A statistical model for quantifying age-reading errors and its application to the Antarctic minke whales

TOSHIHIDE KITAKADO[1], CHRISTINA LOCKYER[2] AND ANDRÉ E. PUNT[3]
Contact email: kitakado@kaiyodai.ac.jp

ABSTRACT

A statistical method for quantifying age-reading error, i.e. the extent of bias and inter-reader variability among readers, is introduced. The method assumes the availability of an independent control reader who produces reference ages for ageing structures which are also read by the subject readers. This reader is assumed to provide unbiased age estimates. Linear structures in bias and variance are incorporated in a conditional probability matrix representing the stochastic nature of age-determination for each reader. A joint likelihood function for the parameters related to ageing bias, variance and nuisance parameters is defined based on observed ageing outcomes from both the control and subject readers. The method is applied to data for Antarctic minke whales taken during Japanese commercial (1971/72-1985/86) and scientific (1986/87-2004/05) whaling. 250 earplugs selected according to a predetermined protocol were used in the analyses to estimate the inter-reader variation for four Japanese readers. One of the authors acted the control reader. The Japanese readers and the control reader differed in terms of both the expected age given the true age, and variance in age-estimates. The expected age and random uncertainty in age-estimates differed among the Japanese readers, although the two readers in charge of age-reading for samples taken during Japanese scientific whaling (JARPA I and JARPA II) provided quite similar ageing outcomes. These results contribute to analyses using catch-at-age data for this species. It should also be noted that the model and approach in this paper can be applied to populations other than the Antarctic minke whales, if a control reader is available, even retrospectively.

KEYWORDS: ANTARCTIC MINKE WHALES; AGE-READING ERROR; EARPLUGS;

INTRODUCTION

The primary source of information on the abundance of many cetacean populations is estimates of abundance from sightings surveys (e.g. Branch, 2007) and from mark-recapture studies (e.g. Larsen and Hammond, 2006). This information allows an evaluation of recent (up to the last 20-30 years) trends in abundance when a time-series of comparable estimates is available. However, inferences regarding the status of populations relative to management reference points are more precise if estimates of abundance and information from other data sources are used to fit population dynamics models. Although a variety of types of population models has been applied to cetacean populations, most of those presented to the IWC Scientific Committee have been age- and sex-structured.

Population dynamics models have been proposed as one way to test the hypothesis that the abundance of Antarctic minke whales (*Balaenoptera bonaerensis*) increased in abundance prior to the start of directed harvesting during the early 1970s, perhaps due to the 'krill surplus' which has been postulated to have arisen because of the substantial declines of species such as blue, fin and sei whales (e.g. Mori and Butterworth, 2006). The output from these models can also provide the information needed to assess the impact of

[1]Tokyo University of Marine Science and Technology, 5-7, Konan 4, Minato-ku, Tokyo 108-8477, Japan.
[2] The North Atlantic Marine Mammal Commission, P.O. Box 6453 Tromsø, N-9294, Norway,
[3] School of Aquatic and Fisheries Sciences, University of Washington, Seattle, WA 98195-5020

environmental factors on, for example, reproductive success, as is common in fisheries assessments (e.g. Maunder and Watters, 2003, but see Haltuch and Punt, 2011).

Age-structured assessment models can be divided into two major 'classes': Virtual Population Analysis (VPA), and 'integrated' methods. VPA methods assume that the catch age-composition is measured with negligible error compared to that associated with the remaining data sources such as estimates of abundance from sightings surveys, while 'integrated' methods allow for sampling and other sources of error with the age-composition data, albeit at the expense of greater complexity. Both of these classes of model have been applied to data for Antarctic minke whales (e.g. Mori *et al*., 2007; Mori and Butterworth, 2008; Punt and Polacheck (2005, 2006, 2007, 2008)), and both suggest an increase in abundance from ~1930 to ~1970. However, Punt and Polacheck (2005) found that there were substantial differences between growth curves estimated internally to the 'integrated' model based on the length and age data collected from the commercial catches or externally using JARPA data. Several possible explanations for this result have been explored, including time-varying growth and time-varying fishery selectivity. However, one key possible reason for this discrepancy is age-reading error.

Age-reading error can be divided into age-reading bias (i.e. the expected age assigned to ageing structures, in this case, ear plugs, for animals of a given age, differ from the actual age) and random age-reading error (i.e. variation about the expected age for a given ageing structure). Both of these types of error can be consequential for assessments based on population models. For example, Reeves (2003) found that random age-reading error led to 'smoothing' of recruitment estimates (i.e. large year-classes 'smoothed' to adjacent, less abundant, year-classes). More, importantly perhaps, a 'drift' of ageing methods could have led population models to estimate spurious trends in recruitment for the Antarctic minke whales (Butterworth and Punt, 2009).

The impact of age-reading error can be included in assessments by specifying a matrix which defines the conditional probability of an animal of true age *a* being aged to be that age or some other age, a', $P(a'|a)$. The model-predictions upon which the likelihood component in the 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 age-reading error. Given $P(a'|a)$, this prediction would be:

$$C_{a'} = \sum_a P(a'|a) C_a \tag{1}$$

where 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. Therefore, it is crucial for the analyses which use catch-at-age data to have information on $P(a'|a)$ for each age reader.

This paper introduces a method for quantifying age-reading errors. It was applied to data from an age-reading experiment conducted for the Antarctic minke whales. The experiment involved one reader (Lockyer) reading 250 ear plugs using a protocol designed by the Scientific Committee of the IWC (Butterworth and Punt, 2009) and comparing the resulting age-estimates with values obtained from past and current age readings by Japanese scientists.

METHODS

The age-reading experiment

The experiment involved reading 250 ear plugs from female minke whales caught in Antarctic Area IV (Table 1). The plugs were chosen from five groups of years (50 from each group), corresponding to periods near the start and the end of commercial whaling, and the start, middle and end of JARPA sampling (1974/5-1976/7, 1982/3-1984/5, 1989/90-1991/2, 1997/8-1999/2000 and 2003/4-2005/6; referred to as Periods I-V respectively). For this random selection, all the female plugs for the period concerned were allocated a sequential number starting at 1 and culminating at N. A random number was then drawn randomly from [1,

N] and that plug selected, unless it was seen to be damaged or to have deteriorated in quality, in which case another random draw was made. The length of the whale, previous age readings, and the names of the original age readers (Kato, Masaki or Zenitani) were recorded for the selected whale. In addition to these three readers, a new reader (Bando) who aged all of the samples taken during JARPAII so far (2005/06-2010/11) also read 100 recent samples three times so that his reading outcomes could be standardized against the others.

