Age estimation of Southern Hemisphere minke whales: Issues and error models based on data from the 1983 IWC Aging Workshop

TOM POLACHECK, CSIRO Marine and Atmospheric Research, GPO Box 1538, Hobart, TAS 7001, Australia
ANDRÉ E. PUNT, School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195-5020, USA

Contact e-mail: Tom.polacheck@csiro.au

ABSTRACT

Age estimates for 360 Southern Hemisphere minke whale earplugs by nine independent readers done in conjunction with the 1983 IWC minke ageing workshop (IWC 1984) are analyzed to provide insights into issues related to the accuracy and precision of the age estimates from the commercial and JARPA catches. Some preliminary models of biases and variances in age estimates are developed based on the comparative age readings as a way to explore the robustness of the population modelling results to errors in the ageing data. Cross comparison of the age estimates by the different readers indicates that systematic inconsistency (i.e. ageing bias) exists for at least some of the readers, and that the amount of bias is related to the estimated (and hence true) age. However, comparisons of this type do not allow the biased readers to be identified, but only to conclude that the readings by at least one, if not all, of the readers are not unbiased. The results also suggest that there is also likely to be a substantial amount of "random" (non-systematic) error in the age estimates of experienced readers. There is also substantial variability among readers in their assessment of the readability of an earplug. Some readers considered that a substantial proportion of the 360 earplugs were unreadable or attached qualifications to their estimates (i.e. as high as 42%), If those readers' estimates are unbiased then increased bias may be associated with the estimates by other readers for these "unreadable" earplugs based on limited data and analyses. Evaluation of the readability of earplugs is a problem requiring further investigation. There is a need for further work to develop appropriate ageing error models for the existing and future age reading readings underpinned by adequate data.

INTRODUCTION

Population modelling of Antarctic minke whales has been and continues to be an important area of research within the IWC Scientific Committee. The initial interest in this work stemmed from the relative lack of older individuals in samples taken when commercial catches began in the 1970s. Virtual population analyses (VPA) of the catch-at-age data from the commercial and subsequent JARPA catches suggested large increases in minke whale recruitment and hence abundance beginning in about 1940. These increases have been attributed to the removal of large numbers of other whale species occurring at this time (Butterworth *et al.* 1996). Subsequent work has also suggested that a continuing time series of age samples, combined with abundance estimates would be able to provide improved estimates of natural mortality when analysed using a population modelling framework (e.g. Butterworth *et al.* 1999). The robustness of the results from the population modelling analyses to alternative assumptions and model formulations (e.g. the form of the selectivity function and natural mortality assumptions) has not been fully resolved (e.g. Punt and Polacheck

2007, 2008). In addition, population modelling was identified as one important component in attempting to understand the large decrease in abundance estimates for minke whales from the IWC IDCR/SOWER sighting surveys based on the "standard" estimation methodology.

Critical to all of the population modelling work that has been done to date are the estimates of the age distribution of the catches from the commercial and JARPA catches. The age estimates are derived from counts of the number of bands in earplugs taken from caught whales. However, little information is available on the reliability (i.e. the accuracy and associated precision) of the age estimates or about whether the bands are always formed annually (as is the underlying assumption). More recently, comparisons of length-at-age data from the commercial and JARPA catches suggest an apparent inconsistency (Punt and Polacheck 2005, Polacheck and Punt 2006). Aging and/or length measurement errors were considered in Polacheck and Punt (2006) as possible hypotheses contributing to this apparent inconsistency, who noted that there were unresolved questions with respect to the age and length data.

The most extensive review of the ageing methodology and age estimation for Antarctic minke whales was undertaken in 1983 in conjunction with an ageing workshop (IWC, 1984). As part of the workshop, independent earplug readings were undertaken by nine scientists. However, the primary focus of the workshop was on whether age at maturity could be determined reliably from earplugs. The extent to which the comparative age readings were analysed in terms of age reading errors was limited (in part due to time constraints). The workshop report left a number of unresolved questions about the accuracy and precision of the ageing data (IWC 1984, Polacheck and Punt 2006).

The key importance of catch-at-age data as input to the population modelling for Antarctic minke whales was noted at the 2006 annual meeting of the IWC Scientific Committee. Among the identified high priority tasks for future work identified at that meeting was to "examine the data from the 1983 ageing workshop to provide insights for the development of error models for the catch-at-age data - particularly with respect to potential biases arising from unreadability of ear plugs being related to age" (IWC, 2007). Problems in obtaining permission to access the workshop data prevented any progress with respect to this task for the 2007 annual meeting. These access problems were subsequently resolved and the data have been made available. The purpose of the present paper is to present analyses of the comparative age readings from the 1983 workshop with respect to their reliability. We also develop some preliminary models of biases and variances in the age estimates based on the comparative age readings as a basis to explore the robustness of the population modelling results to errors in the ageing data. We stress that the models presented are at best preliminary. As discussed further below, there is a need for additional research, data and analyses to be able to evaluate the reliability of the existing age readings and to reasonably quantify the errors associated with them.

Data and Methods

The Comparative Age Reading Data from the 1983 Workshop

The 1983 workshop circulated 360 minke whale ear plugs to nine scientists to obtain comparative age estimates. Six of the readers examined all 360 earplugs, while the

number of earplugs for which age related data[1] are provided by the other three readers ranged from 173 to 319 (Table 1). The readers had varying amount of previous experience in reading whale earplugs in general and those from minke whales in particular. Four of the readers (i.e. readers 5, 6, 8 and 9) had essentially no or little previous experience (Lockyer, personal communication). Readers 1, 2 and 4 had previously trained and worked closely together. Reader 1 was the individual who had undertaken the original earplug readings for the early commercial catches. His estimate were supplied to the IWC and used in analyses undertaken for the Scientific Committee. His readings in the workshop data set are his original readings (i.e. he did not re-read the earplugs for the workshop).

The IWC archive contains a number of files with age reading data – representing in part progressive updating of the data. The readings by readers 5, 6, 8 and 9 (the inexperienced readers) were in fact not coded by the beginning of the workshop and were not considered by it. We have used the data contained in the file "IWC.SC35MAW.READINGS" in the analyses presented here. This file is described as the "definitive dataset of age and transition phase readings arising from Minke Aging Workshop". The file was finalized after the workshop in September 1983. As such, some of the numerical summaries presented here may differ from those contained in the Workshop Report.

The procedure used to select the 360 earplugs used in the workshop is described in Kato (1984). The sampling procedure involved a two-way stratified random sample (i.e. stratified by two readability categories and nine ovulation categories with 20 samples in each stratum). The two readability categories were either poor or good and were determined by Kato at the time of sampling based on his impression of the earplug as seen through the plastic film in which it was contained at the time the sample was drawn. An earplug was excluded from the sampling if it was considered to be incomplete.

Some readers undertook up to three readings of the each earplug for a portion of the available samples (Table 1). Readers 7, 8 and 9 made the multiple readings independently and blind with respect to the other readings. As such, these multiple readings can be used to provide an intra-reader comparison. However, readers 2, 3 and 4 undertook the three reading in sequence. These readings were hence not undertaken "blind" and the 1983 workshop considered that they could not be used for intra-reader comparisons.

The IWC data file also provides a "best" age estimate for each earplug. In cases with only a single reading, this reading was considered "best". When multiple readings were undertaken, the procedure for determining the "best" estimate varied among readers (IWC, 1984). In the case where the multiple readings were done blind, either the mean or mode of the three age estimates was chosen as the "best". In the case of sequential readings, the reader generally selected the reading in which he had the most confidence, except in cases where the estimates varied widely when a mean or no "best" estimate may have been provided.

[1] "age related data" means that either that an age estimate was provided or a coded comment was made related to the readability of the earplug in either as a specific comment or general remark fields (Appendix 3).

The readers also supplied information on their assessment of the readability of the earplugs. The codes for qualifying the readings varied among individuals which confounds inter-reader comparisons. For readers undertaking multiple blind readings separate and independent assessments were made for each reading. In addition to the specific code supplied with respect to readability, readers also supplied general remarks about the earplugs. These too varied among readers as well as the extent to which any comments related to readability were included.

Error Estimation Model

Several alternative quantitative hypotheses for the extent of age reading error (i.e. bias and variance) that might exist in the commercial and JARPA age reading data sets were generated by assuming that either (a) the age estimates of the one of the age readers were the "true" ages or that (b) both readers made imprecise estimates but one of the readers was unbiased. For the latter approach, the true ages in the sample were treated as random effects.

Descriptions of these two methods are provided in the appendices. The first method allowed a wide range of alternative models to be efficiently and easily explored, particularly in light of the number of cross combinations among readers. The second method is statistically more robust because it only requires assuming that one reader's estimates of age are unbiased and does not require making the assumption that the unbiased age estimates are the true ages. It is not our purpose to attempt to resolve which set of estimates are the most appropriate. Instead our purpose is to generate a range of plausible error models that are consistent with the age reading data from the 1983 Workshop. These models are intended to provide an indication of the potential magnitude of the error that may exist in the minke whale ageing data. They can be used to provide an initial assessment of the possible consequences of ageing error (particularly bias) on the robustness of conclusions from the catch-at-age modelling work (see Punt and Polacheck (2008) for examples of this).

