Abstract–Otolith thermal marking is an efficient method for mass marking hatchery-reared salmon and can be used to estimate the proportion of hatchery fish captured in a mixed-stock fishery. Accuracy of the thermal pattern classification depends on the prominence of the pattern, the methods used to prepare and view the patterns, and the training and experience of the personnel who determine the presence or absence of a particular pattern. Estimating accuracy rates is problematic when no secondary marking is available and no error-free standards exist. Agreement measures, such as *kappa* (κ), provide 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 proportion of marked fish. If a third reader is used or if two or more groups of paired readings are examined, latent class models can provide estimates of the error rates of each reader. Applications of κ and latent class models are illustrated by a program providing contribution estimates of hatchery-reared chum and sockeye salmon in Southeast Alaska.

Manuscript accepted 16 April 2001.
Fish. Bull. 100:1–10 (2002).

The use of agreement measures and latent class models to assess the reliability of classifying thermally marked otoliths*

D. James Blick

Peter T. Hagen

Alaska Department of Fish and Game
Division of Commercial Fisheries
10107 Bentwood Place
Juneau, Alaska 99802-5526
E-mail address (for P. T. Hagen, contact author): peter_hagen@fishgame.state.ak.us

The ability to induce patterns in salmon otoliths by manipulating water temperatures has proved to be an efficient means for marking large numbers 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 shown to be particularly cost effective for marking 100% of the releases (Munk et al., 1993).

Several fisheries management programs in Alaska use thermal marking to estimate hatchery contributions to commercial fisheries (Hagen et al., 1995). Typically, several hundred salmon otoliths are systematically collected during each two- or three-day commercial opening during the fishing season. The otoliths and sampling data are shipped to a processing laboratory where a subsample of otoliths (generally 50 to 100) are processed immediately to meet in-season management needs; a portion of the remaining otoliths are processed later to provide an overall estimate of hatchery contribution to the fisheries.

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 fry otoliths that were obtained after thermal marking but prior to their release into the wild. Because there can be wide variation in the appearance of the thermal marks within a mark 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 mark group. The examination is also used to confirm that all the hatchery fish have been successfully marked.

The process of making a determination on otoliths from returning adult salmon can become problematic because wild salmon may also contain otolith patterns that 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 that of a

* Contribution PP-184 of the Alaska Department of Fish and Game, Commercial Fisheries Division, Juneau, Alaska 99802-5526.

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 accuracy 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; therefore the accuracy of a reading cannot directly be determined through conventional methods that make use of a “gold standard” (known origin sample) or other error-free classification methods. To ensure that 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 presumption of accuracy is based on the premise 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, three 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.
3 How the precision of the estimate of the proportion is influenced by classification error.

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 of chum salmon (*Oncorhynchus keta*) and sockeye (*Oncorhynchus nerka*) 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 the quality of otolith readings for thermal marking programs.

Methods

Standard available

Table 1

Notation used to show the cross-classification of a sample of *n* otoliths by two readers to either hatchery (H) or wild stock (W) assignment. Row and column sums are indicated by the subscript “.”

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

A sample of *n* otoliths, which are examined by two readers, can be cross-classified as hatchery (H) or wild stock (W) as in Table 1. Suppose we wish to estimate the accuracy rate (probability of making a correct classification) or conversely, the error rate (probability of making a wrong classification). If we know nothing about reader 1, but reader 2 is infallible (or is considered a “gold standard”), unbiased estimates of the accuracy and error rates of reader 1 and the proportion of hatchery stocks (p) are given by

$$\hat{\pi}^{(1)}_{\mathrm{H|H}} = n_{\mathrm{HH}}/n_{\cdot\mathrm{H}}, \qquad \hat{\pi}^{(1)}_{\mathrm{W|H}} = n_{\mathrm{WH}} / n_{\cdot\mathrm{H}} = 1 - \hat{\pi}_{\mathrm{H|H}}$$
$$\hat{\pi}^{(1)}_{\mathrm{W|W}} = n_{\mathrm{WW}}/n_{\cdot\mathrm{W}}, \qquad \hat{\pi}^{(1)}_{\mathrm{H|W}} = n_{\mathrm{HW}} / n_{\cdot\mathrm{W}} = 1 - \hat{\pi}_{\mathrm{W|W}}$$
$$\hat{p} = n_{\cdot\mathrm{H}}/n,$$