One of us (Lockyer) read all of the plugs twice, with randomized sample order both initially and after each complete set of readings. Fifty of the plugs (10 from each of the five periods), again selected at random, were read a third time. Another two readers (Zenitani and Bando) also read each plug three times. All readings were blind, i.e. the reader had no knowledge of other data pertaining to the whale from which the plug was taken.

Lockyer was unable to obtain an age for all of the plugs (Table 2). Age estimates could be obtained for more than 86% of the plugs for each trial, although the proportion of plugs which could be read decreased between the first and second trials. All the records of ages from the Japanese readers were "valid" except for 1% of values by Bando, which were recorded as "minimum". The bulk of the analyses is based on the "valid" readings only, although sensitivity tests consider the use of the data from the other categories in Table 2 (see below).

Statistical Analysis

Conditional probability for age-reading errors

Suppose that two groups of readers independently obtain age-estimates using a common set of *n* samples (here *n* =250). Group 1 consists of only one reader (Lockyer), who conducted ageing at most three times for all the *n* samples. Group 2 consists of four readers (Masaki, Kato, Zenitani and Bando). Masaki, Kato and Zenitani read different plugs from different time periods, whereas Bando read the 100 samples from Periods IV and V, which Zenitani also read. The sample sizes for Masaki and Kato were respectively n_M =50 and n_K =78 (Table 1), and Zenitani and Bando read plugs three times for their samples of n_Z =122 and n_B =100, respectively.

Let a_{1jk} $(j=1,2,...,n;k=1,...,r_j)$ be the observed ages by Group 1 (assuming that it is a "valid" count) of the *j*-th sample during the *k*-th of r_j trials (r_j =2 or 3). Similarly, for Group 2, let a_{2j} $(j=1,2,...,n_M)$ and a_{2j} $(j=n_M+1,2,...,n_M+n_K)$ respectively denote the observed ages by Masaki and Kato and a_{2jk} $(j=n_M+n_K+1,2,...,n;k=1,2,3)$ denote the observed counts for the *j*-th sample during the *k*-th trial by Zenitani. Also, let $\tilde{a}_{2jk}$ $(j=151,...,n;k=1,2,3)$ be the readings by Bando. As noted above, Lockyer did not assign "valid" ages to all of the samples during all of the trials (Table 2). In such cases, the notation changes accordingly. For example, the data for the "either" or "interval" categories can be denoted $a_{1jk}^{(1)}$ and $a_{2jk}^{(2)}$ respectively (for "either" the age is either $a_{1jk}^{(1)}$ or $a_{1jk}^{(2)}$, while for "interval" the age-estimate is between $a_{1jk}^{(1)}$ and $a_{1jk}^{(2)}$). The "minimum" counts by Lockyer and Bando were treated in a similar manner.

Now, consider the joint probability distribution of the observations. Let $b_i(a;\phi)$ and $\sigma_i(a;\phi)$, respectively, denote the expected age and standard deviation for the age-estimates for the *i*-th Group for an animal of true age *a*, where ϕ is a vector of unknown parameters. The variability in ageing is expressed as a matrix form $\{P_i(a'|a;\phi)\}_{a,a'=L,...,H}$, where:

$$P_i(a'|a;\phi) \propto \exp\left[-\frac{(a'-b_i(a;\phi))^2}{2\sigma_i^2(a;\phi)}\right] \quad (2)$$

is the conditional probability that the *i*-th group draws ageing outcomes *a*' given that the true age of the animal is *a*, and $\sum_{a'=L}^{H} P_i(a' \mid a;\phi) = 1$ for all *a* (Punt *et al*. 2008), where *H* and *L* are respectively maximum and minimum possible ages.

The expected age for Reader 1 is assumed to be proportional to the true age:

$$b_1(a) = (1+x)a \tag{3}$$

On the other hand, the expected age for the readers in Group 2 is a linear function of true age *a*:

$$b_2(a;\phi) = b_L + (b_H - b_L)\frac{a-L}{H-L}\,. \tag{4}$$

This is a 2-parameter model from Punt *et al*. (2008). The parameters of equation (4) should relate to each reader when considering hypotheses related to reader effects. The values of *L* and *H* are pre-specified (*L* =0 and *H*=70) and are not estimated.

The functional form of the ageing error standard deviation for the two Groups is also assumed to be a linear function of true age:

$$\sigma_i(a;\phi) = \sigma_{iL} + (\sigma_{iH} - \sigma_{iL})\frac{a-L}{H-L} \qquad (i = 1,2) \tag{5}$$

As for the expectation, the parameters in the equation (5) are specific to the reader concerned.

Likelihood function

Let $\beta = (\beta_L, \ldots, \beta_H)$ be the true age composition of sampled animals, which is unknown. The contribution of *j*-th sample by Reader 1 to the likelihood given the true age (say a) is:

$$P_1(a_{1j} \mid a;\phi_1) = \sum_{k=1}^{r_j} P_1(a_{1jk} \mid a;\phi_1) \tag{6}$$

where $a_{1j} = (a_{1j1}, ..., a_{1jr_j})$ and ϕ_1 is the vector of parameters of interest. By considering the distribution for Group 2 in a similar way, the joint probability distribution of ageing outcomes by the two groups is provided by a mixture form as:

$$\Pr(a_{1j}, a_{2j};\phi,\beta) = \sum_{a=L}^{H} \beta_a\, P_1(a_{1j} \mid a;\phi_1)\, P_2(a_{2j} \mid a;\phi_2) \qquad (j = 1,2,...,150) \tag{7}$$

$$\Pr(a_{1j}, a_{2j}, \tilde{a}_{2j};\phi,\beta) = \sum_{a=L}^{H} \beta_a\, P_1(a_{1j} \mid a;\phi_1)\, P_2(a_{2j} \mid a;\phi_2) P_2(\tilde{a}_{2j} \mid a;\phi_2) \qquad (j = 151,...,250) \tag{8}$$

where ϕ_2 is the parameter vector for Group 2 and $\phi = (\phi_1, \phi_2)$. Finally, the full likelihood function for the parameters is :

$$Like(\phi,\beta) = \prod_{j=1}^{150} \Pr(a_{1j}, a_{2j};\phi,\beta) \prod_{j=151}^{n} \Pr(a_{1j}, a_{2j}, \tilde{a}_{2j};\phi,\beta)\,. \tag{9}$$

