NPAFC

Doc. No. 370

Rev. No. _____

The Use of Agreement Measures and Latent Class Models to Assess the Reliability of Thermally-marked Otolith Classifications

By

J. Blick and P. Hagen

Alaska Department of Fish and Game
P.O. Box 25526
Juneau, Alaska 99821-5526

Submitted to the
NORTH PACIFIC ANADROMOUS FISH COMMISSION
by the
UNITED STATES PARTY

October 1998

This paper may be cited in the following manner:
Blick, J. and P. Hagen. 1998. The Use of Agreement Measures and Latent Class Models to Assess the Reliability of Thermally-marked Otolith Classifications. (NPAFC Doc 370). 15p. Alaska Dept. Fish and Game, Juneau Alaska. 99801-5526

The Use of Agreement Measures and Latent Class Models to Assess the Reliability of Thermally-marked Otolith Classifications

Jim Blick and Pete Hagen
Alaska Department of Fish and Game
Juneau, Alaska

Abstract

Otolith thermal marking is an efficient method of mass marking hatchery-reared salmon, and with a careful sampling program can be used to determine the proportion of hatchery fish captured in a mixed stock fishery. However, the accuracy of the determination is dependent on a number of factors including the prominence of the thermal pattern, the methods used to prepare and view the patterns, and the training and experience level of the personnel who determine the presence or absence of a particular mark pattern. Estimating accuracy rates is problematic when no secondary marking is available and no error-free standards exist. Agreement measures, such as *kappa* (κ), allow a relative measure of the reliability of the determinations when independent readings by two readers are available, but the magnitude of κ can be influenced by the true proportion of marked fish. With the use of a third reader or when two or more groups of paired readings are examined, the use of latent class models allows estimation of the error rates of each reader. Applications of κ and latent class models are illustrated for multiple readings of chum and sockeye salmon otoliths as part of a quality control assessment of contribution estimates of hatchery-reared salmon to several commercial fisheries openings and site-specific locations in Southeast Alaska.

Introduction

The ability to induce patterns in the otoliths of salmon by manipulating water temperatures has proved to be an efficient means for 100% marking of salmon (Volk et al. 1990). When salmon embryos or alevins are exposed to a rapid drop in temperature, otolith growth is temporarily disrupted and this results in a discontinuity in the otolith's microstructure. When viewed under transmitted light microscopy, this discontinuity appears as a dark ring. By controlling the number of temperature drops and the timing between drops, a coded pattern of dark rings can be recorded on the otolith and this pattern can be recovered from otoliths of older fish by removing the overlaying material and exposing the otolith core. For hatcheries that release a large number of fish, this type of marking method has been shown to be particularly cost effective (Munk et al. 1993).

Several fisheries management programs in Alaska make use of thermal marking to provide estimates of hatchery contribution (Hagen et al. 1995). Typically the procedure that is used is to collect several hundred salmon otoliths systematically from each commercial opening during the season. The otoliths and sampling data are shipped to a processing laboratory where a subsample of otoliths (generally 50 to 100) are processed and read on a rapid schedule to meet in-season management needs, while a portion of the remaining otoliths are later processed to provide an overall contribution estimate.

The process by which a reader determines the presence or absence of a thermal mark in an otolith can be characterized as one of pattern recognition and image matching. Prior to examining otoliths of unknown origin, the readers gain familiarity with the patterns likely to be encountered by carefully examining otoliths of fry that were obtained after thermal marking, and prior to their release into the wild. Because there can be wide variation in the appearance of the thermal marks within a marking group (due in part to differences in developmental stages at marking), a single mark group may be represented by a variety of patterns. As a result, secondary characteristics and measurements of the patterns are sometimes necessary to identify an otolith to a marked group. The examination is also used to confirm if the hatchery fish have been 100% marked.

