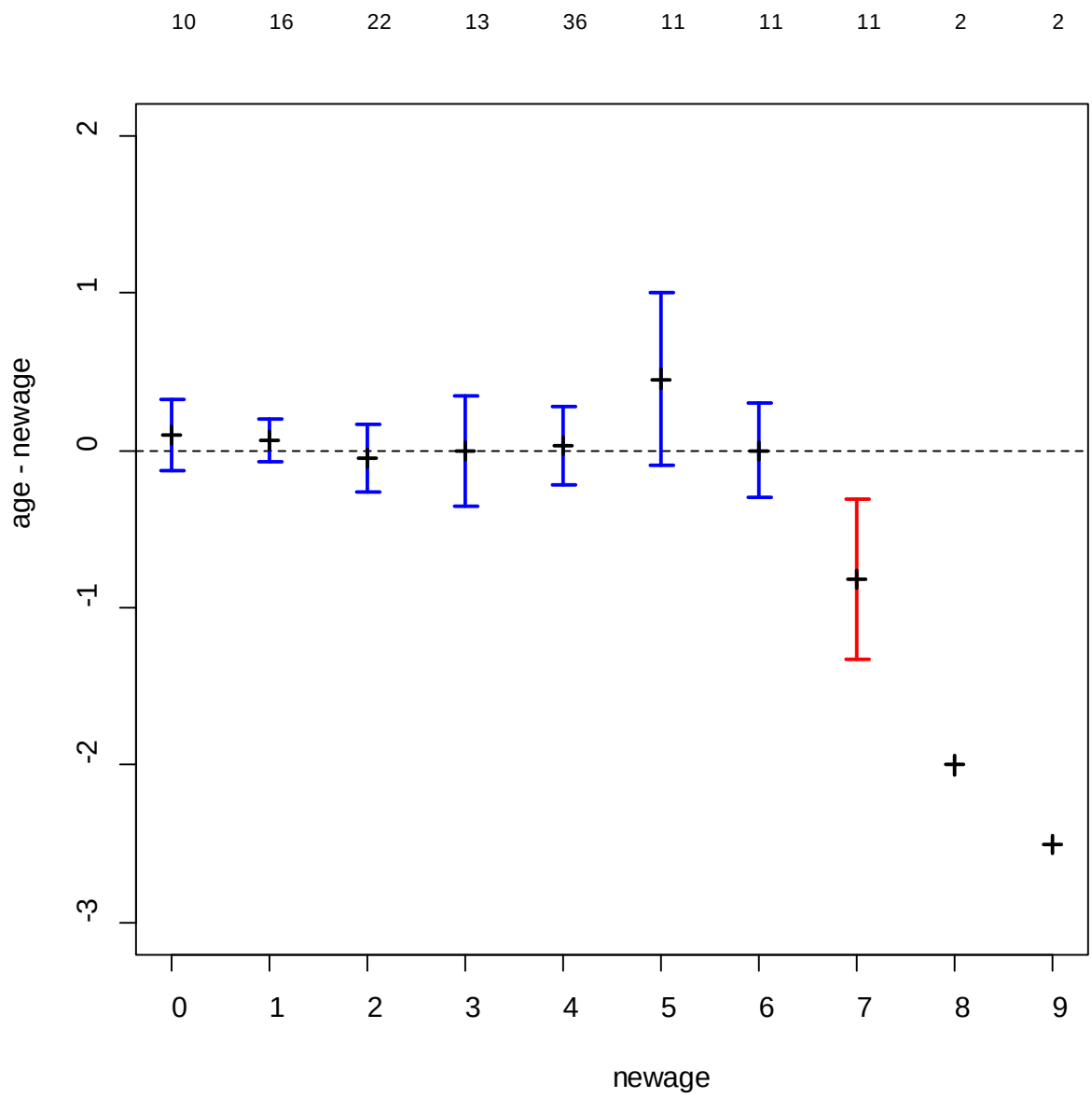


Figure 1F (cont.).