The likelihood contribution for the data that are not in the "valid" category can be expressed as follows; for example, when a data type is "interval" as $[a_{1jk}^{(1)}, a_{1jk}^{(2)}]$, the distribution is:

$$P_1([a_{1jk}^{(1)}, a_{1jk}^{(2)}] \mid a;\phi) = \sum_{a'=a_{1jk}^{(1)}}^{a_{1jk}^{(2)}} P_1(a' \mid a;\phi). \quad (10)$$

The parameters in the expectation and variance structures are of interest in this model, whereas $\beta_L,\ldots,\beta_H$ are nuisance parameters. To make the estimation easier and to reduce the number of nuisance parameters, a functional constraint is incorporated on the parameters for the true age composition of the sample β_a $(a \geq A)$ as $\beta_a = \beta_A \exp(-Z(a-A))$, where A is the largest number which satisfies:

$$\frac{\#\{j = 1,\ldots,n \mid a_{1j1} \geq A\}}{n} > q, \quad (11)$$

and Z is a mortality parameter. The threshold value q is, of course, *ad hoc*, but the constraint is nevertheless useful in cases such as this experiment. We use a value q = 0.20 as a base case assumption.

Scenarios

Table 3 lists the scenarios considered in this paper. Lockyer is taken to be the control reader for Cases 1, 2, 3 and 5, and Zenitani for Case 4. Case 5 examines the sensitivity of the results to using all of the data not only the “valid” data, but also the "either", "interval" and "minimum" data. Several alternative models are considered based on the covariates included in the models for the mean and variance structures for age-reading Group 2 (see Table 4).

RESULTS

Lockyer’s age-reading

A sample of 100 specially selected ear plugs, independent of the experimental sample, was made available to Lockyer, who had expressed a wish to undertake a “trial” reading of minke whale ear plugs in general during 1-2 December 2009. Lockyer read 50 ear plugs from this sample to become re-familiarised with the Growth Layer Group (GLG) counting methods for this species. Although the specimens bore their true ID numbering, they were read “blind”. The results of this trial, although not part of the experimental design, helped to refine the design of a proposed age recording form.

The first reading began on 2 December in the afternoon, and continued each day until completion, with readings on 3, 4, 6 and 7 December 2009, with approximately 50 ear plugs read each day or a maximum of 70 on any one day, with breaks every two hours to rest the eyes. A Nikon binocular microscope was used to examine all ear plugs with an eye objective 10xB22 and zoom magnification x0.8-x8 facility. Even at maximum magnification, it was only just possible to read all GLGs at the plug base of some older animals. Five ear plugs were placed in water in separate petri dishes with individual labelling for examination at any one time. These were then replaced in sample jars before the next set of five.

The second readings began after a 2-day break on 10 December and continued on 11, 12, and 15 December. A break was then taken 16 December, before the third reading of a sub-set of 50 ear plugs took place. These readings were completed on t17 December.

An excel data book (Appendix 1) was updated regularly (usually after reading 10 plugs) throughout the experimental readings compiling all information written on the working form. This also helped to make convenient breaks between each microscope use, and avoid monotony. Reading efficiency depends greatly on the degree of alertness, and the day’s reading session was terminated on two occasions because of onset of tiredness.

The colour of the ear plugs varied from pale ivory through tan to dark brown. The ear plugs for young animals were usually pale cream while most ear plugs for older animals appeared dark. However, this was not always consistent. The pale colouration frequently made it difficult to discern any GLG differentiation,

and plugs for apparently very young animals were often very difficult to age. In addition, accessory laminae were sometimes present and led to difficulties with age determination. For this reason, occasionally two possible alternative readings were provided because the reader could not be certain which to choose. Normally, not in an experimental setting, one might refer to biological data to help resolve such issues.

Statistical analysis

Histograms and scatter plots of the "valid" age-reading outcomes from Lockyer do not suggest evidence for between-trial bias (Figure 1). Similarly, there is no evidence for between-trial bias for Zenitani (Figure 2) and Bando (Figure 3). Consequently, trial was not considered as a covariate in the analyses. The ageing outcomes of the two primary readers for JARPA (Zenitani) and JARPA II (Bando) appeared similar (see Figure 4).

Figure 5 plots the ageing outcomes for each of the Japanese readers ("best" estimates for Masaki and Kato, and the medians of the three estimates for Zenitani and Bando) against the age-estimates by Lockyer. These plots indicate a consistent discrepancy between the age-estimates obtained by Lockyer and those obtained by the Japanese scientists. In fact, under Case 1, where Lockyer's bias is assumed to be zero, the estimated ages by the four Japanese readers appear negatively biased (solid lines in Figure 5). Figure 6 shows the difference in absolute and relative biases among the Japanese readers against the control reader. The standard errors and coefficient of variation for the control and Japanese readers also differ (see Figure 7).

Table 5 summarizes the results of the parameter estimation and model selection for the different models under Case 1. Incorporating a reader effect into the mean component tended to improve the goodness of fit substantially (in terms of model selection criteria) compared to incorporating these effects into the variance structure (i.e. the extent of random age-reading error). Model 3, in which the reader effects were incorporated into both the mean and variance structures, led to the most parsimonious fit to the data. The adequateness of the fits for Model 3 in Case 1 is also confirmed by Figure 5.

Table 6 provides estimates of parameters which could be used to compute ageing error matrices. It should be noted that the differences in parameter estimates between Cases 1 (base case) and 5 (which use data for indexes 0-3 in Table 2) are almost negligible.

Ageing-error matrices based on Model 3 could be incorporated into assessments of the impact of age-determination error on the outputs from age-structured models for Antarctic minke whales (e.g. Punt, 2010, Punt et al., 2013). It should be noted that the analyses of this paper are predicated on Lockyer's age-estimates. It cannot necessarily be assumed that Lockyer provides unbiased estimates of age. Overall, the results suggest that the age-reading errors for Lockyer and the four Japanese readers differ.

DISCUSSION

We have introduced a statistical method for quantifying age-reading error and the extent of inter-reader variability between the subject readers and applied it to a data set for the Antarctic minke whales. Availability of the independent control reader played an important role in standardizing outcomes by the subject readers. It is not possible to evaluate possible biases in the control reader using this experiment, and therefore we conducted sensitivity analyses to determine how the estimates of the parameters of the model change.