(where, for example, $\pi^{(1)}_{\mathrm{W|H}}$ refers to the probability that reader 1 classifies an otolith as W when its true state is H). These estimates reflect the fact that reader 2 is infallible; the accuracy rates ($\hat{\pi}_{\mathrm{H\,|\,H}}$, $\hat{\pi}_{\mathrm{W\,|\,W}}$) and the error rates ($\hat{\pi}_{\mathrm{W\,|\,H}}$, $\hat{\pi}_{\mathrm{H\,|\,W}}$) are conditional on the numbers of hatchery or wild stock otoliths as determined by reader 2.

No standard available

If a standard is not available, an unbiased estimate of p can be obtained if the accuracy rates for reader 1 are known. The estimate is

$$\hat{p}^* = (n_{\mathrm{H}}/n + \pi^{(1)}_{\mathrm{W|W}} - 1)\big/(\pi^{(1)}_{\mathrm{H|H}} + \pi^{(1)}_{\mathrm{W|W}} - 1),$$

where n_{H} is the number of otoliths classified as hatchery otoliths. If the accuracy rates are estimated, then $\hat{p}^*$ will no longer be unbiased, but will be much less biased than the estimator n_{H}/n and will in general have a much smaller mean-squared error (Rogan and Gladen, 1978). For a Bayesian approach to this problem, see Viana et al. (1993) and Joseph et al. (1995).

Agreement measures When accuracy rates are unavailable, statistics that measure “agreement” between readers are often calculated (e.g. Fleiss, 1981). One such index is simply the proportion of observed agreement (P_o), defined as

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

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 = $(n_H n_{\cdot H} + n_W n_W)/n^2$. 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), then κ equals zero. Note that with κ, independence between readers is assumed 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 2874 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, although 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 2). Table 2A 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 κ = 0.89. On the other hand, Table 2B is generated under the same assumptions except that $\pi_{H|H}$ = 0.5. In this case P_o drops only slightly to 0.95, whereas 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.

Now let $\pi_{H|H}$ = 0.9 and $\pi_{W|W}$ = 0.9 for both readers, and P= 0.5 (Table 2C). In this case, P_o = 0.82 and κ = 0.64. On the other hand, Table 2D 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.

In none 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 (Table 2, C–D), even though P_o is the same for C and D, the scale it is being compared with has changed, thus changing the value of κ. This increases the difficulty of comparing κ across populations with different underlying proportions. Note also that Table 2D could have been derived from $\pi_{H|H}$ = 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.

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

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 (LCMs) belong to a family of latent variable models that hypothesize the existence of unobservable "latent" variables, about which information can be obtained only 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, whereas 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 20 years (see Walter and Irwig, 1988; Formann, 1996, for reviews).

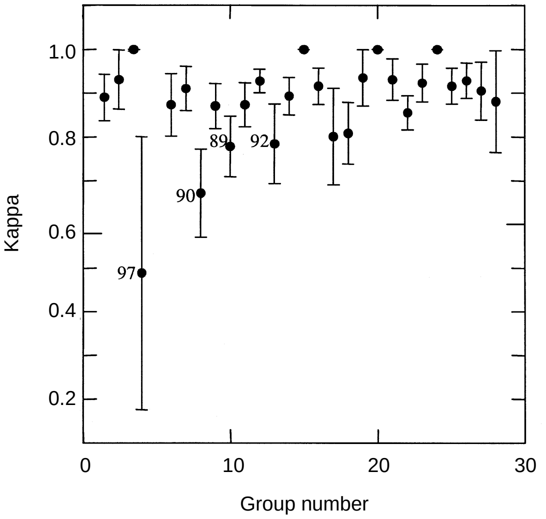

Figure 1

The values of κ (±1 SE) from 27 groups of paired readings of chum salmon otoliths (total=2874). 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 κ.