Results

Cross Comparison Among Readers

Table 2 summarizes the percentage agreement in age readings for the nine readers. As is evident from this table, exact agreement was rare (in general <20% and never >39%). It should be noted that such levels of lack of agreement do not necessarily indicate poor precision. In fact, even for relatively modest levels of reading error (e.g. a CV of 10%), less then 50% of a reader's estimates will correspond to the true age for ages 8 and greater (Figure 1). Moreover, the percentage of times that independent estimates from two readers with the same level of precision would be the same would in general be less then the expected number of times a reading would correspond to the true age (Figure 1).

Figure 2 compares the "best" age estimates for reader 1 with those for the other eight readers. Reader 1 was chosen as the standard for this comparison because his readings are the ones used in the population modelling work. Evident in this figure is that there is substantial variability among readers in their estimates. In particular, this figure and Table 3 suggest that substantial bias may exist in the age readings of some readers (e.g. readers 7 and 8) relative to those of reader 1 and that the bias tends to increase with age (or *vice-versa*). Of course, comparisons of this nature do not indicate that the readings of any particular reader are biased, only that at least one of the readers is biased. The readings of readers 1, 2 and 4 are highly consistent, although this is not

unexpected because readers 1, 2 and 4 have worked together (e.g. are from the same "school" of readers"). This indicates that a relatively high level of consistency in age estimates among readers is likely to be achievable. Note, however, that high levels of consistency do not in themselves constitute high levels of accuracy nor of precision.

Figure 3 shows the differences between the "best" age readings of reader 1 and those of the other eight readers as a function of corpora count. As the corpora count is correlated with age, it is not surprising that the amount of variability (and apparent bias when it appears to exist) appears to increase with the corpora count. What is not clear is the extent to which bias and imprecision depends not only on true age, but also on reproductive history. In particular, the level of reproductive activity might affect the subsequent band formation and hence the readability and/or the "true" relationship between the number of bands and age, since maturity is considered to affect banding patterning in earplugs (i.e. the "transition phase"). This question is considered further below.

Readability

The results above are based on each reader's "best" estimate. However, for some readers, qualifications were associated with their age estimates as a result of concerns regarding the readability of the earplug. Cross comparison among readers with respect to their judgement of readability is confound by an absence of an agreed protocol and a lack of consistency among readers in how and the extent to which such information was recorded. Some readers provided substantive qualification about the readability of an earplug even when they provided a "best" estimate. It should be noted that only readings from 154 of the 360 were considered by the 1983 Workhsop to have been "readable" by all five experienced readers (1WC, 1984).

The criteria used by the Workshop to classify an earplug as "readable" appears not to have documented and we were not easily able to establish a set of criteria that yielded approximately 154 "readable" earplugs. It is clear that the Workshop considered that simply obtaining an age estimate did not automatically imply that an earplug was "readable" Had that been the criterion used by Workshop, 320 of the first readings of earplugs and 282 of the "best" estimates would have been assessed as "readable" by all five experienced readers. (Note that the Workshop used the first readings in their analyses). If earplugs for which specific qualifications were given to an estimate (e.g. Table A3.1 in Appendix 3) were also considered to be "unreadable," then the number of earplugs assessed to be "readable" by all four readers would be 188 for the first readings and 179 for the "best" estimates. The number of "readable" earplugs is reduced to 178 and 170 respectively if earplugs for which there were also general remarks indicating a problem with the earplug (not simply that it was difficult) are also excluded, while the numbers are reduced to 137 and 133 if all earplugs for which there were general remarks indicating that the earplug was difficult to read are assumed to be "unreadable". The reasons for our inability to replicate the Workshop's number of "readable" earplugs may be due to our use of an updated data file and/or the Workshop's use of some other combination of the specific qualification and general remark codes as their criterion. Nevertheless, it would appear that the Workshop considered age estimates for which there were specific qualifications along the lines of those in Table A3.1 as "unreadable". Moreover, it is clear that the set of earplugs available for cross-comparisons will be sensitive to the criteria used for selecting "readable" earplugs.

Table 4 suggests that the assessment of readability and self judgement of the extent of likely error in an age reading is difficult for readers at the time when the readings are being made. For example, less then half of the readings received the same qualification (including no qualification) by the three readers who undertook multiple independent readings. For the inexperienced readers (readers 8 and 9), only 20% of earplugs were given the same qualification in all three readings, while for the experienced reader it was 39% (i.e. the combination of earplugs in which either no qualifications were given for all three earplug readings or the same qualification was given in all three readings in Table 4). Moreover, whether any qualification was attached to a reading varied greatly among readings. For example, 60% of the earplugs read by reader 7 had qualifications for only one of the readings.

The most frequent type of comment or qualification was that the reading was "approximate" and/or "difficult" (Tables 5 and 6). However, in those cases where the comments were more specific, almost all of them indicated that the reader considered that the estimate of age was likely to be an underestimate (Tables 5 and 6).

It seems important to know whether there is likely to be differing levels of bias and/or variability associated with readings that one reader excludes as "unreadable" and another assesses as "readable". If this were the case, estimates of the extent of error from inter-reader comparisons may provide incorrect estimates of the actual bias and imprecision. To obtain an indication of whether this may be a problem, the difference between the "best" estimates for reader 1 and first age-estimates by reader 7 for earplugs were tabulated, categorized into those earplugs subsequently judged to be "unreadable" (i.e. either no "best" estimate was provided or a qualification was attached to the estimate) and those for which no qualification was attached to the "best" estimate The differences in age estimates for the two readers were generally greater for the "unreadable" earplugs (Table 7). This suggests that there may be higher levels of bias associated with "unreadable" earplugs, at least if reader 7's estimates are unbiased.

Consistency of Intra-Reader Age Estimates

The multiple readings provide a way to assess the level of precision (variability) that may exist in the readings of individual readers (Figures 4-9). The readings for readers 7, 8 and 9 were done independently (blind) whereas those for readers 2, 3 and 4 were done sequentially. As would be expected, there is a higher level of consistency among the multiple readings when they are done sequentially. This tends to confirm that non-blind readings provide only minimum estimates of ageing imprecision. Some quantitative estimates of variability based on the multiple readings using the statistical ageing error model are provided below.

There is to be a tendency for the first readings by readers 7-9 (particularly those for older animals) to somewhat "biased" relative to those for the subsequent two readings (Figures 7-9). This was most evident in the estimates for reader 7 for whom earplugs that were assessed to be over ~20 yr on the first reading were generally assigned a higher age on the second and third readings (Figure 7). In contrast, earplugs that were assessed to be over ~30yr on the first reading by reader 9 tended to be assigned a lower age on subsequent readings (Figure 9). It is not clear whether this represents some form of "learning" between readings. Nevertheless, these results suggest that extrapolation of estimates of age reading error from limited multiple reading

experiments may not be straightforward as there may be "experimental" effects confounding the results.

Error Model Estimation

The analyses in this section are based on a reader's "best" age estimate and all earplugs for which a "best" age estimate was provided were included, even if specific and/or general remark comments indicated that there were problems with the estimate.

Comparison of age estimates from the various readers relative to reader 1 based on the assumption that the readings of one reader are without error (Appendix 2) resulted in the best fitting model including a bias term for 6 of the 8 readers. This included three out of four of the experienced readers (experienced readers were 2, 3, 4, and 7; Table 8). In all cases where the best fitting model included a bias term, the magnitude of the bias was estimated to increase with age but the magnitude of the estimated bias varied among readers (Figure 10). Somewhat surprising was the result that a bias component exists between readers 2 and 4 and reader 1, although they are all from the same "school". The estimated biases for these two readers were the lowest and relatively low, except for the older ages. The addition of a bias term in these models for the number of corpora count resulted in a significantly better fit in four out of the six cases where the best fit model included a bias term (Table 9). The estimated biases in these instances were a decreasing function of the number of corpora counts. However, the estimated magnitude of the bias component related to the corpora count was always small (e.g. less then one, Figure 11).

The method in which the true ages are treated as random effects (Appendix 1) was applied to the age estimates for readers 1 and 7 under the assumption that reader 7's estimates were unbiased. These two readers were selected based on the above results to provide an idea of the magnitude of the error that may exist in the minke whale ageing data based on readings from experienced readers. The best fit model in this case also indicates that the bias for reader 1 increases with age as does the estimated standard deviation of the age readings (Table 10, Figure 12). It should be noted that the parameter estimates from this fit are used in Punt and Polacheck (2008) to explore the effect of bias as well as random ageing error on the population model estimates of historical population trends and natural mortality rates. These results are referred to as the "Model #1" ageing error estimates in Punt and Polacheck (2008).