The process of making a determination on the otoliths from the returning adult salmon can become problematic because wild salmon may also contain otolith patterns which can mimic the features imposed through thermal marking. Referred to as noisy patterns, their presence can increase the rate of false positives. Conversely if the hatchery employs poor temperature control, or unintended disruptions occur around the period of marking, it may be difficult to identify the otolith as a hatchery fish and this would increase the rate of false negatives. Differences between readers in skill and training level, and how they process otoliths can add to the uncertainty in estimating the precision of the readings and the rates of false positives and negatives.

Otolith marking generally takes place without any secondary marking, such as fin-clipping or coded-wire tagging, so the accuracy of a reading can not directly be determined through conventional methods which make use of a "gold standard" or other error-free classification methods. In order to ensure the information provided to the Alaskan fisheries managers is accurate, each otolith is independently examined by two readers, and a third reading is used to resolve differences between the first two readings. The resolved readings are used to estimate the contribution of hatchery fish and the precision of the estimate is based on the assumption that, through multiple readings, all marked fish are either correctly identified or that errors, if present, are inconsequential. Developing the analytical tools to determine the veracity of that assumption is the objective of this investigation and by establishing such tools, quality control standards for recovering thermal marks can be developed.

In developing the tools to measure the quality of otolith readings, two questions are addressed:

1. How to assess the reliability of otolith readings when no standards are available?
2. How to estimate the proportion of hatchery marks when there is disagreement between two or more readers?

We are also interested in assessing how the variance of the estimate of the proportion is influenced by classification error. Although we will indicate how such an influence can be estimated, we will not provide a detailed study of this question in this report.

We discuss two approaches: 1) indices of agreement typically used in reliability studies, and 2) latent class models where classification errors are estimated for each reader even though the true error rate is considered unknown. The data requirements and their attendant assumptions are presented for each approach. The methods are illustrated by examining among-reader comparisons on chum and sockeye salmon otoliths collected from programs

that monitor inseason contributions of hatchery fish in several commercial fisheries in Southeast Alaska (Hagen et al. 1995). The results are used to provide recommendations for monitoring quality of otolith readings for thermal marking programs.

Notation

The following notation will be used:

H	denotes hatchery stock(s)
W	denotes wild stock(s)
n	sample size
$\pi_{i\|j}^{(k)}$	probability that reader k classifies an otolith as i when its true state is j
p	proportion of hatchery stock in the catch
^	represents an estimate when placed over π or p
.	used as a subscript to denote a marginal total
df	degrees of freedom

Standard Available

A sample of n otoliths, which are examined by two readers, can be cross-classified as follows:

		Reader 2 H	Reader 2 W	
Reader 1	H	n_{HH}	n_{HW}	$n_{H\cdot}$
	W	n_{WH}	n_{WW}	$n_{W\cdot}$
		$n_{\cdot H}$	$n_{\cdot W}$	n

If Reader 2 is infallible (or is considered a "gold standard"), unbiased estimates of the accuracy and error rates of Reader 1 are given by:

$$\hat{\pi}_{H|H} = n_{HH}/\, n_{\cdot H}\,, \qquad \hat{\pi}_{W|H} = n_{WH}/\, n_{\cdot H} = 1 - \hat{\pi}_{H|H}$$

$$\hat{\pi}_{W|W} = n_{WW}/\, n_{\cdot W}\,, \qquad \hat{\pi}_{H|W} = n_{HW}/\, n_{\cdot W} = 1 - \hat{\pi}_{W|W}$$

$$\hat{p} = n_{\cdot H}/n$$

These estimates reflect the fact that Reader 2 is infallible: the accuracy rates ($\hat{\pi}_{H|H}$, $\hat{\pi}_{W|W}$) and the error rates($\hat{\pi}_{W|H}$, $\hat{\pi}_{H|W}$) are conditional on the numbers of hatchery or wild stock otoliths as determined by Reader 2.

No Standard Available

If Reader 2 is also subject to error, the above estimates are no longer unbiased. If the accuracy rates are known, an unbiased estimate of p is:

$$\hat{p}^* = (n_{.H}/n + \pi_{W|W} - 1)/(\pi_{H|H} + \pi_{W|W} - 1)$$

If the accuracy rates are estimated, then $\hat{p}^*$ will no longer be unbiased, but it will be much less biased than the estimator $n_{.H}/n$, and will in general have much smaller mean squared error (Rogan and Gladen 1978).