Ages from fish collected 1992-2005

	newage									
age	0	1	2	3	4	5	6	7	8	9
0	14	2	0	0	0	0	0	0	0	0
1	0	17	0	0	0	0	0	0	0	0
2	0	0	18	1	0	0	0	0	0	0
3	0	0	0	18	2	0	0	0	0	0
4	0	0	0	1	12	5	2	0	0	0
5	0	0	0	0	6	7	5	2	0	0
6	0	0	0	0	1	6	18	4	0	1
7	0	0	0	0	0	1	6	11	6	2
8	0	0	0	0	0	0	0	8	8	2
9	0	0	0	0	0	0	1	2	1	3

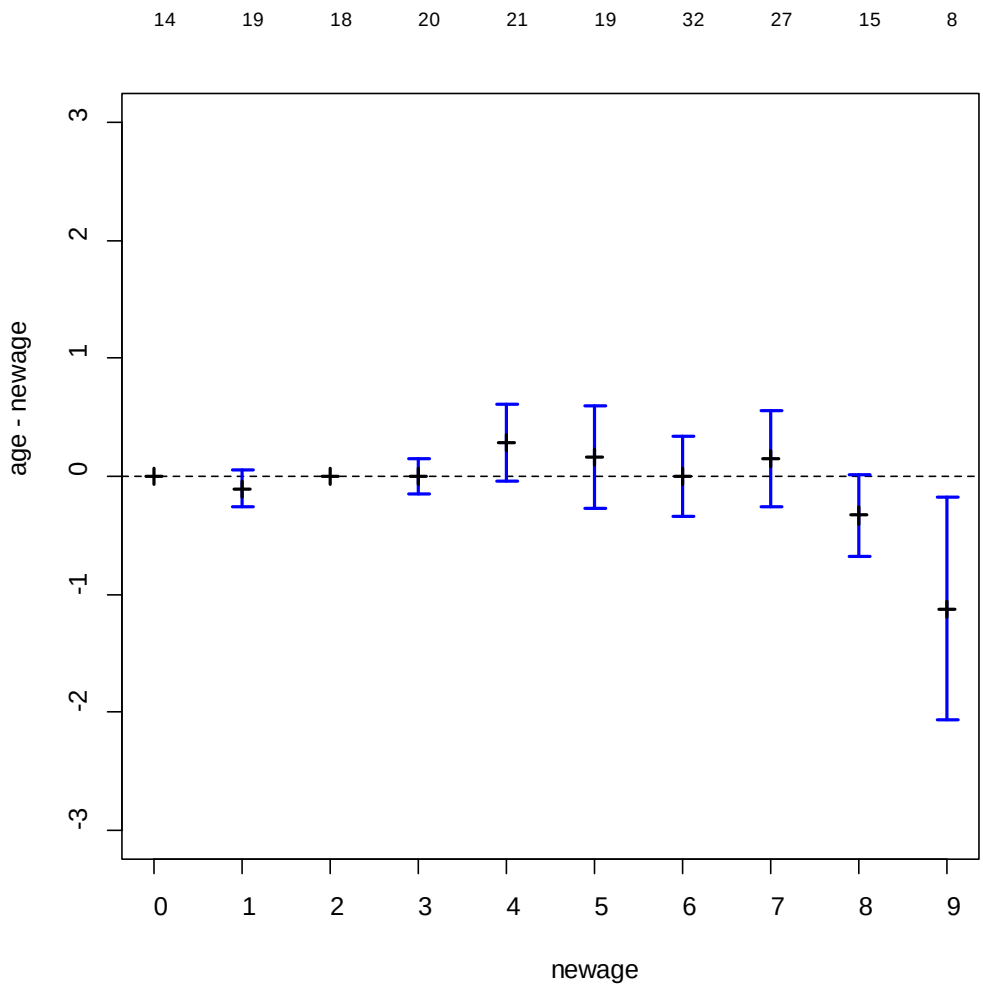
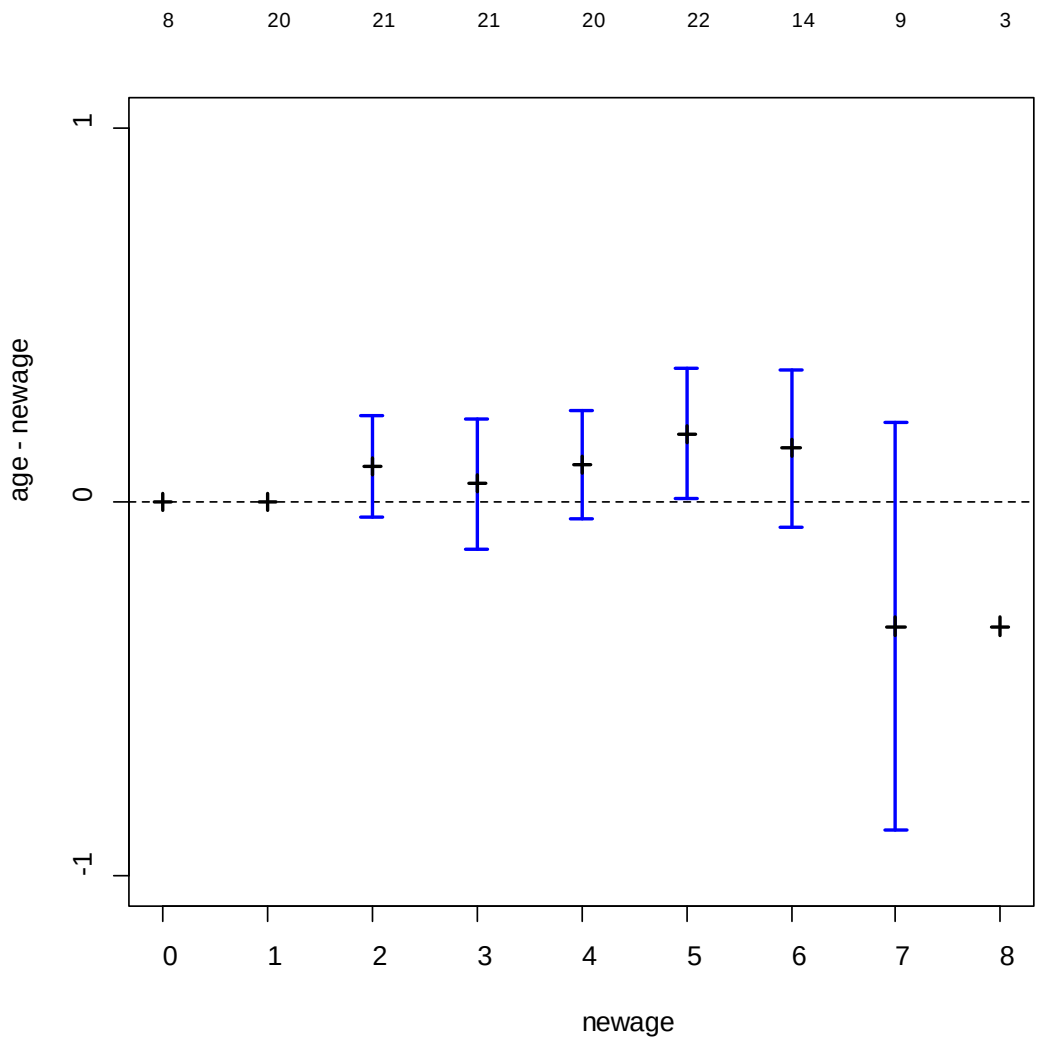