We also examined the relationship between age and ageing error standard deviation for readers 2, 3, 4, and 7 using the method in Appendix 1 and the multiple reads of each earplug (Figure 13). The estimates for readers 2, 3 and 4 are minimal estimates as they are based on sequential, non-independent readings. As such, it is not surprising that estimated errors for reader 7 (based on independent readings) are substantially larger than those for other readers. In this context, it is also worth noting that the estimated standard deviation for reader 1 (for whom no multiple readings are available) based on the comparison of readers 1 and 7 are similar to those for reader 7 (e.g. Figures 12 and 13). Also, the magnitude of the estimated random error component for reader 1 is estimated to be somewhat greater using the method in Appendix 1 if the model is fitted assuming that reader 1's estimates are unbiased and reader 7's are biased (Figure 14).

Discussion

The results presented in this paper highlight problems in the interpretation of the age reading data for Southern Hemisphere minke whales based on earplug readings and problems in developing appropriate error models for use in population modelling. The results based on age reading by different readers suggest that systematic inconsistency (i.e. bias) can be expected in their estimates. However, comparisons of this type do not allow the biased readers to be identified. The results also suggest that there is also likely to be a substantial amount of "random" (non-systematic) error in the age estimates of experienced readers. The magnitude of random error and possible biases estimated from the cross comparison of different readers as well as the more limited independent age readings by the same reader appear to be of sufficient magnitude to potentially effect the results from catch-at-age population modelling (both the expected values of parameter estimates and their variances). The currently available data do not allow the determination of the actual relationship between the age estimates from any individual reader and the true age of an animal as data for "ground truthing" the relationship are lacking.

Development of appropriate aging error models underpinned by adequate data is critical for current and future catch-at-age population modelling of Southern Hemisphere minke whales The analyses in this paper highlight a number of issues that need to be considered in the collection of data for and the development of quantitative ageing error models. These include:

- development of approaches for verification of the relationship between the estimated age from earplugs and the true age of an animal;
- appropriate protocols for regular multiple and truly independent reading of the same earplugs by individual readers;
- development of appropriate standard classifications of earplugs with respect to readability;
- investigation of the relationship between the readability of an earplug and age; and
- development of robust methods for estimating aging error if substantial portions of the earplugs have associated problems with readability.

ACKNOWLEDGEMENTS

Partial funding for this work was provided through grants from the International Whaling Commission. We wish to thank Greg Donovan for his efforts in facilitating access to these data and to the age readers who participated in the 1983 IWC workshop.

Literature Cited

Butterworth, D.S. and A.E. Punt. 1999. An initial examination of possible inferences concerning MSYR for Southern Hemisphere minke whales from recruitment trends estimated in catch-at-age analyses. *J. Cetacean Res. Manage*. 1(1): 33-9.

Butterworth, D.S., Punt, A.E., Geromont, H.F., Kato, H. and T. Miyashita. 1996. An ADAPT approach to the analysis of catch-at-age information for Southern Hemisphere minke whales. *Rep. int. Whal. Commn* 46:349–59.

Kato, H. 1984. Sampling Procedure for the earplugs provided to the minke whale aging workshop. *Rep. Int. Whal. Comm*. 34:717–718.

IWC. 1984. Report of the Minke Whale Aging Workshop. *Rep. Int. Whal. Comm*. 34:675–99.

IWC. 2007. Report of the Scientific Committee, Annex G. Report of the Working Group on Population Modelling. Appendix 4 to the Report of the Sub-committee on In-depth Assessment (IA). *J. Cetacean. Res. Manage.* 9(Suppl.): 177–8.

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

Punt, A.E. and T. Polacheck. 2005. Application of statistical catch-at-age analysis for Southern Hemisphere minke whales in Antarctic Areas IV and V. Document SC/57/IA9 presented to the IWC Scientific Committee, June 2005. (unpublished). 71pp.

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

Punt, A.E. and T. Polacheck. 2008. Further Analyses Related to Application of Statistical Catch-at-age Analysis to Southern Hemisphere Minke Whales. SC/60/IA2. to be presented to the IWC Scientific Committee, June 2008. (unpublished). 46pp.

Table 1: The number of earplugs read by each reader, the number of earplugs for which a "best" age estimate was supplied, and the number of earplugs for which either specific comments and/or general remarks were supplied.

	Number Read*				Number with comments				Number
Reader	1st reading	2nd reading	3rd reading	Number aged**	1st reading	2nd reading	3rd reading	"best" reading	w/ general remarks
1	360	0	0	360	0	0	0	0	360+
2	360	359	357	356	10	7	7	6	16
3	360	320	320	290	29	27	27	0	146
4	360	359	359	359	6	6	6	6	2
5	299	0	0	298	7	0	0	7	1
6	319	0	0	316	20	0	0	20	49
7	360	51	51	352	139	22	20	150	360
8	360	51	51	357	246	43	46	231	328
9	173	51	51	165	23	11	10	14	173

* "number read" for the first reading is the number of earplugs for which an age estimate was provided plus the number for which no age estimate was provided, but a specific comment was supplied for a reading or general remark was supplied. For the second and third reads, the "number read" is earplugs only for which an age estimate was provided plus the number for which no age estimate was provided, but a specific comment was supplied for a reading

** "number aged" is the number of earplugs for which a best estimate of age was provided.

\+ the general remarks associated with reader 1 were the assignment made by Kato (1984) as to whether an earplug had good or bad readability based on a visual inspection of the earplug at time of selection

Table 2: Percent agreement in the "best" age estimate for the nine readers. The lower triangle includes all earplugs while the upper triangle includes only those reading for which a "best" reading was provided by both readers.

	Reader								
Reader	1	2	3	4	5	6	7	8	9
1	-	37	17	23	9	11	15	11	12
2	37	-	17	39	15	14	16	14	14
3	13	14	-	17	9	10	13	10	8
4	23	39	14	-	9	13	16	13	13
5	8	13	10	8	-	11	10	8	12
6	9	12	11	12	20	-	14	9	9
7	14	15	11	16	8	13	-	17	8
8	11	13	9	13	7	8	16	-	11
9	5	7	14	6	5	5	4	6	-

Table 3: Percent of the "best" age readings for each age reader that were either less than, equal to, or greater than the age reading by reader 1. Only readings for which a "best" reading was provided by both readers are included.

Reader	Below	Equal	Above	N
2	29	37	34	356
3	48	17	36	290
4	33	24	43	341
5	42	9	49	298
6	48	11	41	316
7	22	15	63	352
8	21	11	68	357
9	36	12	53	165

Table 4: The consistency in the qualification (i.e. specific comments) associated with the multiple independent readings for readers 7, 8 and 9 (the only readers for which independent multiple readings are available).

	No Qualification in at Least One Reading			Qualifications in All Readings		
Reader	All 3 readings	2 out 3 readings	1 out 3 readings	Number	Same in all 3 readings	Total
7	17	14	12	8	3	51
8	2	4	12	33	9	51
9	0	5	4	2	2	11

Table 5: The frequency of different comment codes associated with the "best" estimates for each reader (see Appendix 3 for the definition of each code). Note that the reader 1 estimates were made prior to the workshop and so no comments codes are available for this reader.

		Reader							
	Comment	2	3	4	5	6	7	8	9
"Best" age estimate provided	0	350	290	353	292	299	210	129	159
	1	1	0	0	6	1	0	0	0
	2	3	0	3	0	0	0	0	0
	4	2	0	3	0	0	0	0	0
	6	0	0	0	0	11	16	0	1
	7	0	0	0	0	0	27	28	5
	8	0	0	0	0	0	69	112	0
	20	0	0	0	0	0	21	0	0
	21	0	0	0	0	5	8	51	0
	22	0	0	0	0	0	1	1	0
	24	0	0	0	0	0	0	2	0
	25	0	0	0	0	0	0	2	0
	26	0	0	0	0	0	0	27	0
	27	0	0	0	0	0	0	2	0
	29	0	0	0	0	0	0	3	0
Numb. no estimate*		4	70	1	62	44	8	3	7
% w/o qualification		97	81	98	81	83	58	36	92

* i.e. a specific comment was supplied but no "best" age estimate was included.

Table 6: The frequency of different general remark codes associated with individual earplugs for each reader (see Appendix 3 for the definition of each code). Note that the reader 1 estimates were made prior to the workshop and so no comments codes are available for this reader.