The results suggested that the expected age and random uncertainty in age-estimates differed among the Japanese readers, although the two readers in charge of age-reading for JARPA and JARPA II provided quite similar ageing outcomes. This is likely because the new reader Bando had a training period to develop his reading skill using JARPA samples which had previously been aged by Zenitani (these samples were, of course, not chosen from the 250 samples on which this study is based).

The analyses of this paper assumed that the age-composition of the catch was the same over more than 30 years to reduce the number of nuisance parameters. However, the assessment of age-reading errors is subject to confounding if the catch age-composition changes over time. The impact of possible violation of this assumption was examined by assuming different age-compositions between Periods I-III and Periods IV and V. The fit of the model to data was better than for the base-case, but Model 3 remained the best model. The values of parameters which determine age-reading erros were also quite similar (see Figure 8) suggesting that how the nuisance parameters are treated only has a small impact on the final results.

The original motivation of the experiment and analysis was to provide quantitative information on age-reading error for use in the statistical catch-at-age analysis. Results by Punt et al. (2013) confirm that the results of this analysis are sensitive to whether age-reading error is accounted for.

It should also be noted that the model and approach shown in this paper are applicable to populations other than the Antarctic minke whales provided that a control reader is available, even retrospectively as was the case in this study.

ACKNOWLEDGEMENTS

The authors are also grateful to Hidehiro Kato, Ryoko Zenitani, Takeharu Bando, Toshiya Kishiro, Hikari Maeda and Yui Zennyoji for their preparation of Lockyer's age-reading experiment according to the protocol. Hidehiro Kato, Ryoko Zenitani and Takeharu Bando are especially thanked for their dedicated coordination of the experiment and giving the authors helpful information on past Japanese age-reading experiments. The IWC provided support to Lockyer under contract.

REFERENCES

Branch, T.A. 2007. Abundance of Antarctic blue whales south of 60°S from three complete circumpolar sets of surveys. *J. Cetacean Res. Manage*. 9:253-62.

Butterworth, D.S. and Punt, A.E. 2009. Proposed further work to aid resolution of questions concerning ageing of Antarctic minke whales. Appendix 4 to the Report of the Sub-Committee on In-Depth Assessment. *J. Cetacean Res. Manage*. (*Suppl*. 1) 11:209.

Haltuch, M.A. and Punt, A.E. 2011. The promises and pitfalls of including decadal-scale climate forcing of recruitment in groundfish stock assessment. *Can. J. Fish. Aquat. Sci*. 68: 912-26.

Larsen, F. and Hammond, P.S. 2006. Distribution and abundance of West Greenland humpback whales (*Megaptera novaeangliae*). J. Zoology 263: 343-58.

Maunder, M.N. and Watters, G.M. 2003. A general framework for integrating environmental time series into stock assessment models: Model description, simulation testing, and example. *Fish. Bull*. 101: 89-99.

Mori, M. and Butterworth, D.S. 2006. A first step towards modelling the krill-predator dynamics of the Antarctic ecosystem. CCAMLR Science, 13:217-277

Mori, M., Butterworth, D.S. and Kitakado, T. 2007. Further progress on application of ADAPT-VPA to Antarctic minke whales. Paper SC/59/IA13 presented to the IWC Scientific Committee, May 2007 (unpublished). 32pp.

Mori, M. and Butterworth, D.S. 2008. Some modifications to the current ADAPT-VPA model for Antarctic minke whales Paper SC/60/IA13 presented to the IWC Scientific Committee, June 2008 (unpublished). 5pp.

Polacheck, T. and A.E. Punt. 2006. Minke whale growth models for use in statistical catch-at-age models. Paper SC/58/IA3 presented to the IWC Scientific Committee, May 2006 (unpublished). 36pp.

Punt, A.E. 2010. A note on the impact of accounting for age-determination error on the outcome of the statistical catch-at-age analysis for Antarctic minke whales. Paper SC/60/IA6 presented to the IWC Scientific Committee, June 2008 (unpublished). 6pp.

Punt, A.E., Bando, T., Hakamada, T. and Kishiro, T. 2013. Assessment of Southern Hemisphere Minke Whales using Statistical Catch-at-age Analysis. Paper SC/65a/IA1 presented to the IWC Scientific Committee, June 2013(unpublished).

Punt, A.E. and Polacheck, T. 2005. Application of statistical catch-at-age analysis for Southern Hemisphere minke whales in Antarctic Areas IV and V. paper SC/57/IA9 presented to the IWC Scientific Committee, June 2005 (unpublished). 40pp.

Punt, A.E. and Polacheck, T. 2006. Further statistical catch-at-age analyses for Southern Hemisphere minke whales. Paper SC/58/IA2 presented to the IWC Scientific Committee, May 2006. (unpublished). 40pp.

Punt, A.E. and Polacheck, T. 2007. Further development of statistical catch-at-age models for southern hemisphere minke whales. Paper SC/59/IA4 presented to the IWC Scientific Committee, May 2007 (unpublished). 42pp.

Punt, A.E. and Polacheck, T. 2008. Further analyses related to the application of statistical catch-at-age analysis to Southern Hemisphere minke whales. Paper SC/60/IA2 presented to the IWC Scientific Committee, June 2008 (unpublished). 46pp.

Punt, A.E., Smith, D.C., Tuck, G.N. and Methot, R.D. 2006. Including discard data in fisheries stock assessments: Two case studies from South-eastern Australia. *Fish. Res*. 79:239–50.

Reeves, S.A. 2003. A simulation study of the implications of age-reading errors for stock assessment and management advice. ICES J. Mar. Sci. **60**: 314–328.

Table 1. The number of samples employed in this experiment

	Period	Group 1	Group 2			
		Lockyer*	Masaki	Kato	Zenitani	Bando
Period I	74/75-76/77	50 (10)	50	0	0	0
Period II	82/83-84/85	50 (10)	0	50	0	0
Period III	89/90-91/92	50 (10)	0	28	22	0
Period IV	97/98-99/00	50 (10)	0	0	50	50
Period V	03/04-05/06	50 (10)	0	0	50	50

* The numbers in brackets indicate how many plugs were read three times by Lockyer.