Figure 1G (cont.).

Ages from fish collected 2006-07

	newage								
age	0	1	2	3	4	5	6	7	8
0	8	0	0	0	0	0	0	0	0
1	0	20	0	0	0	0	0	0	0
2	0	0	19	1	0	0	0	0	0
3	0	0	2	18	0	0	0	0	0
4	0	0	0	2	18	0	0	0	0
5	0	0	0	0	2	18	0	0	0
6	0	0	0	0	0	4	12	4	0
7	0	0	0	0	0	0	2	4	1
8	0	0	0	0	0	0	0	1	2



Appendix. An outline of methods to measure and display agreement in fish ages or some other variable measured by two methods or different readers.

Richard S. McBride

Northeast Fisheries Science Center, NOAA Fisheries, Woods Hole, MA

17 March 2014

Graphical depictions of paired age data, indices of precision, and tests of symmetry are explained by focusing on age estimates from 27 Yellowtail Flounder[3] as determined by three methods: scale impressions, whole otoliths and cross-sectioned otoliths.

Tabulation and graphics

A tabulation of the data, below, is typically included when reporting paired ages.

Tabulation of Yellowtail Flounder ages by aging hardpart and method (X-S = cross sectioned)

Whole Otolith Age	Scale Age					
	2	**3**	**4**	**5**	**6**	**7**
2	8		1			
3		3	1		1	
4			3	3		
5				1	5	
6						1
7						

Whole Otolith Age	X-S Otolith Age					
	2	**3**	**4**	**5**	**6**	**7**
2	8	1				
3		3	1			1
4		1		4	1	
5				3	3	
6						1
7						

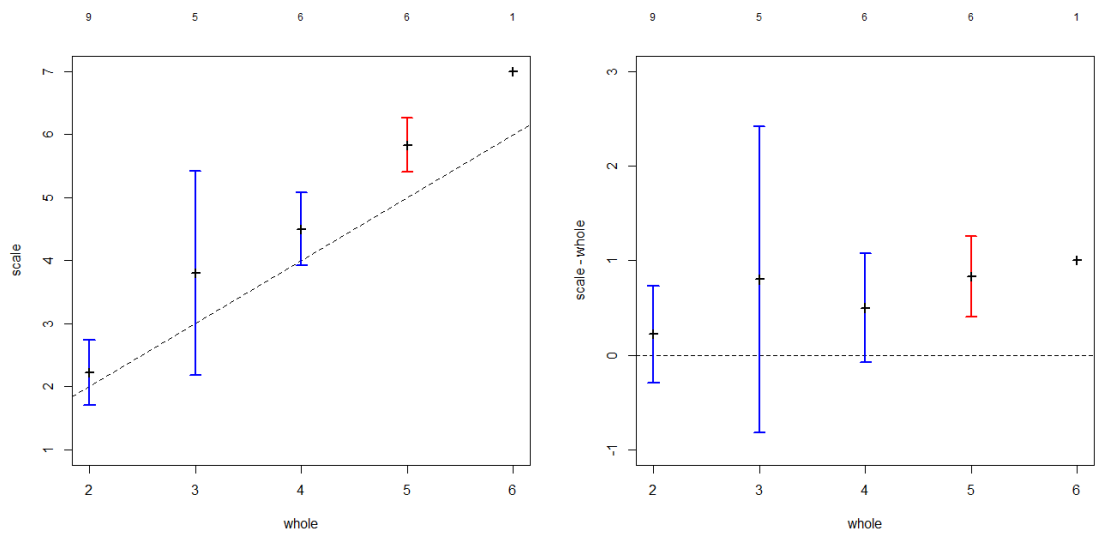
X-S Otolith Age	Scale Age					
	2	**3**	**4**	**5**	**6**	**7**
2	7		1			
3	1	3	1			
4			1			
5			2	3	2	
6				1	3	
7					1	1

[3] Ages of commercially caught Georges Bank Yellowtail Flounder (*Limanda ferruginea*). Source: Walsh, S.J. and J. Burnett. 2002. The Canada-United States yellowtail flounder age reading workshop: 28-30 November 2000, St. John's, Newfoundland. North Atlantic Fisheries Organization. Scientific Council Studies 35:1-59 (tables in Annex 3, http://archive.nafo.int/open/studies/s35/walsh.pdf. Data are also online at www.rforge.net/FSAdata/ (see Appendix Table).

The values in gray, along the diagonal, indicate where agreement exists between paired methods. Values off the diagonal indicate disagreement. Higher ages are estimated by scales and sectioned otoliths than whole otoliths, particularly at older ages.

Another standard graphical approach is to display an 'age-bias' plot (Campana et al. 1995), as shown below, using the 'ageBias' function written in R by Derek Ogle (https://sites.google.com/site/fishrfiles/gnrl/AgeComparisons.pdf?attredirects=0).