When accuracy rates are unavailable, statistics which measure "agreement" between readers are often calculated (e.g., Fleiss 1981, ch. 13). One such index is simply the proportion of observed agreement (P_o) defined as:

$$P_o = (n_{HH} + n_{WW})/n$$

Another index, called *kappa* (κ), corrects P_o for the degree of agreement that is expected by chance alone. It is defined as:

$$\kappa = (P_o - P_e)/(1 - P_e)$$

where P_e = expected agreement $= (n_H.n_{.H} + n_W.n_{.W})/n^2$. The divisor, $1 - P_e$, constrains κ to be less than or equal to one, and if all agreement is due to chance ($P_o = P_e$), κ equals zero. Note that κ assumes independence between readers in order to calculate expected agreement.

An example of how agreement indices can be used to monitor readings is shown in Figure 1, which displays κ and its standard error for 2,874 chum otoliths readings divided into 27 groups based on different reader pairs and capture locations. Included are P_o's for four of the groups. The results indicate that κ levels were similar between the different groups suggesting overall consistency in readings, though some of the groups had lower values which, in practice, would invite further investigation.

The P_o's in Figure 1 have a different rank order than the κ values. This apparent discrepancy highlights a potential problem in interpretation when using agreement indices to draw conclusions. To help illustrate this point, consider the following examples.

Table (1a) is generated as the expected counts given $\pi_{H|H} = 0.9$ and $\pi_{W|W} = 1.0$ for both readers, and p = 0.1. In this case, $P_o = 0.98$ and $\kappa = 0.89$. Table (1b) is generated under the same assumptions except that $\pi_{H|H} = 0.5$. In this case P_o drops only slightly to 0.95, while κ drops to 0.47. Because the hatchery stock is rare, the inability of the readers to detect the

mark is not well reflected by P_o, whereas κ reflects it better by correcting for the high level of chance agreement.

1a)

		Reader 2 H	Reader 2 W	
Reader 1	H	81	9	90
	W	9	901	910
		90	910	1000

1b)

		Reader 2 H	Reader 2 W	
Reader 1	H	25	25	50
	W	25	925	950
		50	950	1000

On the other hand, Table (2a) is generated as the expected counts given $\pi_{H|H} = 0.9$ and $\pi_{W|W} = 0.9$ for both readers, and $p = 0.5$. In this case, $P_o = 0.82$ and $\kappa = 0.64$. Table (2b) is generated under the same assumptions except that $p = 0.05$. In this case P_o remains unchanged at 0.82, but κ drops to 0.25.

2a)

		Reader 2 H	Reader 2 W	
Reader 1	H	410	90	500
	W	90	410	500
		500	500	1000

2b)

		Reader 2		
		H	W	
Reader 1	H	50	90	140
	W	90	770	860
		140	860	1000

In neither of the above examples is the index "wrong". Rather, as is the case with most indices, interpretation is affected by the values of the underlying parameters. In the latter example, even though P_o is the same for both tables, the scale it is being compared to has changed, thus changing the value of κ. This makes difficult the comparison of κ across populations with different underlying proportions. Note also that Table (2b) could have been derived from the following parameter values: $\pi_{H|H} = 0.5$ and $\pi_{W|W} = 0.944$ for both readers, and $p = 0.19$. Thus, without additional information, it is impossible to draw reliable conclusions about reader accuracies or the proportion of hatchery marks.

While agreement measures can be subject to ambiguous interpretations, in practice they can still serve a useful monitoring role during routine comparisons when the circumstances of the readings are fairly well characterized. It is when trying to translate agreement measures into statements about the accuracy of different readers and the influence of reading error on the contribution estimates, that the interpretive difficulties with indices such κ and P_o become apparent.