Table 2. Types of data provided by Lockyer

Index	**Category**[$]	**Data type**	**1st trial**	**2nd trial**	**3rd trial**
0	Valid	Age	228 (91.2%)	216 (86.4%)	43 (86%)
1	Either	Age1 or Age 2	2 (0.8%)	4 (1.6%)	0
2	Minimum	Age >=	11 (4.4%)	10 (4.0%)	3 (6.0%)
3	Interval	(Age1, Age2)	1 (0.4%)	0	0
4	may be missing	Age	1 (0.4%)	0	0
10	Uncertain	Age	2 (0.8%)	12 (4.8%)	3 (6.0%)
100	unreadable	NA	5 (2.0%)	8 (3.2%)	1 (2.0%)

[$] "valid": a single age was recorded; "either": two possible ages were offered; "minimum": only a minimum age was counted; "interval": a range of possible ages was given; "missing": part of the plug was missing; "uncertain": the reader was not confident in the age estimate. The numbers in brackets are percentages for each trial.

Table 3. The scenarios considered in the analyses.

	Bias in control reader	**Data**
Case 1 (Base)	Lockyer; 0%	"valid" only
Case 2	Lockyer; 10%	"valid" only
Case 3	Lockyer; -10%	"valid" only
Case 4	Zenitani; 0%	"valid" only
Case 5	Lockyer; 0%	Index=0,1,2,3

Table 4. List of assumptions regarding the covariate effects and parameters. In all the models, constraints $\sigma_{iH} \leq \sigma_{iH}$ for all the readers are incorporated.

Model	Assumption
0	No reader effects
1	Reader effects in Group 2 only in the mean structure
2	Reader effects in Group 2 only in the variance structure
3	Reader effects in Group 2 both in the mean and variance structures
4	Reader effects both in the mean and variance structures, but the expected ages by the recent two readers, Zenitani and Bando, are same

Table 5. Results of the analysis (Upper value=estimate; Lower value= SE) for the Case 1, where ageing by the control reader (Lockyer) is assumed to be unbiased. Note that the column "#parameters" does not include the number of nuisance parameters for the age composition.

Model	Loglike	#parameters	Δ-AIC	Δ-AICc	Reader 1 (Lockyer)		Reader 2-1 (Masaki)		Reader 2-2 (Kato)		Reader 2-3 (Zenitani)		Reader 2-4 (Bando)	
					bL1	bH1	bL21	bH21	bL22	bH22	bL23	bH23	bL24	bH24
0	-3004.0	6	106.18	104.54	0	70	1.39	61.17						
							0.10	0.51						
1	-2958.6	12	27.38	26.50	0	70	2.65	60.51	2.33	56.21	1.03	62.92	1.58	63.93
							0.44	2.06	0.38	1.27	0.10	0.55	0.10	0.57
2	-2978.1	12	66.40	65.52	0	70	1.68	61.86						
							0.28	0.76						
3	-2938.9	18	0.00	0.00	0	70	3.08	58.79	2.45	56.01	1.03	62.85	1.64	63.64
							0.53	1.91	1.36	4.10	0.09	0.54	0.11	0.61
4	-2964.7	16	47.60	47.29	0	70	3.10	58.74	2.63	55.49	1.24	63.07		
							0.53	1.90	0.60	2.00	0.08	0.52		

Model	Reader 1 (Lockyer)		Reader 2-1 (Masaki)		Reader 2-2 (Kato)		Reader 2-3 (Zenitani)		Reader 2-4 (Bando)	
	sigL1	sigH1	sigL21	sigH21	sigL22	sigH22	sigL23	sigH23	sigL24	sigH24
0	1.62	3.43	0.56	4.17						
	0.18	0.66	0.06	0.34						
1	1.64	3.36	0.52	3.98						
	0.17	0.60	0.06	0.28						
2	1.46	3.38	1.66	1.66	0.08	9.53	0.57	3.39	0.84	3.22
	0.15	0.50	0.27	0.27	0.35	1.56	0.08	0.41	0.12	0.44
3	1.55	3.14	1.55	1.55	0.54	7.37	0.46	3.66	0.66	3.41
	0.17	0.51	0.24	0.24	1.64	4.40	0.06	0.38	0.09	0.40
4	1.50	3.21	1.57	1.57	0.75	6.85	0.42	3.83	0.91	3.07
	0.15	0.49	0.24	0.24	0.44	1.39	0.06	0.38	0.11	0.42

Table 6. Results of the analysis (Upper value=estimate; Lower value= SE) for the Cases 1-5 under the best model (Model 3). The values in italics are pre-specified.

Model 3	Reader 1 (Lockyer)		Reader 2-1 (Masaki)		Reader 2-2 (Kato)		Reader 2-3 (Zenitani)		Reader 2-4 (Bando)	
	bL1	bH1	bL21	bH21	bL22	bH22	bL23	bH23	bL24	bH24
Case 1	*0.00*	*70.00*	3.08	58.79	2.45	56.01	1.03	62.85	1.64	63.64
(Lockyer: unbaised)			0.53	1.91	1.36	4.10	0.09	0.54	0.11	0.61
Case 2	*0.00*	*77.00*	3.10	64.27	2.67	60.66	1.09	69.04	1.70	69.85
(Lockyer: 10% bias)			0.52	2.11	0.56	2.16	0.10	0.64	0.12	0.70
Case 3	*0.00*	*64.00*	3.08	53.22	2.62	50.27	1.14	56.35	1.74	57.15
(Lockyer: -10% bias)			0.53	1.66	0.65	1.85	0.09	0.49	0.11	0.54
Case 4	0.00	75.52	3.00	63.51	2.68	59.58	*0.00*	*70.00*	0.71	70.26
(Zenitani: unbaised)	0.00	0.62	0.52	2.11	0.55	2.10			0.12	0.62
Case 5	*0.00*	*70.00*	3.01	59.03	2.36	55.99	1.02	62.65	1.63	63.41
(Case 1 with Index=0~3)			0.51	1.87	0.47	1.78	0.09	0.57	0.11	0.63