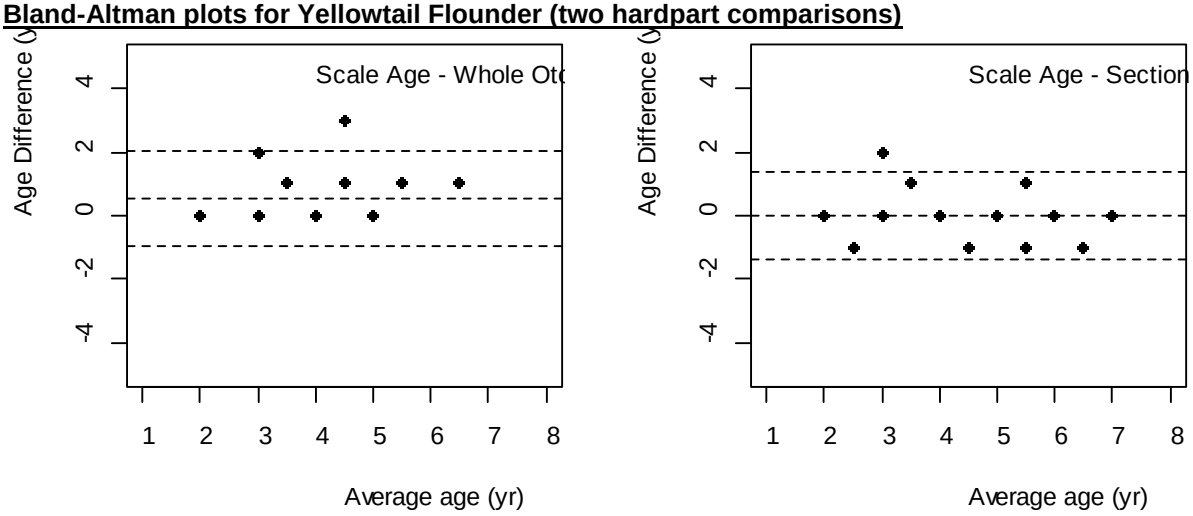
Age-bias plots for Yellowtail Flounder scales versus whole otoliths



The figure on the left plots the values against a 45° slope of perfect agreement (dashed line), whereas the figure on the right plots the reference line of perfect agreement as a dashed hortizontal line. The colored vertical bars are 95% confidence intervals of scale ages compared to whole otolith ages, showing if this interval is (red) or is not (blue) different than 1:1 agreement. The cross in the center of the confidence interval is the mean, or if alone, is the value of a single point. Sample size is depicted along the top of each graph. The NEFSC tabulates ages and includes an age-bias plot in their routine quality assurance and quality control procedures (http://www.nefsc.noaa.gov/fbp/age-prec/).

A similar graphical approach is to use Bland and Altman plots (Bland and Altman 1986). These plot the individual differences between two methods of measurement against the mean of each paired measurement. The presence of bias is indicated by the distribution of points relative to the mean difference between them (middle, dashed hortizontal line in the figure below). The magnitude of bias is indicated by the

Bland-Altman plots for Yellowtail Flounder (two hardpart comparisons)



distribution of points relative to ± 2 standard deviations of this distribution of differences (top and bottom dashed lines). The plot on the left depicts that ages estimated from whole otoliths are more biased relative to scales than ages estimated from sectioned otoliths (right figure). Like the age-biased plot, the Bland-Altman plot directly addresses the issue of agreement between the real values, and not whether the points are related in some general way, as in regression analysis, or affected by the range of values, as in correlation analysis (Altman and Bland 1983; Bland and Altman 1986). The Bland-Altman plot is familiar and well accepted in medical research; however, age-biased plots are well established in depicting age agreement and have the advantage of depicting ages as discrete classes relative to a reference value (Campana 2001).

Indices of Precision

A number of indices of agreement are used to summarize age agreement between paired ages. In the case of the Yellowtail Flounder example, these are:

Index	Scales-Whole	Whole-Sectioned	Scales-Sectioned
PA	55.6	51.8	66.7
APE	6.2	7.0	4.6
CV	8.8	9.9	6.5

The index of percent agreement (PA) is the easiest index to calculate and understand:

$\mathrm{PA} = 100 \times \frac{F}{N}$, where *F* is the number of fish whose paired ages agreed, and *N* is the number of fish whose age were estimated. The table above presents these indices as calculated in exact years. This can also be calculated within 1, 2, or 3 years (92.6, 96.3, 100%, respectively, for the Yellowtail Flounder PA between scales and whole otoliths). Higher values convey higher agreement. Although simple and relatively intuitive, a high value for a short-lived fish is not comparable to a similar value for a much longer-lived fish, so this index is rarely reported alone. It is one of two indices used by the NEFSC aging program.