Alternative Approach: Latent Class Models

An alternative approach is to try to estimate $\pi_{H|H}$ and $\pi_{W|W}$ for each reader, along with p. Although at first thought this may seem impossible, it can be shown that either by setting a few constraints or by collecting additional information, estimation is indeed possible. This problem falls into the category of latent class modeling (e.g., Everitt 1984; Bartholomew 1987; McCutcheon 1987; Clogg, 1995). Latent class models (LCM) belong to a family of latent variable models which hypothesize the existence of unobservable "latent" variables about which information can only be obtained though measurements on observable "manifest" variables. LCMs specifically restrict the latent and manifest variables to be categorical. In the present situation, the latent variable is the true class (H or W) to which the otolith belongs, while the manifest variables are the readers' classifications. Such models have been used for assessing reliability of diagnostic tests in the medical field over the last twenty years (see Walter and Irwig 1988; Formann 1996, for reviews).

Returning to the problem with two readers, neither of which is a standard, note that there are five essential parameters to estimate: $\pi_{H|H}^{(1)}$, $\pi_{H|H}^{(2)}$, $\pi_{W|W}^{(1)}$, $\pi_{W|W}^{(2)}$, and p, with only 3 *df*

(four pieces of data, n_{HH}, n_{HW}, n_{WH}, n_{WW}, minus one since the sample size, n, is fixed). Thus, the model is overparameterized and either constraints on the parameters or more data are needed.

Possible constraints include: (a) considering two of the parameters known; e.g., $\pi_{W|W}^{(1)} = \pi_{W|W}^{(2)} = 1$ (i.e., both readers always call a wild stock correctly – there are no "false positives"), or (b) considering two sets of parameters equal; e.g., $\pi_{H|H}^{(1)} = \pi_{H|H}^{(2)}$, $\pi_{W|W}^{(1)} = \pi_{W|W}^{(2)}$ (i.e., the accuracy rates are the same for both readers).

Although there may be times when such constraints are realistic, in general they will not, so that more information will be necessary. One way to generate more information is to have a third independent reader (Walter 1984). With three readers, there are seven essential parameters: $\pi_{H|H}^{(1),(2),(3)}$, $\pi_{W|W}^{(1),(2),(3)}$, and p. There are also $2^3 - 1 = 7$ *df*, so all the parameters are estimable. Estimation is most commonly done by the method of maximum likelihood.

Assuming readings are independent among readers and among otoliths, the likelihood function is:

$$\prod_{i=H,W}\prod_{j=H,W}\prod_{k=H,W}\{p\pi_{i|H}^{(1)}\pi_{j|H}^{(2)}\pi_{k|H}^{(3)} + (1-p)\pi_{i|W}^{(1)}\pi_{j|W}^{(2)}\pi_{k|W}^{(3)}\}^{n_{ijk}}$$

This likelihood function must be maximized numerically and methods for this computation will be discussed later.

If more than three readers are used, there are extra *df* which can be used to assess goodness-of-fit. For example, with four readers there will be nine parameters with 15 *df*, leaving 6 *df* for goodness-of-fit. Pearson chi-square or likelihood ratio G^2 tests would both be applicable.

Another way to generate additional information was proposed by Hui and Walter (1980). Suppose there are two or more strata with different hatchery proportions in each strata. For example, catch could be stratified temporally or spatially. If it is assumed that $\pi_{H|H}^{(k)}$ and $\pi_{W|W}^{(k)}$ remain constant over strata, then a solution for just two readers may be obtained. For example, if there are two readers and two strata, then there are six parameters: $\pi_{H|H}^{(1),(2)}$, $\pi_{W|W}^{(1),(2)}$, p_1, and p_2, with $2(2^2 - 1) = 6$ *df*. Increasing the number of strata increases the df; e.g., three strata for two readers gives $3(2^2 - 1) = 9$ *df* for 7 parameters. The likelihood function for two readers and S strata is:

$$\prod_{g=1}^{S}\prod_{i=H,W}\prod_{j=H,W}\{p_g\pi_{i|H}^{(1)}\pi_{j|H}^{(2)} + (1-p_g)\pi_{i|W}^{(1)}\pi_{j|W}^{(2)}\}^{n_{gij}}$$

A third way to supply additional information is to take a Bayesian approach. By specifying prior distributions of the model parameters, unique estimates can be obtained (Evans et al. 1989).