Reader 1 (Lockyer)		Reader 2-1 (Masaki)		Reader 2-2 (Kato)		Reader 2-3 (Zenitani)		Reader 2-4 (Bando)	
sigL1	sigH1	sigL21	sigH21	sigL22	sigH22	sigL23	sigH23	sigL24	sigH24
1.55	3.14	1.55	1.55	0.54	7.37	0.46	3.66	0.66	3.41
0.17	0.51	0.24	0.24	1.64	4.40	0.06	0.38	0.09	0.40
1.47	3.44	1.56	1.56	0.75	7.46	0.45	4.10	0.69	3.63
0.16	0.58	0.24	0.24	0.40	1.50	0.06	0.43	0.09	0.46
1.49	3.02	1.56	1.56	0.73	6.25	0.48	3.33	0.63	3.22
0.15	0.45	0.24	0.24	0.51	1.35	0.06	0.36	0.09	0.35
1.87	2.65	1.48	1.48	0.70	7.44	0.41	3.97	0.64	3.62
0.16	0.51	0.25	0.25	0.38	1.45	0.06	0.41	0.10	0.44
1.60	3.05	1.49	1.49	0.45	7.40	0.47	3.58	0.65	3.51
0.15	0.48	0.24	0.24	0.41	1.47	0.06	0.37	0.09	0.42

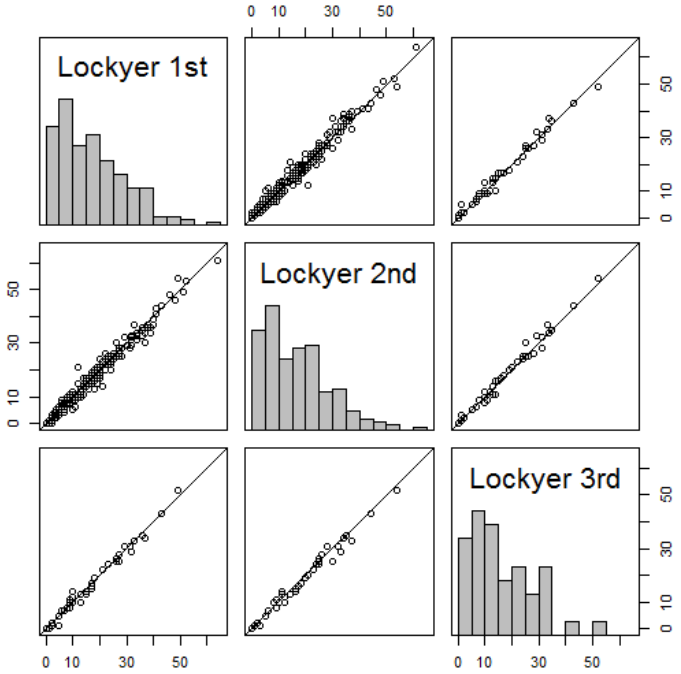

Figure 1. Scatter plots and histograms for Lockyer's ageing data for her three trials.

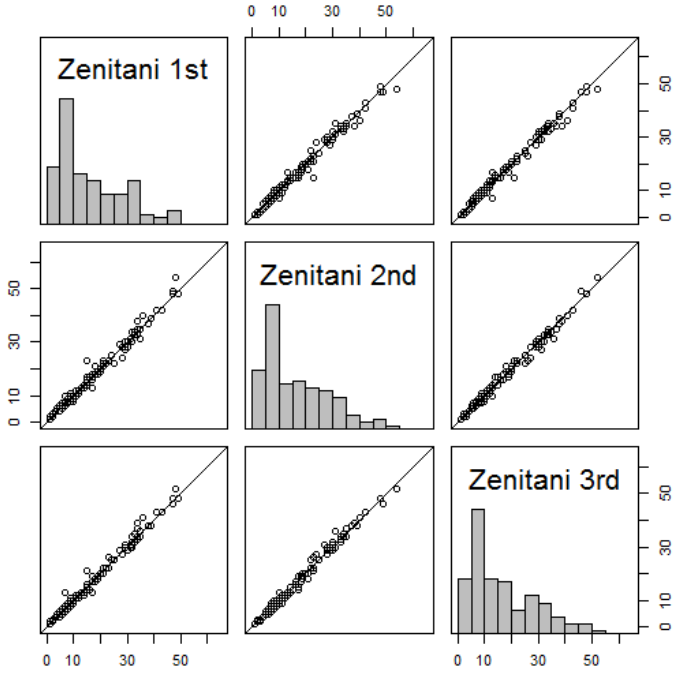

Figure 2. Scatter plots and histograms for Zenitani's ageing data for her three trials.

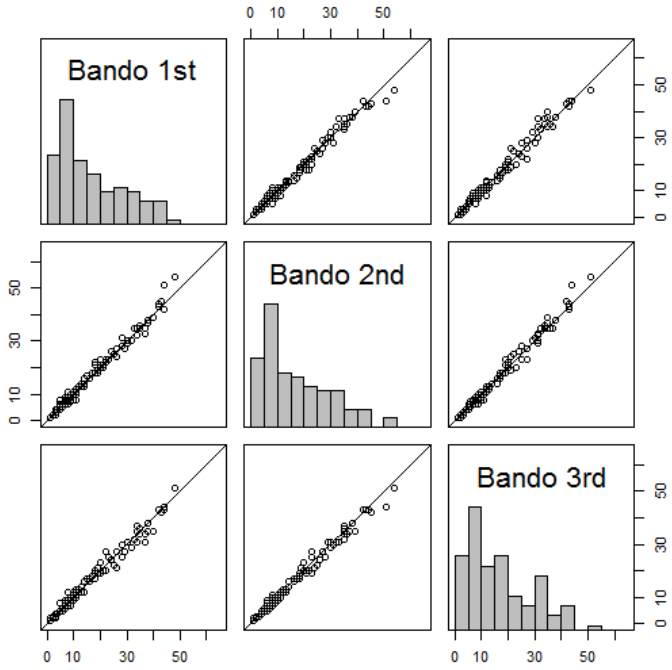

Figure 3. Scatter plots and histograms for Bando's ageing data for his three trials.

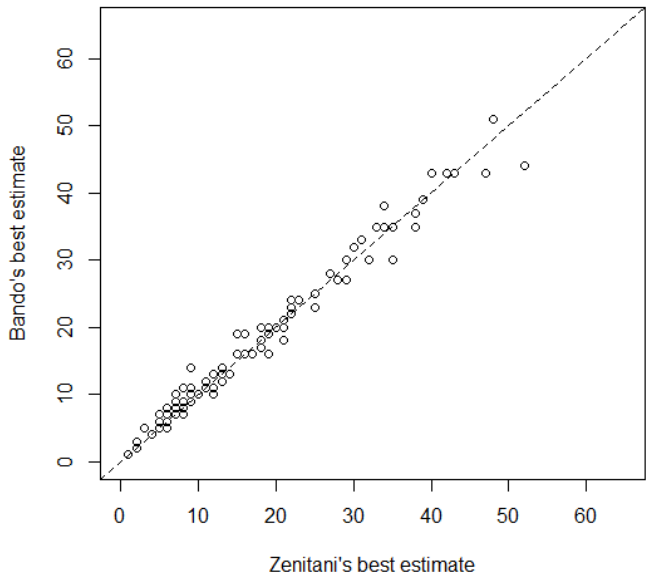

Figure 4. Scatter plot of the best estimates for the two primary readers in JARPA and JARPA II (Zenitani and Bando).