Beamish and Fournier (1981) introduced the index of average percent error (APE) as an index of precision that is less dependent on absolute age of the fish than PA: $\mathrm{APE} = 100 \times \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{R}\sum_{i=1}^{R}\frac{\left|Y_{ij} - \overline{Y}_j\right|}{\overline{Y}_j}\right)$, where *N* is the number of fish aged, *R* is the number of replicated age estimates per fish, Y_{ij} is the *i*th age determination of the *j*th fish, and $\overline{Y}_j$ is the average age for the *j*th fish. Lower values convey higher agreement.

PA and APE are reported together by many aging laboratories, but the NEFSC uses the following, modified index instead.

Chang (1982) proposed using the index of the coefficient of variation (CV), $\mathrm{CV} = 100 \times \frac{1}{N} \sum_{j=1}^{N} \frac{\sqrt{\sum_{i=1}^{R} \frac{\left(Y_{ij} - \bar{Y}_j\right)^2}{R-1}}}{\bar{Y}_j}$, to eliminate the effect of fish age on measures of precision. This is very similar to an actual coefficient of variation (i.e., of replicate age estimates for an individual fish: $CV = s \times 100 / \bar{Y}$), where s is the standard deviation: $s = \sqrt{\sum_{i=1}^{R} \frac{\left(Y - \bar{Y}\right)^2}{R-1}}$. Similar to APE, lower values convey higher agreement. The formulation of CV, when the number of reads is two, makes it about 41% higher than APE (Chang 1982; Campana 2001; Appendix Table), so reporting both is redundant. The NEFSC typically reports PA and Chang's CV in summarizing the reproducibility of ages (http://www.nefsc.noaa.gov/fbp/age-prec/).

APE and Chang's CV may be reported for individual fish or by age class, especially if a particular individual or trend among age classes is suspected; however, they are typically reported as a single value for all fish aged. The simplicity of a single value makes these indices simple and useful, unless it obscures important variation between replicate reads for the same fish, among fish, and among age classes (Hoenig et al. 1995). In addition, there is no theoretical basis that a specific value of PA, APE, or CV serves as a threshold to accept or reject a sample (Campana 2001). These shortcomings of an index approach are addressed by using tests of symmetry, which use the $\chi 2$ statistic to determine if individual ages that do not agree are randomly

distributed among readers or age classes (the null hypothsis) or if they are asymmetrically distributed.

Tests of symmetry

There are three tests of symmetry available for age reading comparisons. The simplest test of symmetry was introduced by McNemar (1947) in relation to an *m* × *m* contingency table. The test statistic is calculated as: $\chi^2 = \frac{(\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(n_{ij} - n_{ji}))^2}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} n_{ij} + n_{ji}}$, where n_{ij} is the observed frequency in the *i*th row and the *j*th column and n_{ji} is the observed frequency in the *j*th row and *i*th column. This is also referred to as a 'maximally pooled' test of symmetry because it adds all the squared differences of the values on each size of the diagonal (i.e., yellow vs. green, see matrix below[4]) and divides that by the sum of the values on each size of the diagonal. The central diagonal (unmarked white spaces in the 3 × 3 age frequency tables, below) is not part of the calculation, and the degrees of freedom is always 1. In this simple example, $\chi^2 = \frac{(3+3-1)^2}{3+3+1}$ = 3.6 (df = 1; *P* = 0.059) – where χ^2 is the chi-squared statistic, df is the degrees of freedom (always 1 with McNemar's test), and *P* is the probability related to rejecting the null hypothesis of no asymmetry among cells off the diagonal. In this example 3 × 3 matrix, McNemar's test does not reject the null hypothesis using a criterion of α = 0.05.

[4] These simple example matrices are from Evans, G. T., and J. M. Hoenig. 1998. Testing and viewing symmetry in contingency tables, with application to readers of fish ages. Biometrics 54:620-629.