Returning to the problem with two readers, neither of which is a standard, there are five essential parameters to estimate: $\pi^{(1)}_{H|H}, \pi^{(2)}_{H|H}, \pi^{(1)}_{W|W}, \pi^{(2)}_{W|W}$, and p, with only 3 df (four pieces of data, n_{HH}, n_{HW}, n_{WH}, n_{WW}, minus one because 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 1) considering that two of the parameters are known (e.g. $\pi^{(1)}_{W|W} = \pi^{(2)}_{W|W} = 1$; i.e. both readers always call a wild stock correctly, there are no "false positives"), or 2) considering that two sets of parameters are equal (e.g. $\pi^{(1)}_{H|H}, \pi^{(2)}_{H|H}, \pi^{(1)}_{W|W} = \pi^{(2)}_{W|W}$; 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 be; therefore 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: $\pi^{(1),(2),(3)}_{\mathrm{H|H}}, \pi^{(1),(2),(3)}_{\mathrm{W|W}}$ and p. There is also $2^3 - 1 = 7$ df, so that all the parameters are estimable. Estimation is most commonly done by the method of maximum likelihood.

If readings are assumed to be independent among readers and among otoliths, the likelihood function is

$$\prod_{i=\mathrm{H,W}}\prod_{j=\mathrm{H,W}}\prod_{k=\mathrm{H,W}}\left\{p\pi^{(1)}_{i|\mathrm{H}}\,\pi^{(2)}_{j|\mathrm{H}}\pi^{(3)}_{k|\mathrm{H}}{}^{-}+(1-p)\pi^{(1)}_{i|\mathrm{W}}\,\pi^{(2)}_{j|\mathrm{W}}\pi^{(3)}_{k|\mathrm{W}}\right\}^{n_{ijk}}.$$

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 degrees of freedom that 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.

Table 2

Examples from cross-classification data generated as expected counts from a sample of 1000 otoliths based on different accuracy rates for identifying hatchery fish ($\pi_{\mathrm{H\,|\,H}}$) and wild fish ($\pi_{\mathrm{W\,|\,W}}$) under different mark proportions (p). The examples used illustrate differences between observed agreement (P_o) and chance-corrected agreement (κ) under different underlying conditions.

A		Reader 2				
		H	W			
Reader 1	H	81	9	90	$\pi_{\mathrm{H\,|\,H}} = 0.9$	$P_o = 0.98$
	W	9	901	910	$\pi_{\mathrm{W\,|\,W}} = 1.0$	$\kappa = 0.89$
	Total	90	910	1000	$p = 0.1$	

B		Reader 2				
		H	W			
Reader 1	H	25	25	50	$\pi_{\mathrm{H\,|\,H}} = 0.5$	$P_o = 0.95$
	W	25	925	950	$\pi_{\mathrm{W\,|\,W}} = 1.0$	$\kappa = 0.47$
	Total	50	950	1000	$p = 0.1$	

C		Reader 2				
		H	W			
Reader 1	H	410	90	500	$\pi_{\mathrm{H\,|\,H}} = 0.9$	$P_o = 0.82$
	W	90	410	500	$\pi_{\mathrm{W\,|\,W}} = 0.9$	$\kappa = 0.64$
	Total	500	500	1000	$p = 0.5$	

D		Reader 2				
		H	W			
Reader 1	H	50	90	140	$\pi_{\mathrm{H\,|\,H}} = 0.9$	$P_o = 0.82$
	W	90	770	860	$\pi_{\mathrm{W\,|\,W}} = 0.9$	$\kappa = 0.25$
	Total	140	860	1000	$p = 0.05$	

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 $\pi^{(k)}_{\mathrm{H|H}}$ and $\pi^{(k)}_{\mathrm{W|W}}$ 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: $\pi^{(1),(2)}_{\mathrm{H|H}}, \pi^{(1),(2)}_{\mathrm{W|W}}$, p_1, and p_2, with $2(2^2 - 1) = 6$ df. Increasing the number of strata increases the degrees of freedom; 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

$$\prod_{g=1}^{S}\prod_{i=\mathrm{H,W}}\prod_{j=\mathrm{H,W}}\left\{p_s\,\pi^{(1)}_{i|\mathrm{H}}\,\pi^{(2)}_{j|\mathrm{H}}+(1-p_g)\,\pi^{(1)}_{i|\mathrm{W}}\,\pi^{(2)}_{j|\mathrm{W}}\right\}^{n_{gij}}.$$

A third way to supply additional information is to take a Bayesian approach (see "Discussion" section). By specifying prior distributions of the model parameters, unique estimates can be obtained (Joseph et al., 1995).