		Reader							
	Remark	2	3	4	5	6	7	8	9
Best age estimate provided	0	344	193	358	298	270	0	32	0
	31	0	1	0	0	1	0	0	0
	32	0	0	0	0	2	0	0	0
	33	2	0	0	0	0	0	0	0
	34	3	0	0	0	0	0	0	0
	37	0	5	0	0	0	0	0	0
	38	0	6	0	0	0	0	0	0
	39	1	1	0	0	0	0	0	0
	40	0	64	0	0	27	0	0	0
	41	1	1	0	0	0	0	0	0
	42	0	2	0	0	0	0	0	0
	43	0	3	0	0	0	0	0	0
	44	0	6	0	0	0	0	0	0
	45	0	5	0	0	0	0	0	0
	51	0	0	0	0	15	9	2	13
	52	0	0	0	0	0	58	82	41
	53	0	0	0	0	0	179	169	92
	54	0	0	0	0	0	98	63	16
	55	0	0	0	0	0	8	9	3
	61	0	0	0	0	1	0	0	0
	99	5	3	1	0	0	0	0	0
No age estimate		4	70	1	62	44	8	3	7

Table 7: The mean difference between the first age estimate by reader 7 and the age estimate by reader 1, stratified by age-class and by whether the earplug was judged to be readable or unreadable by reader 7 (i.e. a qualification with the estimate provided in the specific remark field for reader 7's "best" reading).

	Unreadable		Readable	
Age	Mean Difference	*N*	Mean Difference	*N*
1-5	-0.8	4	-0.4	11
6-10	-0.3	26	-0.6	41
11-15	0.2	27	0.1	34
16-20	2.4	22	1.5	33
21-25	2.7	21	2.4	33
26-30	7.3	19	3.3	24
31-35	6.2	13	4.9	18
36-40	9.6	8	7.6	10
>40	17.4	9	9.8	6

.

Table 8: AIC values for various fits of reader 1's "best" age estimates to those of the other readers using the approach in Appendix 2 (i.e. the estimates by each of the other readers are equal to the true ages). The grey shaded value indicates the best fit models for each reader. The mathematical form of the SD and bias components for each of the numerical codes is provided in Appendix 2) .

SD	Bias	Reader							
Model	Model	2	3	4	5	6	7	8	9
0	1	1823.5	1688.3	1959.2	1823.8	1961.6	2284.6	2384.1	1019.6
0	2	1665.8	2123.0	1795.2	1861.1	2111.2	2082.3	2231.0	988.4
0	3	1659.0	1689.7	1784.0	1813.6	1918.5	2074.0	2202.8	956.6
0	4	1661.0	1685.1	1786.0	1815.1	1920.5	2076.0	2204.8	958.6
0	5	1737.8	1683.7	1773.0	1815.6	1904.3	2048.0	2170.5	946.4
1	1	1818.1	1690.3	1943.0	1825.8	1963.5	2205.8	2293.7	1013.7
1	2	1667.6	2125.0	1797.1	1863.1	2113.2	2082.4	2231.8	990.4
1	3	1660.8	1691.7	1785.3	1815.6	1920.5	2068.3	2188.0	958.6
1	4	1822.2	1687.1	1947.0	1829.8	1965.2	2070.3	2264.1	966.6
1	5	1822.1	1694.3	1947.0	1829.8	1967.5	2209.8	2297.7	1017.7
2	1	1830.8	1720.3	1941.1	1836.3	1973.2	2118.9	2192.0	1006.6
2	2	1734.5	2179.4	1858.8	1878.6	2188.4	2101.0	2243.1	1019.5
2	3	1659.1	1693.6	1775.2	1817.6	1920.6	2000.6	2102.8	954.5
2	4	1663.4	1696.2	1782.6	1819.1	1955.4	2018.7	2104.8	962.4
2	5	1808.5	1696.2	1919.5	1831.8	1951.7	2095.2	2169.2	999.5
3	1	1822.1	1694.3	1925.4	1829.7	1950.8	2128.5	2229.3	1002.1
3	2	1668.5	2129.0	1798.7	1865.2	2117.2	2086.2	2207.6	994.4
3	3	1663.3	1695.7	1788.6	1819.4	1924.5	2028.9	2152.1	962.6
3	4	1665.3	1698.3	1790.7	1820.9	1926.5	2004.6	2106.8	958.5
3	5	1826.1	1698.2	1921.5	1833.8	1953.7	2097.2	2171.2	1001.5
4	1	1778.0	1686.2	1890.7	1829.8	1907.6	2032.8	2081.5	976.8
4	2	1657.3	2128.8	1777.4	1867.1	2112.3	1993.9	2145.2	983.3
4	3	1646.0	1688.1	1757.6	1819.6	1892.8	1940.5	2035.0	938.6
4	4	1651.9	1690.2	1759.6	1821.1	1909.4	1942.5	2037.0	940.6
4	5	1782.0	1690.2	1894.7	1833.8	1911.6	2036.8	2085.5	980.8

Table 9: AIC values for various fits of reader 1's "best" age estimates to those of the other readers using the approach in Appendix 2 for models in which an additional bias component related to the number of corpus counts was included in the model. The grey shaded value indicates the best fit model including those models in Table 8. The mathematical form of the SD, bias and corpus components for the various numerical codes is provided in Appendix 2).

SD	Age	Corpora	Reader					
Model	Bias	Bias	2	4	6	7	8	9
1	3	3	1794.7	1896.3	1842.2	2027.6	2086.6	970.2
2	5	5	1667.2	1782.9	1914.9	1963.4	2068.6	962.4
3	3	3	1656.0	1760.3	1789.9	2359.7	2001.2	926.6
4	5	5	1658.0	1762.3	1914.3	1920.6	2003.2	928.6
5	3	3	1798.7	1900.3	1846.2	2031.6	2090.6	974.2
1	5	5	1775.9	1893.7	1834.7	2014.6	2058.0	970.2
2	5	5	1669.2	1775.9	1915.5	1913.4	2065.4	-*
3	5	5	1648.8	1757.9	1860.4	1932.5	2034.1	935.4
4	5	5	1652.2	1757.5	1792.4	1890.2	1982.0	922.7
5	5	5	1780.1	1897.7	1838.7	2018.6	2062.0	974.2

* failed to converge

Table 10: Application of model selection criteria based on AIC to the first and second age readings for reader 7 and the "best" age estimates for reader 1 based on the estimation method described in Appendix 1, assuming Reader 7's estimates are unbiased. The grey shaded value indicates the best fit model.

Model #	Bias	Standard deviation	# pars	Likelihood	AIC
1	Von Bertalanffy	Von Bertalanffy	41	401.108	884.216
2	Linear	Von Bertalanffy	39	403.742	885.484
3	Exponential	Von Bertalanffy	41	414.565	911.130
4	Von Bertalanffy	Linear	37	407.147	888.294
5	Linear	Linear	35	408.966	887.932
6	Exponential	Linear	37	423.658	921.316

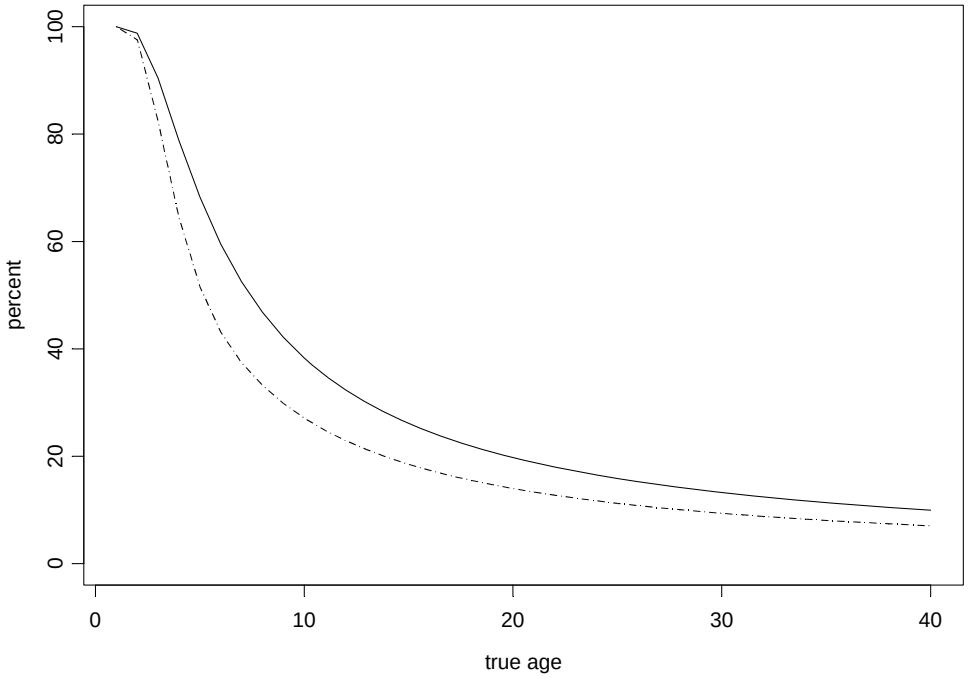

Figure 1: The expected percentage of times a reader's age estimate would be expected to be equal to the true age (solid line) and the percentage of times the estimates from two independent readers would be expected to be equal (dotted line), assuming that the estimates are unbiased and follow a Berkson's distribution (Appendix 2) with a coefficient of variation of 0.10.