Age frequency tables depicting different pooling methods for calculating χ^2

Age Frequencies		
	3	0
0		3
1	0	

McNemar's maximally pooled

Age Frequencies		
	3	0
0		3
1	0	

Evans and Hoenig's diagonally projected

Age Frequencies		
	3	0
0		3
1	0	

Bowker's unpooled paired cells

Evans and Hoenig (1998) modified this test of symmetry approach to calculate the statistic as $\chi^2 = \sum_{p=1}^{m-1} \frac{\left(\sum_{i=1}^{m-p}\left(n_{p+j,j} - n_{j,p+j}\right)\right)^2}{\sum_{i=1}^{m-p}\left(n_{p+j,j} + n_{j,p+j}\right)}$, where $p = i - j$. They referred to this as a diagonally projected test of symmetry, because the values are summed along a series of diagonal cells (first yellow, then green, see 3 × 3 matrix) projecting outward from the central diagonal. The degrees of freedom equal the number of paired projected diagonals that have a difference in ages. In this example 3 × 3 matrix, Evans and Hoenig's $\chi^2 = \frac{6^2}{6} + \frac{-1^2}{1} = 7$ (df = 2 [the number of paired diagonal rows with non-zero values]; $P = 0.030$) and the null hypothesis is rejected.

The most common test of symmetry used in fish age studies is that of Bowker (1948), which calculates $\chi^2 = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{\left(n_{ij} - n_{ji}\right)^2}{n_{ij} + n_{ji}}$. Summation is done without pooling cells. In the matrix (below), all cells off the diagonal are used in the calculation (yellow, green, and blue). The degrees of freedom equal the number of values in at least one paired

cells. In this example 3 × 3 matrix, Bowker's $\chi^2 = \frac{3^2}{3} + \frac{-1^2}{1} + \frac{3^2}{3} = 7$ (df = 3 [the number of paired cells with non-zero values]; *P* = 0.07) and the null hypothesis is not rejected.

The results of the Yellowtail Flounder matrices show no evidence of asymmetry between scales and sectioned otoliths (below, right), whereas whole otoliths are significantly biased, particularly when examined with McNemar's test.

Index	Scales-Whole	Whole-Sectioned	Scales-Sectioned
McNemar	χ2 = 12 (df = 1; P = 0.0005)	χ2 = 9.3 (df = 1; P = 0.002)	χ2 = 0.1 (df = 1; P = 0.74)
Evans & Hoenig	χ2 = 12 (df = 3; P = 0.007)	χ2 = 9.4 (df = 4; P = 0.052)	χ2 = 1.5 (df = 2; P = 0.47)
Bowker	χ2 = 12 (df = 6; P = 0.06)	χ2 = 11 (df = 7; P = 0.14)	χ2 = 6.3 (df = 6; P = 0.39)

In practice, the three tests of symmetry may agree in rejecting the null hypothesis or not, because of differences in pooling the cells to calculate the statistic and the degrees of freedom. McNemar's maximally-pooled statistic is most sensitive to even small differences on one side of the diagonal, if there are many cells where these small differences accumulate. Evans and Hoenig's diagonally-projected statistic is most sensitive to a matrix with lots of imprecision projecting off at least one of the off-center diagonals. Bowker's unpooled statistic is most sensitive to just a few differences, even a single cell, if large enough.

Tabulations, plotting, or calculating indices or the χ^2 statistic are tedious to do by hand but fairly easy to do with software. The NEFSC uses a pre-programmed EXCEL workbook (http://www.nefsc.noaa.gov/fbp/age-prec/). Dr. Derek H. Ogle, Northland College, supports a number of R programs in the FSA package, version

0.4.3[5], for visualizing and calculating precision and bias in age agreements (https://www.rforge.net/doc/packages/FSA, see ageBias, ageComp, agePrec). Dr. Gary Nelson, Commonwealth of Massachussets Division of Marine Fisheries, supports a number of R programs in his fishmethods package, version 1.6-0, including compare2, written by Dr. John Hoenig, which compares paired sets of ages by Evans and Hoenig's and McNemar's tests of symmetry (http://cran.r-project.org/web/packages/fishmethods/index.html).