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 is independent given the true state of the otolith. In principle, the latter assumption may be difficult to meet especially if all readers examine the same otolith. The fact that the otolith is not prepared independently by each reader could 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 for the problem of dependence due to preparation is to require independent preparations. This however, requires additional otoliths and with only two otoliths per fish, this would limit the number of readers to two. But in practice, this may not be a large concern. Typically, the second reader has the option to provide additional processing effort to the first otolith or, if needed, to process the second otolith. In almost all cases additional preparation is not done and readers feel they are able to extract sufficient information about the presence or absence of a mark from each other's preparations. In addition, reader accuracy rates obtained by LCM do not appear to vary systematically with the reading order, which also suggests that preparation-induced dependency is not a significant factor.

Dependency associated with variability in the appearance of the mark may be harder to address. A general solution is to model the dependence with additional parameters (e.g. Vacek, 1985; Qu et al., 1996; Yang and Becker, 1997; Qu and Hagdu; 1998; Albert et al., 2001). Modeling dependence requires either more readers or more strata. These modeling approaches are complicated and are currently evolving (see Albert et al., 2001). Alternatively, additional latent classes may be added (Christensen et al., 1992; Formann, 1994), e.g. a third class of otoliths from ambiguous sources.

In the previous discussion concerning three or more readers, we implied that readers were different individuals. This need not be so; what is required are three or more independent readings. If it were possible for the same individual to read the same otolith more than once, independently, then the number of different readers could be reduced. If independence could not be met, 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 $\pi_{W|W} > \pi_{H|W}$, which implies that $\pi_{H|H}$ and $\pi_{W|W}$ are either greater than or less than 0.5) 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 Formulas for estimating κ and its standard error are straightforward (Fleiss, 1981). Estimates can also be obtained from several software packages including PROC FREQ in SAS (SAS Institute, 1989).

Maximizing either of the likelihood functions for the LCMs requires a numerical procedure. 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 (Agresti, 1990). This method is relatively easy to implement in software such as PROC NLIN in SAS (SAS Institute, 1989). Perhaps the most direct and efficient way would be to use LCM software. We are not aware of any routines for LCMs in any major statistical package at present, but several independent LCM packages exist (for a review, see Clogg, 1995; and for an Internet listing see http://ourworld.compuserve.com/homepages/jsuebersax/index.htm).

As with many maximum likelihood problems, where numerical methods must be used, complications can arise. Constraints may at times be needed to ensure that parameter estimates fall in acceptable intervals (e.g. [0,1] for p and [0.5,1] for the π's). 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, the jackknife being much less computationally intensive.

Finally, the Bayesian programs discussed in Joseph et al. (1995) can be found at http://www.epi.mcgill.ca/Joseph/software.html.

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 without knowledge of prior readings. The data, along with pairwise κ estimates and the LCM parameter estimates (using PROC NLIN in SAS; see appendix for code) are presented in Table 3.

These results indicate that the third reader is significantly (α=0.05) less able to correctly identify a hatchery mark when it is present and that there are no significant differences among readers in their ability to detect a wild mark when it is present. These conclusions are readily apparent from the table of results, and although the pairwise κ's are consistent with these results, they are more difficult to interpret. With the variance due to sampling estimated to be (0.7379)(1 – 0.7379)/(570 – 1) = 0.0003399, misclassification error contributes only 0.36% to the total variance.

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 and the LCM estimates are shown in Table 4. These estimates indicate that the readers are not statistically different in their ability to detect hatchery marks, whereas the second reader is better able to distinguish wild marks. With eight parameters and 12 df, there are 4 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. Misclassification error contributes from about 8% to 14% to the total variance in the estimates of the proportion of hatchery stock.