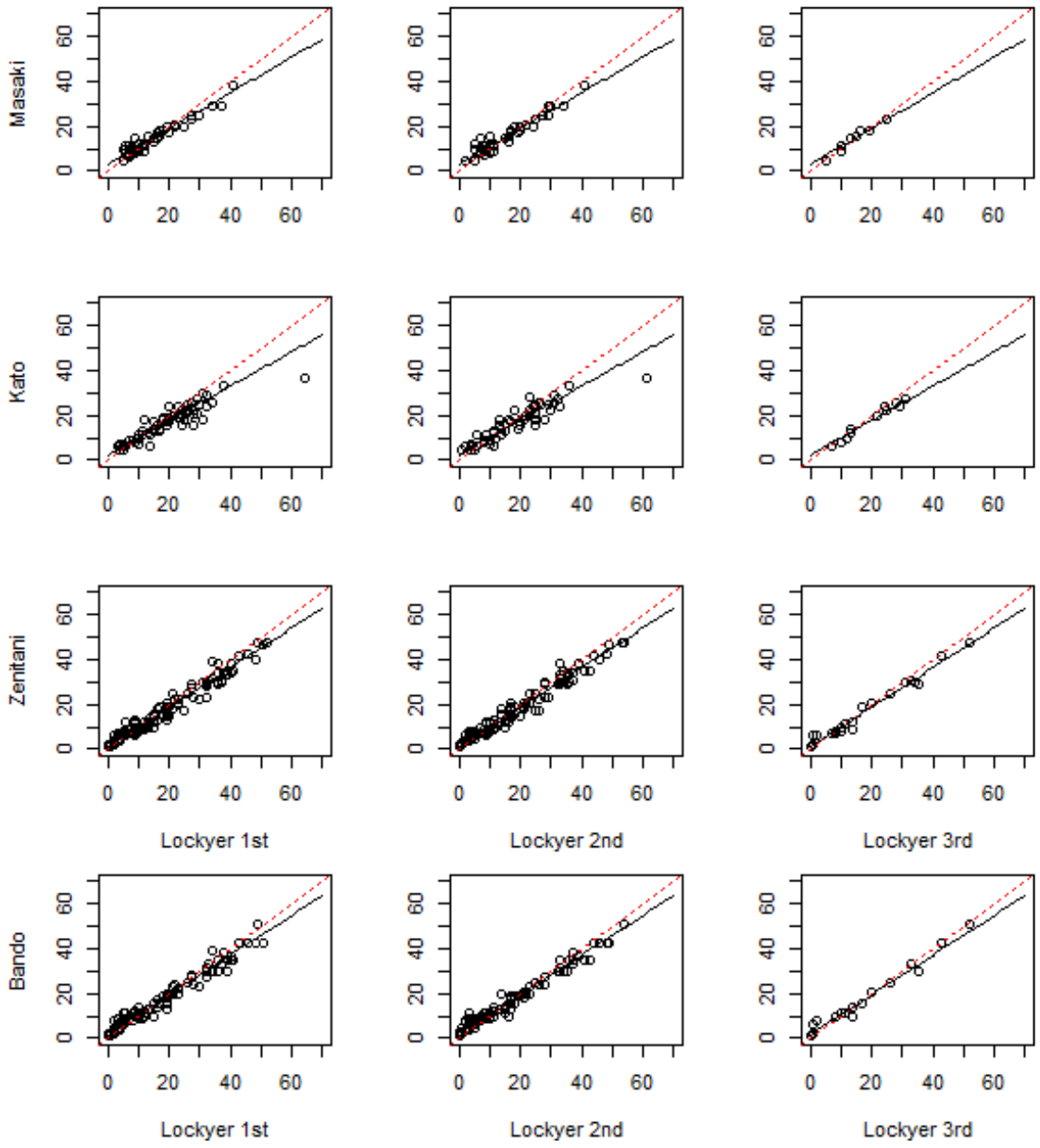

Figure 5. Scatter plots of the "best" age-estimates from the four Japanese readers against Lockyer's 1st, 2nd and 3rd trials ("valid" data only). The dashed lines show the 1-1 lines. Lockyer's age-estimation is assumed to be unbiased (Case 1).

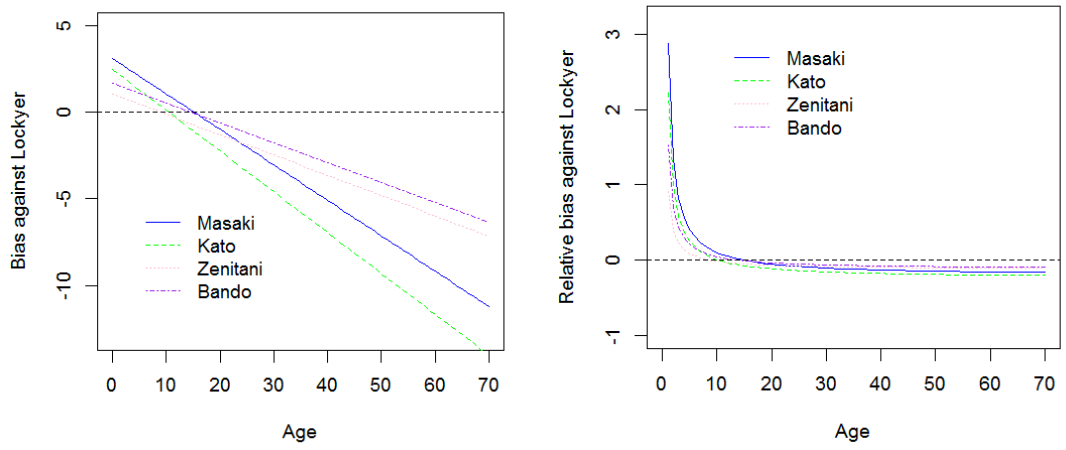

Figure 6. Absolute (left) and relative (right) biases for the Japanese readers relative to the control reader (Lockyer), who is assumed to be unbiased (Case 1).