In summary, tabulations and graphical displays, indices of precision, and tests of symmetry are used to assess aging agreement between readers, methods, or periods of time. There are pros and cons to each assessment method, so aging laboratories typically report one or more types of each category: graphic, index, or test. The NEFSC routinely reports a tabulation of paired ages, plots an age-bias plot, and lists the PA and Chang's CV. If the CV is high, then Bowker's statistic is also calculated and evaluated to report bias.

[5] Version as of March, 2014.

Appendix Table. Summary of indices and tests of symmetry for 14 data sets available in the FSA data package maintained by Dr. Derek H. Ogle, Northland College, in the FSA package, version 0.1.4. (http://rforge.net/doc/packages/FSAdata/00Index.html).

Data	Sample		Indices			McNemar			Evans and Hoenig			Bowker		
	N	Range	PA	APE	CV	χ2	Df	*P*	χ2	df	*P*	χ2	df	*P*
AlewifeLH[a]	104	0-10	58.7	8.9	12.5	17.0	1	< 0.0001	22.1	4	0.0002	34.5	16	0.0047
BluefishAge[b]	445	0-10	87.0	1.6	2.3	0.6	1	0.4308	3.1	2	0.2077	11.6	10	0.3136
Croaker1[c]	317	0-11	93.1	0.6	0.9	0.0	1	1.0000	0.0	1	1.0000	10.6	8	0.2242
Morwong1[d]	217	0-12	52.1	4.1	5.8	17.0	1	< 0.0001	20.6	3	0.0001	42.1	30	0.0693
Morwong2[d]	136	3-23	70.6	2.4	3.3	1.6	1	0.2059	2.4	3	0.4936	20.6	19	0.3582
Morwong3[d]	58	3-13	89.7	1.4	2.0	6.0	1	0.0143	6.0	1	0.0143	6.0	2	0.0498
MulletBS[e]	51	2-6	29.4	13.7	19.4	36.0	1	< 0.0001	36.0	3	< 0.0001	36.0	7	< 0.0001
StripedBass4[g]	1202	2-20	61.8	2.8	4.0	9.2	1	0.0024	19.8	5	0.0013	72.7	37	0.0004
StripedBass5[g]	458	2-21	85.8	1.1	1.5	3.5	1	0.0628	3.5	2	0.1719	14.9	19	0.7271
StripedBass6[g]	451	2-20	55.4	4.3	6.1	37.7	1	< 0.0001	42.3	6	< 0.0001	98.9	38	< 0.0001
WalleyePS[h]	60	1-13	53.3	8.9	12.6	24.1	1	< 0.0001	24.3	7	0.0010	24.7	16	0.0759
YTFlounder[j] (sw)	27	2-7	55.6	6.2	8.8	12.0	1	0.0005	12.0	3	0.0073	12.0	6	0.0620
YTFlounder[j] (wc)	27	2-7	51.8	7.0	9.9	9.3	1	0.0023	9.4	4	0.0518	11.0	7	0.1386
YTFlounder[j] (cs)	27	2-7	66.7	4.6	6.5	0.1	1	0.7388	1.5	2	0.4723	6.3	6	0.3869

[a] Alewife (*Alosa pseudoharengus*) otoliths versus scales
[b] Bluefish (*Pomatomus saltatrix*) otoliths by 2 readers
[c] Atlantic croaker (*Micropogonias undulatus*) otoliths by 2 readers
[d] Jackass Morwong (*Nemadactylus macropterus*) otoliths reader twice by Reader A (1) and Reader B (2), and by Reader A versus B (3).
[e] Red Mullet (*Mullus barbatus ponticus*) whole versus broken-burnt otoliths
[g] Striped bass (*Morone saxatilis*) scales versus otoliths (1, 6), scales by 2 readers (4), otoliths by 2 readers (5).
[h] Walleye (*Sander vitreus*) sectioned otoliths versus scale impressions.
[j] Yellowtail Flounder (*Limanda ferruginea*) scales versus whole otoliths (sw), whole versus cross-sectioned otoliths (wc), and cross-sectioned otoliths versus scales (cs).