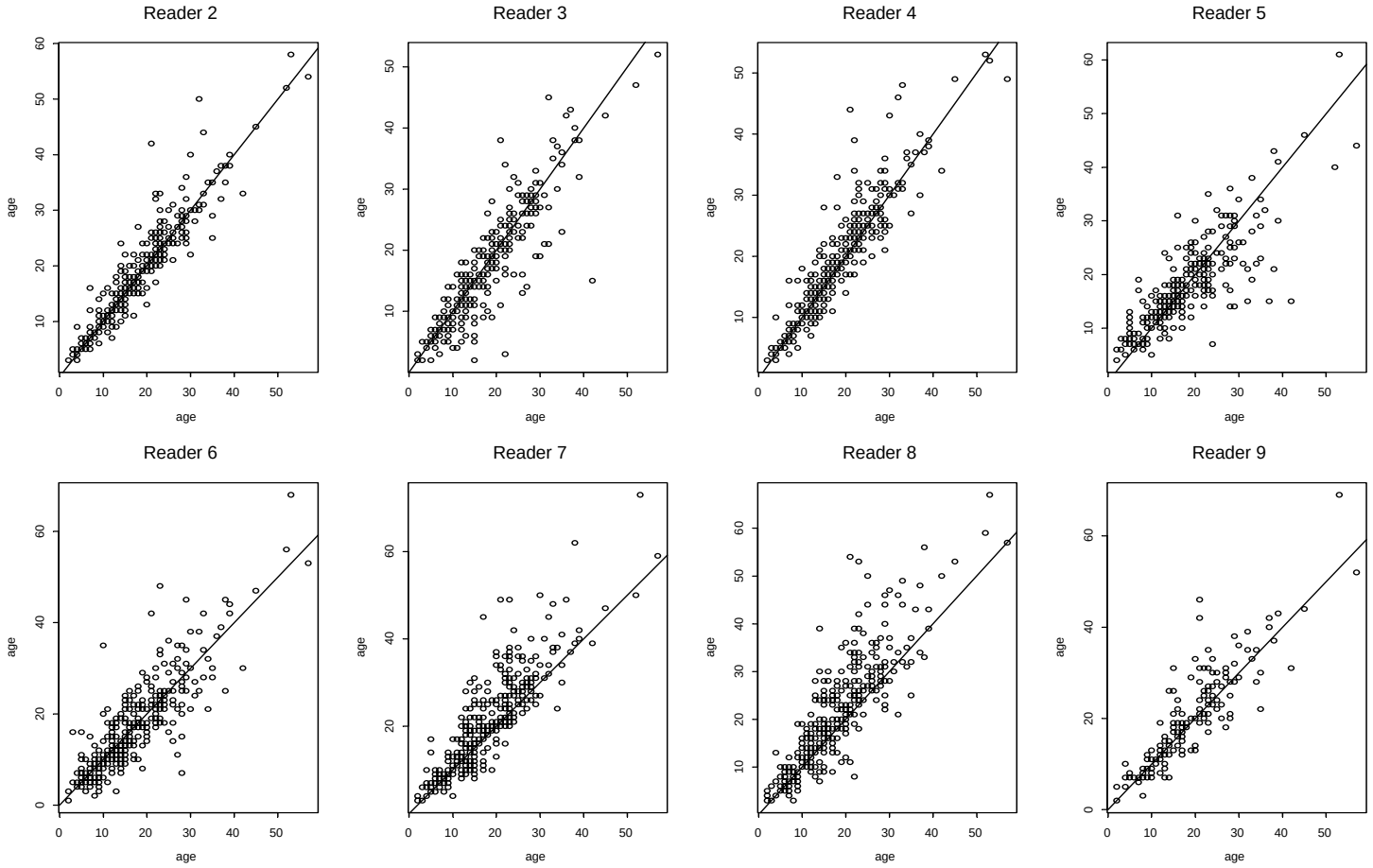

Figure 2: Comparison of reader 1's best estimates (x-axis) with the "best" estimates for the other eight readers (y-axis).

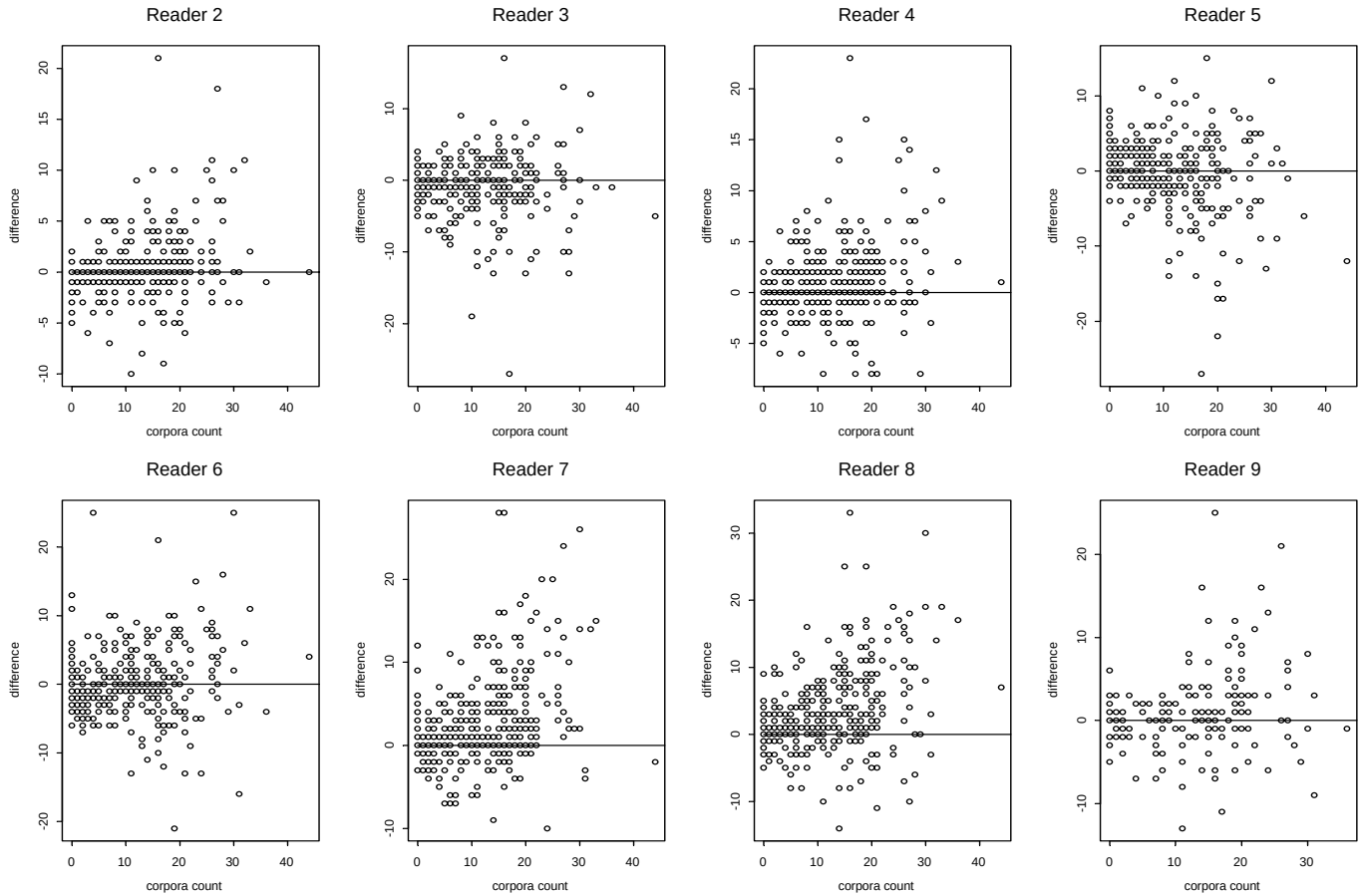

Figure 3: Difference between reader 1's best estimates and those for the other eight readers as a function of corpora count.

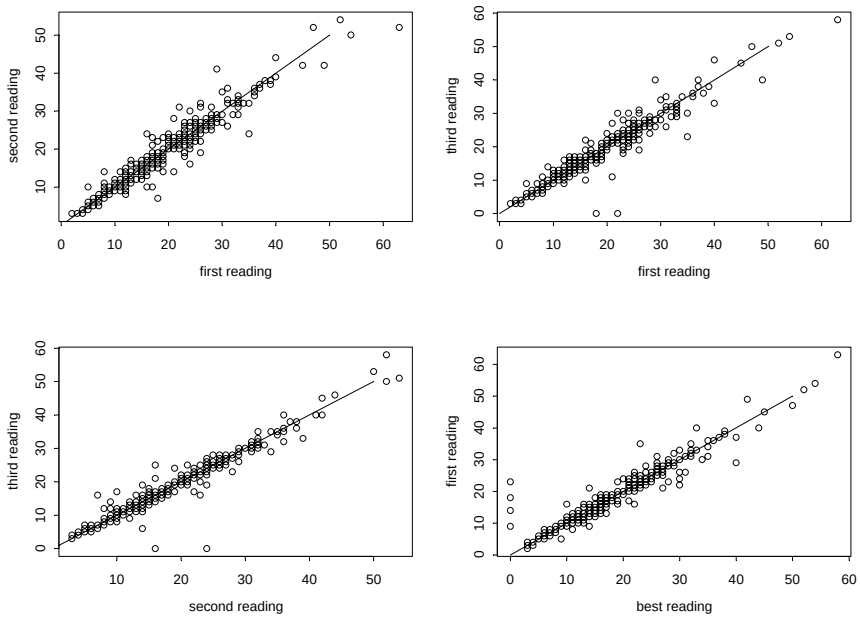

Figure 4: Comparison of the first, second and third age readings by reader 2. Each set of readings was done in turn (i.e. not blind). Also shown is the relationship between the first reading and the "best" estimate provided by reader 2.