Assumptions of the Latent Class Model

A critical assumption in the above models is that readings are independent. Specifically, the reading of each otolith by a given reader is independent of any other reading by the same reader, and each reading by various readers on a given otolith are independent given the true state of the otolith. The latter assumption may be hard to meet especially if all readers examine the same otolith. The fact that the otolith is not prepared independently by each reader will probably induce a dependence among the readers. Also, variability in the readability of the mark due to the marking process can induce a dependence. Such dependence can bias the estimators of π and p (Vacek 1985). Note that this latter assumption of independence is also required for κ.

One remedy to the problem of dependence due to preparation is to require independent preparations. With only two otoliths per fish, this would limit the number of readers to two. A more general solution is to model the dependence with additional parameters (e.g., Vacek 1985, Qu et al. 1996, Qu and Hagdu 1997, Yang and Becker 1997). Modeling dependence requires either more readers or more strata. Alternatively, additional latent classes may be added (Formann 1994); e.g., a third class of otoliths whose origin is ambiguous.

In the previous discussion concerning three or more readers, it was implied that readers were different individuals. This need not be so: what is required is three or more independent readings. If it were possible for the same individual to read more than once independently, the number of different readers could be reduced. It seems unlikely, however, that independence could be achieved, although the dependence could be modeled as discussed above.

Another critical assumption, but one that should be met most of the time, is that the individual accuracy rates are known to be either greater than or less than the error rates (e.g., $\pi_{H|H} > \pi_{W|H}$ and $\pi_{W|W} > \pi_{H|W}$). This is because of an inherent symmetry in the problem that results in the same likelihood function being generated when the error rates are switched with the accuracy rates.

Computation

Maximizing either of the likelihood functions given above requires a numerical procedure. Several possibilities exist. The most straightforward is to use an optimization routine such as "Solver" in Excel (Microsoft Corporation 1993) or "nlminb" in S-PLUS (Statistical Sciences 1995).

Alternatively, the EM algorithm (Dempster et al. 1977; Dawid and Skene 1979; McLachlan and Krishnan 1997) can be easily used. The simplicity of the EM algorithm follows from the recognition that the LCM is an example of a finite mixture problem, specifically in this case a mixture of multivariate Bernoulli distributions with mixing parameter p (Everitt 1984). Use of the EM algorithm for such mixture problems in fisheries is well documented; e.g., for stock composition estimates (Millar 1987, Pella et al. 1996), and for age-length keys (Kimura and Chikuni 1987).

A more efficient alternative to the EM algorithm is to use iteratively reweighted least squares (Green 1984, Agresti 1990). This method is relatively easy to implement in software such as PROC NLIN in SAS (SAS Institute 1989).

Finally, perhaps the most direct and efficient way would be to use LCM software, if it were available. Although several independent LCM packages exist (see Clogg 1995, for review), we are not aware of any routines for LCMs in any major statistical package at present.

As with many maximum likelihood problems where numerical methods must be used, complications can arise. Although the EM algorithm guarantees that the parameter estimates will fall in the interval [0,1], the other methods may at times need constraints. Also the likelihood function may have local maxima which means that several runs with varying starting values may be necessary to identify the global maximum. Finally, estimates of standard errors may entail additional computing. PROC NLIN in SAS provides asymptotic (i.e., large-sample) standard errors. Jackknife and bootstrap estimates are relatively easy to program with the jackknife being much less computationally intensive.