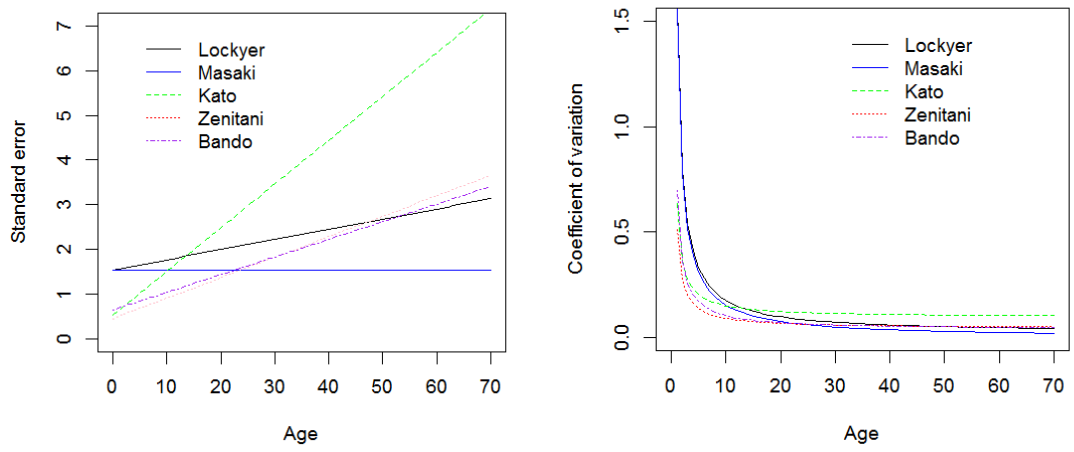

Figure 7. Standard errors (left) and coefficient of variations (right) for the control and Japanese readers if the control reader (Lockyer) is assumed to be unbiased (Case 1).

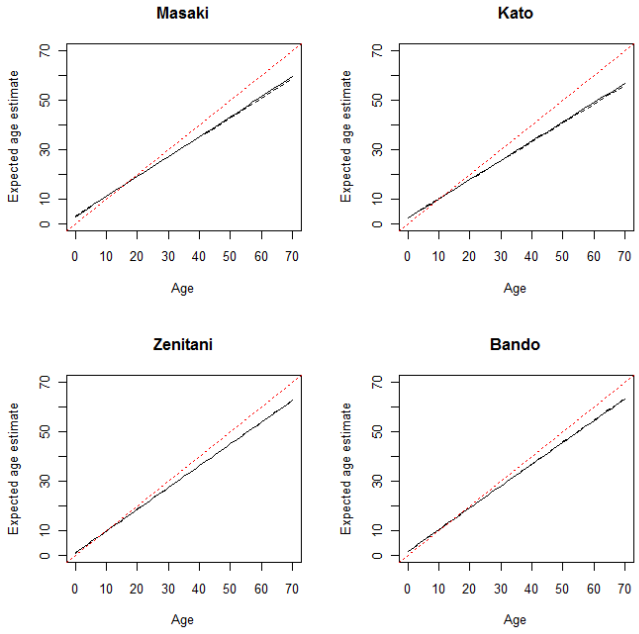

Figure 8. Sensitivity of the relationship between expected age and true age for the base-case analysis (solid black lines) and for the sensitivity test in which the catch age-composition is assumed to change over time (shaded black lines).

Appendix 1

Format used by Lockyer to summarize age readings

Specimen ID no	Age readings			Comments					
	Trial counts - given in sequence	Agreed count from trials based on weighted mean (CHL)	Best count according to Japanese method of average of counts	Plug complete? Yes or No; comment	Neonatal line present? Yes or No	Central cut? Yes or No	General appearance	Readability - Excellent; Good: Poor; Unreadable	Other

Specimen ID no
This refers to the experimental number provided for this reading stage of the experiment.

Age readings
In general following descriptors were used. When there is uncertainty about age, the age is prefixed by ca – e.g. ca N. When part of the plug is missing, + is suffixed on the age. However, + can also be applied in young animals (range up to 6 GLGs) where a new GLG is forming at the edge but maybe incomplete. Other ways of giving this are e.g. N - N+1 – in other words a range. Sometimes two possible ages are offered because of difficulties in reading. Here the ages will be e.g. N or P. Where only a minimum age is counted in difficult to read plugs, the age will be given as e.g. >N. Sometimes this notation is also used for incomplete plugs.

Trial counts - given in sequence
This gives the numbers of GLGs counted in sequence. The minimum number of trials is three, but may be many more depending on the confidence of the reader in what is being seen. It should be noted that before recording counts, the ear plug has been scanned several times to get a feel for the GLG patterns with rough counts made. The written counts reflect when the reader is more confident in the counting.

Agreed count from trials based on weighted mean (CHL)
In cases where there is no consistency of count, the mean may be weighted to the most recent count depending on the relative confidence in the reading.

Best count according to Japanese method of average of counts
The mean here is a simple mean and treats all readings equally.

Plug complete? Yes or No; comment
Yes denotes that all parts of the core were found, even if in two or more pieces. A comment will usually describe how many pieces or what is missing.

Neonatal line present? Yes or No
Yes means that at least part of the neonatal Line has been identified.

Central cut? Yes or No
Yes means that the core is adequately exposed at the centre line.

General appearance
Information on colouration, relative size, etc. is given here. However, this has not been consistently provided, but has often been added if there has been a problem with reading. If the plug or part of it is attached to the glove finger, this is noted.

Readability - Excellent; Good: Poor; Unreadable
E – Excellent means very clear GLGs and little error likely in reading.
G – Good means generally quite readable with mostly clear GLGs. However, there may be some error.
P – Poor means parts of the plug are difficult to read because GLGs are obscure or irregular. A large margin of error is likely in GLGs.
U – Unreadable means that the clarity of GLGs is so poor and/or confusing, that any GLG count provided is likely to be erroneous or incomplete.
Combinations of categories e.g. G/P mean partly good and partly poor – often which part will be specified e.g. P(top)/G (base).

Other
Here expanded information on readability may be given; also possible transition phase age if determined.