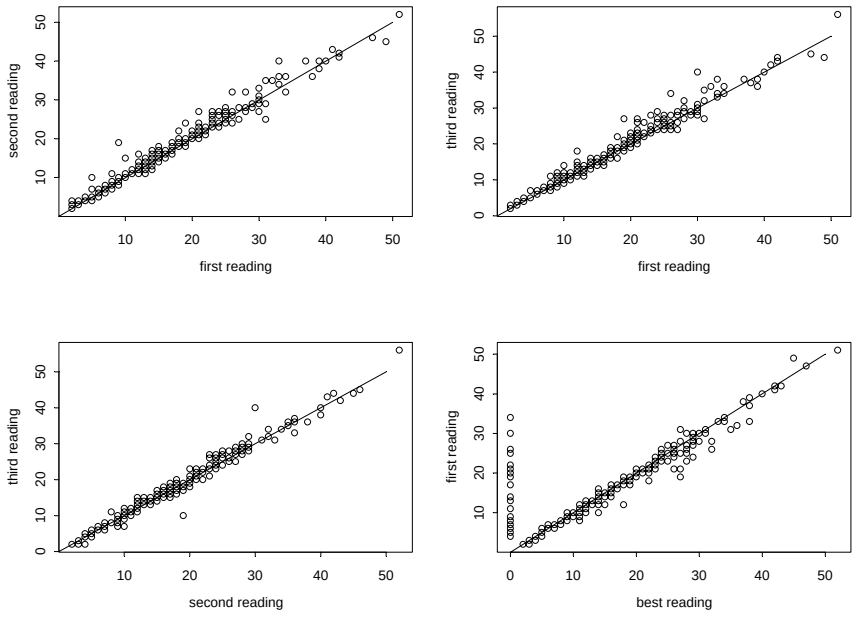

Figure 5: Comparison of the first, second and third age readings by reader 3. Each set of readings was done in turn (i.e. not blind). Also shown is the relationship between the first reading and the "best" estimate provided by reader 2.

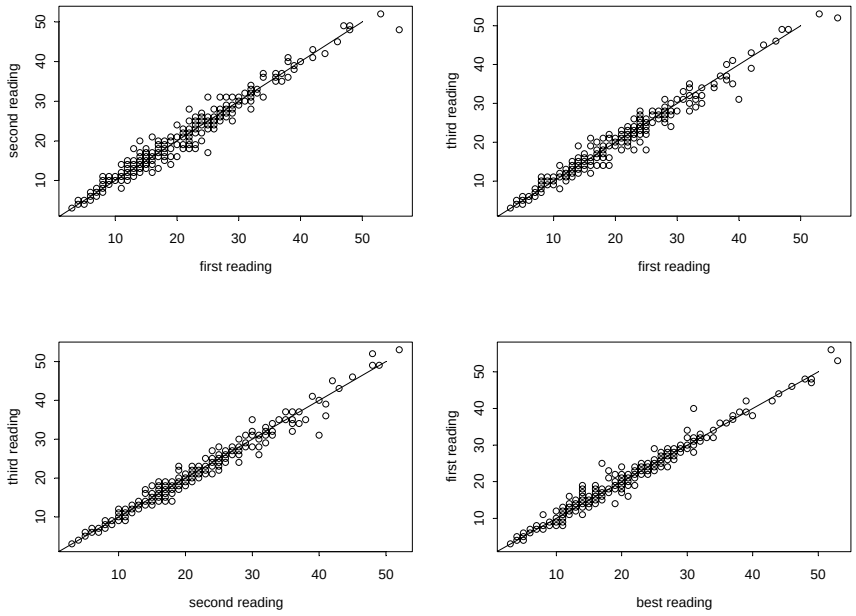

Figure 6: Comparison of the first, second and third age readings by reader 4. Each set of reading was done in turn (i.e. not blind). Also shown is the relationship between the first reading and the "best" estimate provided by reader 4.

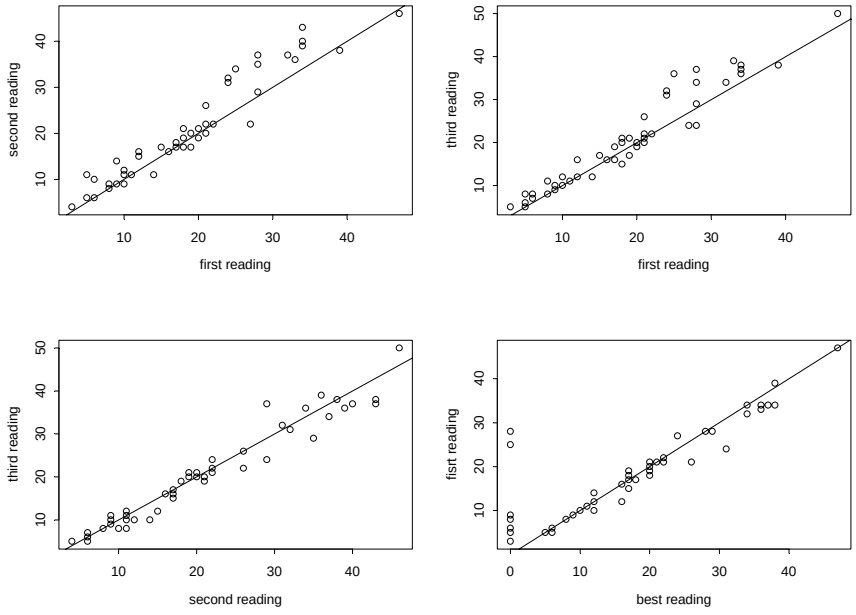

Figure 7: Comparison of the first, second and third age readings by reader 7. Each set of readings was done blind. Also shown is the relationship between the first reading and the "best" estimate provided by reader 7.

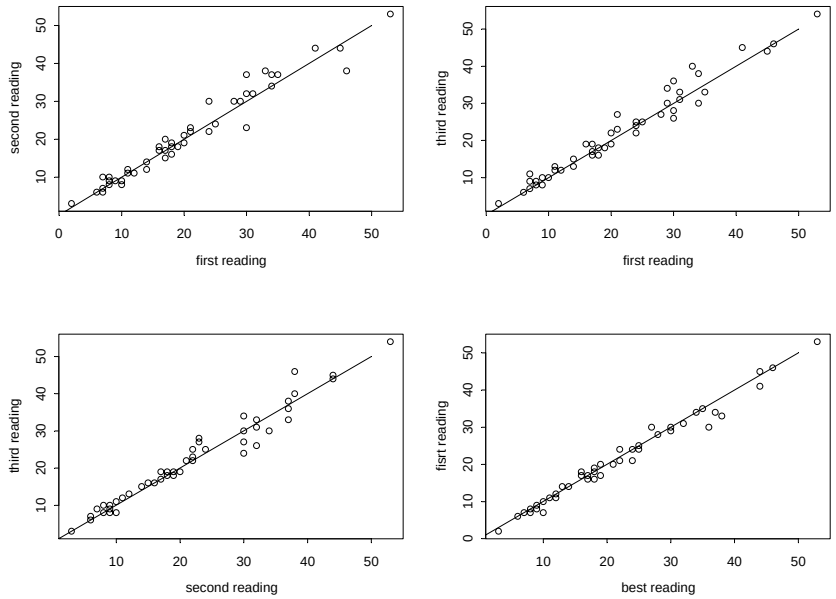

Figure 8: Comparison of the first, second and third age readings by reader 8. Each set of readings was done blind. Also shown is the relationship between the first reading and the "best" estimate provided by reader 8.