Examples

The first example analyzes the results of three readers examining 570 chum otoliths. The samples were taken from a common location and the readers were familiar with the patterns. Each reading was made independent of the others readings. The data are presented below along with pairwise *kappa* estimates and their standard errors:

HHH	406
HHW	13
HWH	1
WHH	1
HWW	6
WHW	2
WWH	6
WWW	135

Reader Pairs	κ	SE(κ)
1 & 2:	0.95	0.014
1 & 3:	0.88	0.022
2 & 3:	0.90	0.021

Estimates of the LCM parameters (using PROC NLIN in SAS; see appendix for code) are given below with the asymptotic standard error:

Parameter	Estimate	SE
$\pi_{H\|H}^{(1)}$	0.998	0.002
$\pi_{H\|H}^{(2)}$	0.998	0.002
$\pi_{H\|H}^{(3)}$	0.969	0.008
$\pi_{W\|W}^{(1)}$	0.958	0.017
$\pi_{W\|W}^{(2)}$	0.986	0.010
$\pi_{W\|W}^{(3)}$	0.957	0.017
p	0.738	0.018

These results indicate that the third reader is significantly less able to correctly identify a hatchery mark when it is present and that the second reader appears (though it is not statistically significant) to be better able to detect a wild mark when it is present. These conclusions are readily apparent from the table of results, and although the pairwise *kappa*s are consistent with these results, they are more difficult to interpret. If an infallible method had estimated the same value for p, the variance would have been estimated as (.7379)(1 - .7379)/(570 – 1) = .0003399, compared to $.01847^2$ = .0003411 from the LCM. This is less than 0.5% loss of efficiency for this particular example.

The second example consists of two readers with four spatial strata. Samples were obtained from sockeye salmon caught in four neighboring Alaskan gillnet fisheries in central Southeast Alaska. The data are shown below:

	Statistical Area			
	108-30	**108-50**	**106-41**	**106-30**
HH	152	127	85	20
HW	11	9	21	5
WH	2	6	5	1
WW	271	382	832	411
n	436	524	943	437

LCM estimates are as follows:

Parameter	Estimate	Standard Error
$\pi_{H\|H}^{(1)}$	0.980	0.013
$\pi_{H\|H}^{(2)}$	0.964	0.021
$\pi_{W\|W}^{(1)}$	0.984	0.005
$\pi_{W\|W}^{(2)}$	0.997	0.003
$p_{108\text{-}30}$	0.366	0.024
$p_{108\text{-}50}$	0.257	0.020
$p_{106\text{-}41}$	0.096	0.010
$p_{106\text{-}30}$	0.047	0.011

These estimates indicate that the first reader is better able to detect hatchery marks, while the second reader is better able to distinguish wild marks. With eight parameters and twelve *df*, there are four *df* available for a goodness-of-fit test. Pearson's chi-square yields 4.83 which with 4 *df* has a p-value of 0.306, thus indicating an acceptable model fit. The loss in efficiency of the LCM estimates of the various p's compared to an infallible estimate (computed as in the previous example) ranges from about 8% to 14%, although the magnitude of the loss is very small.

Discussion

There are numerous classification problems in fisheries that require the judgement of trained individuals. In many of those situations no "gold standard" is available to test those judgements and it becomes necessary to apply other methods to determine the veracity of the classifications. Reading thermally marked otoliths is a particularly good example of that problem because thousands of classification decisions are needed each year to provide estimates of hatchery contributions.

The common approach for assessing the quality of the readings, in the absence of having samples of known origin, has been to collect independent and multiple readings on the samples, and presume that agreement between readings can serve as a proxy for reading accuracy. Agreement indices such as *kappa* are very easy to compute and they have utility in that they can serve as flags to indicate reading problems. However, as was shown here, they also suffer difficulties in interpretation. In addition, if a particular suite of *kappas* is found to be consistent, it is not clear what that says about the influence of reader disagreement on the contribution estimate. Also, the indices in themselves do not provide inferences about the relative skill of one reader versus another in pulling out a particular set of patterns.

Latent class models provide an alternative approach with readily interpretable quantities for a modest computational cost. Classification accuracies or errors are direct, meaningful parameters unlike an index of agreement. In addition, estimates of p are available. These models can be readily extended to the case of more than two outcomes; e.g., multiple hatchery marks. These models could also be useful in other applications such as aging fish or in the identification of any character for which there is no "gold standard" (e.g., field identification of species or sex). A somewhat similar analysis has been proposed for aging (Richards et al. 1992), though the link to LCMs was not discussed. LCMs can handle fairly complicated situations, including ordered classes (Croon 1990), continuous manifest variables, and parameter constraints (see Clogg 1995, Krzanowski and Marriott 1995 for reviews). We feel that further study of the application of LCMs to fishery classification problems would be highly worthwhile.