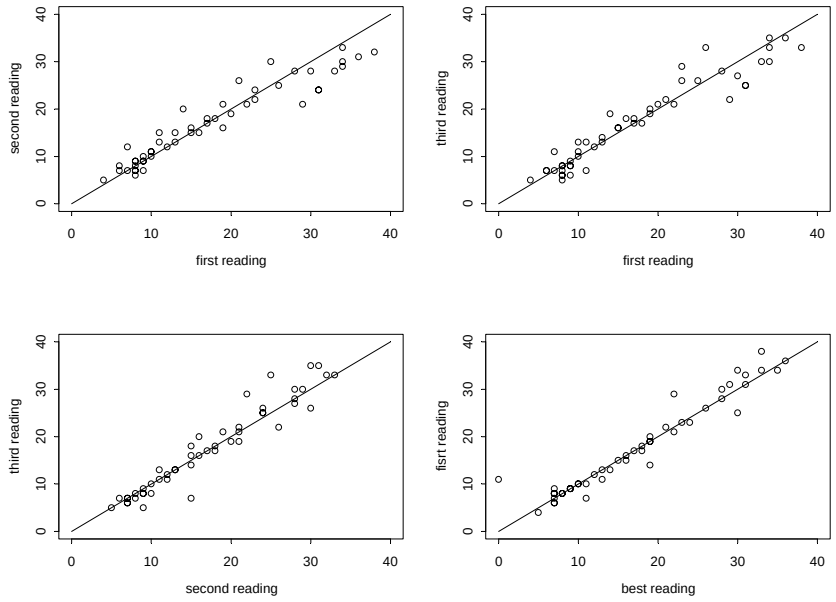

Figure 9: Comparison of the first, second and third age readings by reader 9. Each set of readings was done blind Also shown is the relationship between the first reading and the "best" estimate provided by reader 9.

.

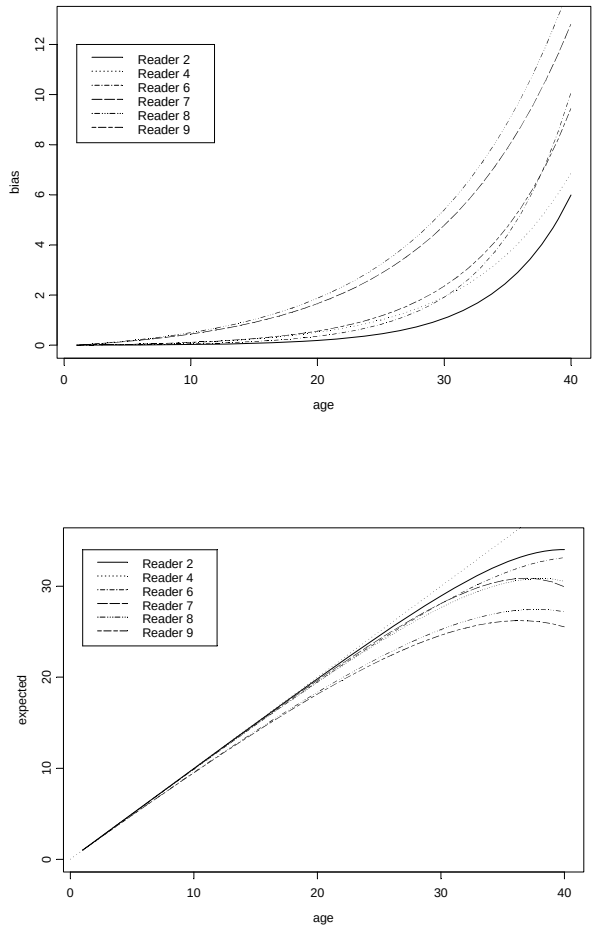

Figure 10: Estimated bias (upper panel) and expected age (lower panel) for reader 1 as a function of true age for the best fit model using the error estimation model in Appendix 2 when the readings of different age readers are assumed to be equal to the true age.

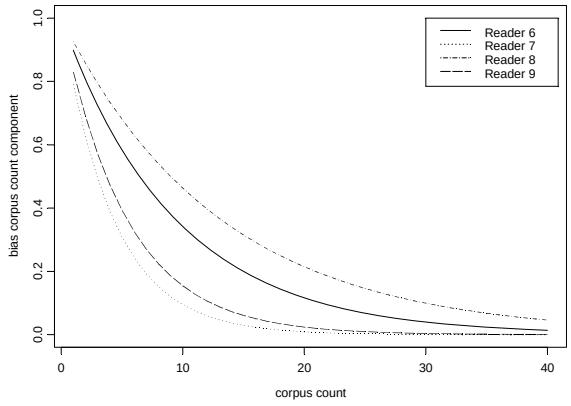

Figure 11: Estimated bias component for reader 1 as a function of corpus count for the best fit model using the error estimation model in Appendix 2 when the readings of different age readers are assumed to be equal to the true age.

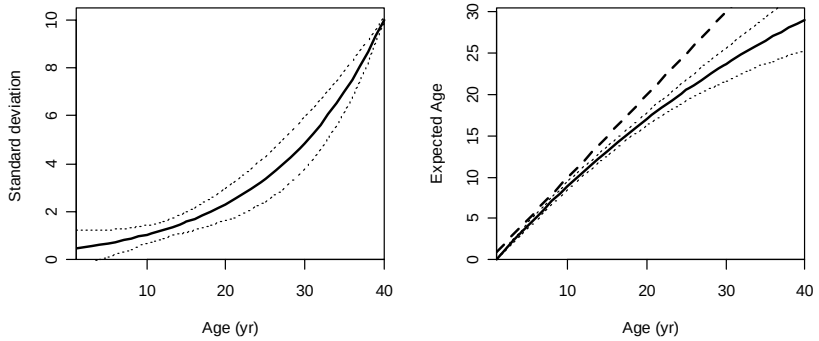

Figure 12: Standard deviation (left panel) and bias (right panel) of estimated age-reading error for reader 1 based on the approach in Appendix 1 and for the AIC-selected model in Table 10 (Model #1). The solid lines denote the maximum likelihood estimates and the dotted lines the asymptotic 90% confidence intervals. The dashed line in the right panel is the 1-1 line.

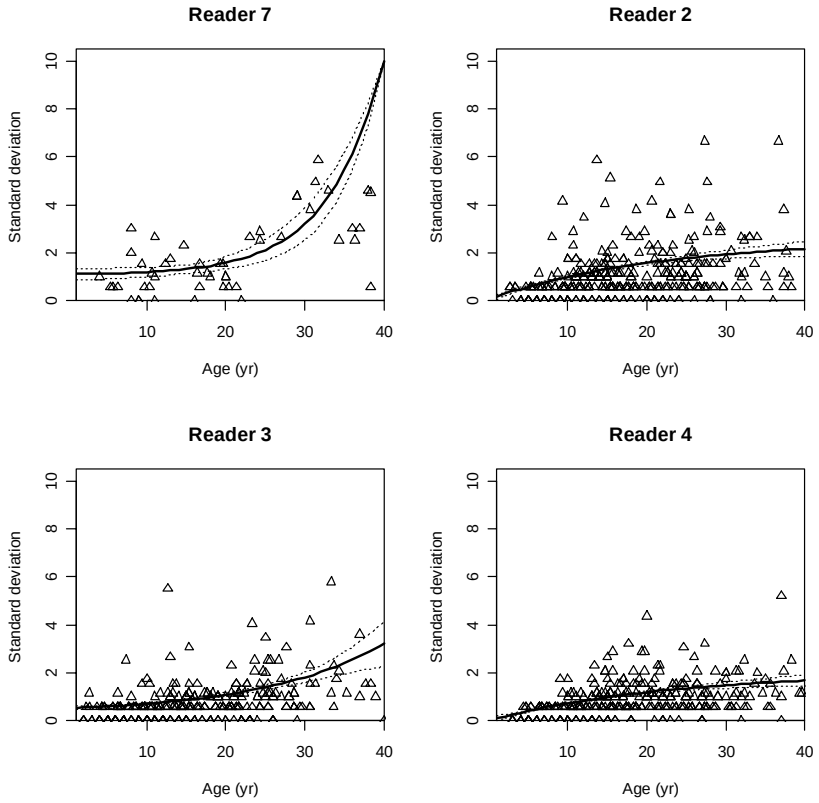

Figure 13: Fit of model #1 to the multiple age readings by readers 7, 2, 3, and 4. The triangles indicate the standard deviations of age-reading error based on those animals that were read three times, the solid lines the best estimates of the standard deviations of age-reading error, and the dotted lines the asymptotic 90% confidence intervals for standard deviation.

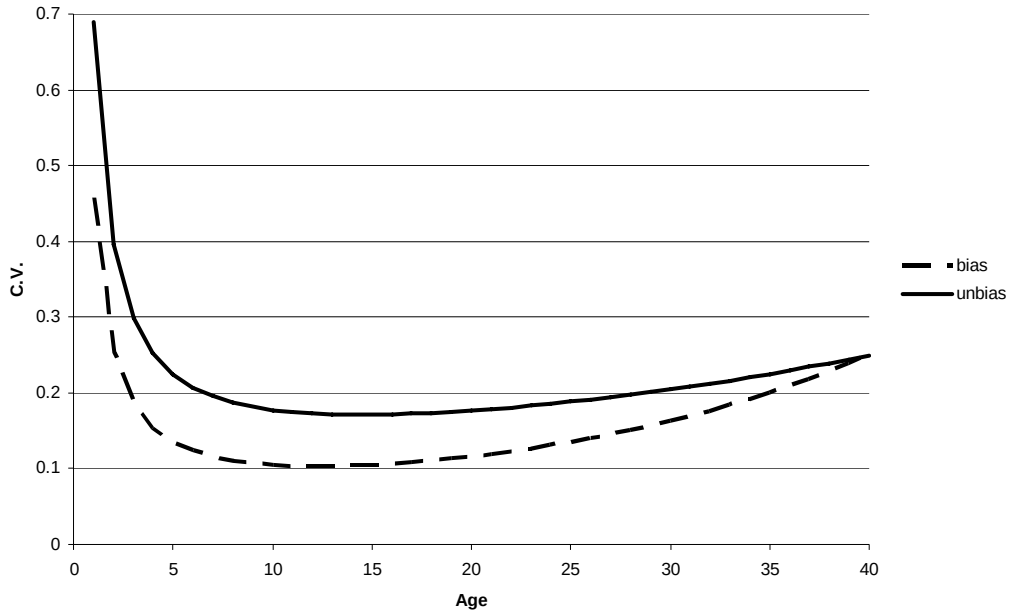

Figure 14: Comparison of the estimated age-reading error CV for reader 1 based on model #1 when the estimates from reader 1 are assumed to be biased (dashed line) and unbiased (solid line).

Appendix 1: The estimator of age-reading error

The functional form on which the probability of reader *i* (of *I* readers – a "reader" in this context could be the 1st, 2nd or 3rd read of an earplug for a single person) assigning an animal of true age *a* an age of *a'* is based, $P^i(a'|a)$, can be very general, but needs to satisfy the constraint that $\sum_{a'} P^i(a'|a) = 1$. Assuming that age-reading error is normally distributed about the expected age and that (i) ageing bias depends on reader and the true age of an animal, and (ii) the age-reading error standard deviation depends on true age and reader, leads to the following model for $P^i(a'|a)$:

$$P^i(a'|a,\underline{\phi}) \propto \exp\left[\frac{-(a'-b_a^i(\underline{\phi}))^2}{2(\sigma_a^i(\underline{\phi}))^2}\right] \qquad \text{(App. 1.1)}$$

where b_a^i is expected age when reader *i* determines the age of an animal of true age *a*, σ_a^i is the standard deviation for reader *i* of the age-reading error for animals whose true age is *a*, and $\underline{\phi}$ is the vector of parameters that determines the age-reading error matrix.

The values for the parameters that determine the age-reading error matrix for each age-reader, $P^i(a'|a)$ are estimated by maximizing the following likelihood function:

$$L(\mathbf{A}|\underline{\beta},\underline{\phi}) = \prod_{j=1}^{J}\sum_{a=L}^{H}\beta_a \prod_{i=1}^{I} P^i(a_{i,j}|a,\underline{\phi}) \qquad \text{(App.1.2)}$$

where *L* is a minimum age (1 for this paper), *H* is a maximum age (40 for this paper), *J* is the number of earplugs that have been read by all *I* readers, β_a are nuisance parameters which model the true age-structure of the sample of earplugs, and **A** denotes the total set of age readings.

The applications of this paper consider three functional forms for both the expected age, b_a, and the standard deviation of age-reading error, σ_a. They are (expressed for the expected age):

a) "Von Bertalanffy"

$$b_a = \begin{cases} b_L + (b_H - b_L)\dfrac{1-e^{-\lambda(a-L)}}{1-e^{-\lambda(H-L)}} & \text{if } \lambda \neq 0 \\ b_L + (b_H - b_L)\dfrac{a-L}{H-L} & \text{if } \lambda = 0 \end{cases} \qquad \text{(App.1.3a)}$$

where b_L is the expected age of an animal of true age *L*, b_H is the expected age of an animal of true age *H*, and λ determines the extent of non-linearity between age and the expected age.

b) "Linear"

$$b_a = b_L + (b_H - b_L)\frac{a - L}{H - L} \qquad \text{(App.1.3b)}$$

c) "Exponential"

$$b_a = b_L + (b_H - b_L)\frac{e^{\gamma a} - e^{\gamma L}}{e^{\gamma H} - e^{\gamma L}} \qquad \text{(App.1.3c)}$$

where γ is the exponential rate at which expected age increases with true age.

Appendix 2: The True Age Estimator of Age-Reading Error

Ageing error estimates were generated by assuming that the age readings of one age reader were in fact equal to the true ages and that the difference between these readings and another reader's estimates were due to a combination of systematic (i.e. bias) and non-systematic (i.e. variance) errors. Estimates of the bias and error were estimated using a Berkson error model:

$$\Pr[A \mid A^*] = \Phi((A^* + 0.5 - A_i - \beta_{A^*})/\sigma_{A^*}) - \Phi((A^* - 0.5 - A - \beta_{A^*})/\sigma_{A^*}) \qquad \text{(App.2.1)}$$

where A is the estimated age, A^* is the true age, Φ is the cumulative normal probability distribution, β_{A^*} is the bias which may be a function of the true age and σ_{A^*} = the standard deviation and may be a function of age.

Note that when $A^* = 0$, the second expression on the right hand side of the equation is set to zero. This error model takes the discrete nature of the age reading process into account. Various formulations for β and σ as functions of the true age were considered (Table App.2.1). The model number associated with the various functional relationship in this table are used in the main text to refer to the functional form associated with specific results.

Estimates for the values for the parameters for the bias and standard deviation parameters for the reader whose estimates are not assumed to be equal to the true ages are obtained by maximizing the following likelihood function:

$$L(\mathbf{A} \mid p_\beta, p_\sigma \mathbf{A}^*) = \prod_{i=1}^{I} \Pr[A_i \mid A_i^*] \qquad \text{(App.2.2)}$$

where p_β, and p_σ are the parameter vectors that determine the bias and standard deviation for the age estimates for a given true age for the reader for whom the age estimates are assumed not equal to the true age estimates, **A** and $\mathbf{A}^*$ are the vector of

estimated and assumed true age readings and I is the number of earplugs for which both readers provided an age estimate.

Table App.2.1: Functional Form used to model the bias and standard deviation. Variable definitions are as per those in Appendix 1.

Model	Descriptor	Functional expression
1	constant	B
2	constant CV	$b_a = a/b$
3	Linear*	$b_a = b_L + (b_H - b_L)\frac{a-L}{H-L}$
4	Von Bertalanffy	$b_a = \begin{cases} b_L + (b_H - b_L)\frac{1-e^{-\lambda(a-L)}}{1-e^{-\lambda(H-L)}} & \text{if } \lambda \neq 0 \\ b_L + (b_H - b_L)\frac{a-L}{H-L} & \text{if } \lambda = 0 \end{cases}$
5	Exponential*	$b_a = b_L + (b_H - b_L)\frac{e^{\gamma a} - e^{\gamma L}}{e^{\gamma H} - e^{\gamma L}}$

*Note that when a bias component related to corpus counts was included in the model no additional constant parameter was included.

Appendix 3: Definition of Codes for the Specific and General Remark Fields

Table A3.1: SPECIFIC REMARKS FIELDS - The codes used in these fields are used to clarify individual age or transition phase readings.

0 : no qualification
1 : questionable
2 : +G (i.e. plus missing germinal layers)
3 : +G? (i.e. plus missing germinal layers?)
4 : +N (i.e. plus missing neonatal layers)
5 : +N? (i.e. plus missing neonatal layers?)
6 : +? (i.e. plus missing layers?)
7 : unreliable
8 : approximately
9 : not used
10 : transition phase not present
11 : transition phase unreadable
12-19 : not used
20 : + (i.e. possibly extra layers difficult to identify)
21 : +1 (i.e. possibly one extra layer difficult to identify)
22 : +2 (i.e. possibly two extra layers difficult to identify)
23 : +3 (i.e. possibly three extra layers difficult to identify)
24 : +4 (i.e. possibly four extra layers difficult to identify)
25 : - (i.e. possibly minus some layers)
26 : -1 (i.e. possibly minus one layer)
27 : -2 (i.e. possibly minus two layers)
28 : -3 (i.e. possibly minus three layers)
29 : -4 (i.e. possibly minus four layers)

Table A3.2: GENERAL REMARKS FIELDS - The codes in these fields are used for a general description of readability and condition of earplugs.

30 : (not used)
31 : earplug consisted of two or more pieces
32 : unreadable
33 : neonatal side damaged
34 : germinal side damaged
35 : unreadable at top
36 : unreadable at bottom
37 : difficult at top
38 : difficult at bottom
39 : alternative interpretation possible (may be given in text field)
40 : difficult
41 : irregular laminae
42 : count incomplete? (comments in text field)
43 : not sure whether neonatal layer is present
44 : neonatal layer difficult to establish
45 : not sure whether germinal epithelium is present/complete
46 : readability good (These codes only apply
47 : readability poor to reader number 01)
51 : hopeless readability
52 : poor readability
53 : average readability
54 : clear readability
55 : very clear readability
61 : top broken or insufficiently cut down
62 : base damaged
63 : broken core
64 : top obscure
65 : base obscure
66 : complete
99 : see verbal remarks in the text field