References

Agresti, A. 1990. Categorical data analysis. John Wiley, New York.

Bartholomew, D.J. 1987. Latent variable models and factor analysis. Oxford Univ. Press., New York

Clogg, C.C. 1995. Latent class models. Chapter 6 *in* G. Arminger, C.C. Clogg, and M.E. Sobel, editors. Handbook of statistical modeling for the social and behavioral sciences. Plenum Press, New York.

Croon, M. 1990. Latent class analysis with ordered classes. Brit. J. Math. Stat. Psych. 43:171-192.

Dawid, A.P. and A.M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. Appl. Statist. 28:20-28.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Royal Stat. Soc. B 39:1-38.

Evans, M.J., Z. Gilula, and I. Guttman. 1989. Latent class analysis of two-way contingency tables by Bayesian methods. Biometrika 76:557-63.

Everitt, B.S. 1984. An introduction to latent variable models. Chapman and Hall, London

Fleiss, J.L. 1981. Statistical methods for rates and proportions, 2nd ed. John Wiley, New York.

Formann, A.K. 1994. Measurement errors in caries diagnosis: some further latent class models. Biometrics 50:865-871.

Formann, A.K. 1996. Latent class analysis in medical research. Stat. Meth. Med. Res. 5:179-211.

Green, P.J. 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). J. Royal Stat. Soc. B 46:149-192.

Hagen, P., K. Munk, B. Van Alen, and B. White. 1995. Thermal mark technology for inseason fisheries management: a case study. Alaska Fishery Research Bulletin 2:143-158.

Hui, S.L. and S.D.Walter. 1980. Estimating the error rates of diagnostic tests. Biometrics 36:167-171.

Kimura, D.K. and S. Chikuni. 1987. Mixtures of empirical distributions: an iterative application of the age-length key. Biometrics 43:23-35.

Krzanowski, W.J. and F.H.C. Marriott. 1995. Multivariate analysis, part 2: classification, covariance structures and repeated measurements. Arnold, London.

McCutcheon, A.L. 1987. Latent class analysis. Sage, Beverly Hills, California.

McLachlan, G.J. and T. Krishnan. 1997. The EM algorithm and extensions. John Wiley, New York.

Microsoft Corporation. 1993. Microsoft Excel user's guide. Microsoft Corporation, Redmond, Washington.

Millar, R.B. 1987. Maximum likelihood estimation of mixed stock fishery composition. Can. J. Fish. Aquat. Sci. 44:583-590.

Munk, K.M., W.W. Smoker, D.R. Beard, and R.W. Mattson. 1993. A hatchery water-heating system and its application to 100% thermal marking of incubating salmon. Progressive Fish-Culturist 55:284-288.

Pella, J., M. Masuda, and S. Nelson. 1996. Search algorithms for computing stock composition of a mixture from traits of individuals by maximum likelihood. U.S. Dept. Commerce, NOAA Tech. Memo. NMFS-AFSC-61.

Qu, Y., M. Tan, and M.H. Kutner. 1996. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 52:797-810.

Qu, Y. and A. Hagdu. 1997. Modeling correlations between diagnostic tests in efficacy studies with an imperfect reference test. *in* Gregoire, T.G., D.R. Brillinger, P.J. Diggle, E. Russek-Cohen, W.G. Warren, and R.D. Wolfinger. Modelling longitudinal and spatially correlated data. Springer, New York.

Richards, L.J., J.T. Schnute, A.R. Kronlund, and R.J. Beamish. 1992. Statistical models for the analysis of ageing error. Can. J. Fish. Aquat. Sci. 49:1801-1815.

Rogan, W.J. and B. Gladen. 1978. Estimating prevalence from the results of a screening test. Amer. J. Epidemiology 107:71-76.

SAS Institute. 1989. SAS/STAT user's guide, version 6, 4th ed, vol. 2. SAS Institute, Cary, North Carolina.

Statistical Sciences. 1995. S-PLUS guide to statistical and mathematical analysis, version 3.3. StatSci, Seattle.

Vacek, P.M. 1985. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 41:959-968.

Volk, E.C., S.L. Schroder, and K.L. Fresh. 1990. Inducement of unique otolith banding patterns as a practical means to mass-mark juvenile Pacific salmon. American Fisheries Society Symposium 7:203-215.

Walter, S.D. 1984. Measuring the reliability of clinical data: the case for using three observers. Rev. Epidém. et Santé Publ. 32:206-211.

Walter, S.D. and L.M. Irwig. 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J. Clin. Epidemiol. 41:923-937.

Yang, I. and M.P. Becker. 1997. Latent variable modeling of diagnostic accuracy. Biometrics 53:948-958.

Appendix

The following SAS (version 6.12) code was used to estimate parameters in the 3-reader model discussed above. This program makes use of iteratively reweighted least squares to maximize the likelihood function. Observed values (e.g., the number of HHH) are equated with the corresponding expected value from the model and a weighted least squares fit is computed using PROC NLIN. This is iterated to convergence of the parameter estimates. Weights are inverses of the predicted values at each iteration. Indicator variables for each possible outcome are generated so that a model in typical regression form can be written. Bounds on the parameter estimates may be needed to constrain the estimates to the interval [0,1]. Note that the asymptotic standard errors provided by SAS will be correct if the option 'SIGSQ=1' is specified. However, the printed degrees of freedom and the associated confidence intervals are not correct. The residual weighted sum of squares listed by SAS is the chi-squared goodness-of-fit-statistic.

SAS code for the multi-reader/multi-strata is also available from the authors.

SAS code for the 3-reader (one stratum) model:

```
%let n=570;

data a;
  array x{8} x1-x8;
  input y;
  do i=1 to 8;
    if i=_n_ then x{i}=1; else x{i}=0;     /* set up indicator variables */
  end;
  cards;
 406          /* H H H */
  13          /* H H W */
   1          /* H W H */
   1          /* W H H */
   6          /* H W W */
   2          /* W H W */
   6          /* W W H */
 135          /* W W W */
;

proc nlin data=a nohalve sigsq=1;    /* sigsq=1 necessary for correct se's  */
  parms a1=.9 a2=.9 a3=.9 b1=.9 b2=.9 b3=.9 p=.6;  /* starting values */
  e1=a1*a2*a3*p+(1-b1)*(1-b2)*(1-b3)*(1-p);
  e2=a1*a2*(1-a3)*p+(1-b1)*(1-b2)*b3*(1-p);
  e3=a1*(1-a2)*a3*p+(1-b1)*b2*(1-b3)*(1-p);
  e4=(1-a1)*a2*a3*p+b1*(1-b2)*(1-b3)*(1-p);
  e5=a1*(1-a2)*(1-a3)*p+(1-b1)*b2*b3*(1-p);
  e6=(1-a1)*a2*(1-a3)*p+b1*(1-b2)*b3*(1-p);
  e7=(1-a1)*(1-a2)*a3*p+b1*b2*(1-b3)*(1-p);
  e8=(1-a1)*(1-a2)*(1-a3)*p+b1*b2*b3*(1-p);
  model y=(e1*x1+e2*x2+e3*x3+e4*x4+e5*x5+e6*x6+e7*x7+e8*x8)*&n;
  bounds 0<=a1<=1,0<=a2<=1,0<=a3<=1,0<=b1<=1,0<=b2<=1,0<=b3<=1,0<=p<=1;
  _weight_=1/model.y;
run;
```

Figure 1. The values of *kappa* $\pm$ *1 SE* from 27 groups of paired readings of chum salmon otoliths (total = 2,874). The groups are based on pairs of different readers examining otoliths collected at different times and locations. The proportion of agreement (P_o) is shown next to group 4, 7, 9 and 12 for comparison with the value of *kappa*.