Design considerations

Design of an otolith reading program is complicated by misclassification error. An important consideration is the precision of the estimates, in particular the precision of the estimate of p. Table 5 shows the asymptotic standard error of $\hat{p}$ for various combinations of p, $\pi_{H|H}$, and $\pi_{W|W}$ for the three-reader model with unknown accuracies, and the one-, two-, and three-reader models with accuracies assumed known. Although this table is derived for a sample of 1000 otoliths, the ratio of any two standard errors within the table would be the same for any sample size (assuming the sample size is large enough to approximate the asymptotic conditions). It is evident that misclassification inflates the standard error over the usual binomial case (right-most column). The table also makes clear the increase in the uncertainty of estimating p when the accuracies also have

Table 3

Cross-classification data and results for 570 chum otoliths examined by three readers showing the parameter estimates and standard errors from the latent class model, followed by a comparison of the differences among reader pairs by using kappa and the latent class model (LCM) accuracy rates. The data show that the high agreement among readers as to hatchery and wild classification (e.g. HHH=406 and WWW=135) is reflected in the overall high accuracy rates estimated from the LCM. However the model also shows that reader 3 has a significantly lower accuracy rate in detecting hatchery marks ($\pi^{(3)}_{H|H}$=0.969) than the other readers.

Reading	Count	LCM Parameter	Estimate	SE
HHH	406	$\pi^{(1)}_{H\|H}$	0.998	0.002
HHW	13	$\pi^{(2)}_{H\|H}$	0.998	0.002
HWH	1	$\pi^{(3)}_{H\|H}$	0.969	0.008
WHH	1	$\pi^{(1)}_{W\|W}$	0.958	0.017
HWW	6	$\pi^{(2)}_{W\|W}$	0.986	0.010
WHW	2	$\pi^{(3)}_{W\|W}$	0.957	0.017
WWH	6	p	0.738	0.018
WWW	135			

Reader pairs	κ	SE	Difference in $\pi_{H\|H}$	SE	Difference in $\pi_{W\|W}$	SE
1 and 2	0.954	0.014	0.000	0.004	–0.028	0.020
1 and 3	0.882	0.022	0.029	0.009	0.000	0.024
2 and 3	0.901	0.021	0.029	0.009	0.028	0.020

Table 4

Cross-classification data for 2340 sockeye otoliths examined by two readers and stratified by four fishing districts, showing the estimates of the latent class parameters and their standard errors. Between-reader comparison is based on whether the difference in accuracy estimates are significantly different than zero. The results indicate that the readers were not statistically different in detecting hatchery marks ($\pi_{H|H}$) but were statistically different in detecting wild marks ($\pi_{W|W}$). LCM = latent class model.

	Fishing districts			
	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 parameter	Estimate	SE	Reader difference	SE
$\pi_{H\|H}^{(1)}$	0.980	0.013	0.017	0.025
$\pi_{H\|H}^{(2)}$	0.964	0.021		
$\pi_{W\|W}^{(1)}$	0.984	0.005	–0.013	0.006
$\pi_{W\|W}^{(2)}$	0.997	0.003		
$p_{108–30}$	0.366	0.024		
$p_{108–50}$	0.257	0.020		
$p_{106–41}$	0.096	0.010		
$p_{106–30}$	0.047	0.011		

to be estimated in the three-reader case. For example, if $\pi_{H|H} = \pi_{W|W} = 0.8$ for all three readers, one would have to have almost twice (0.035/.019=1.84) the sample size to estimate a p of about 0.5. Once accuracy estimates for the readers are obtained, dropping one or even two readers may be appropriate, although the assumption must be made that the accuracy rates will be constant for the remainder of the program. Maintaining two readers will allow for that

Table 5

Asymptotic standard errors for the estimated proportion of marked fish, $\hat{p}$, for various combinations of accuracy rates in identifying hatchery fish, $\pi_{\mathrm{H}\mid\mathrm{H}}$, and wild fish, $\pi_{\mathrm{W}\mid\mathrm{W}}$, and mark proportion p, for a sample of 1000 otoliths. Values are reported for the cases where accuracy rates, π, are the same and assumed known for one, two, or three readers, and for the case where π's are estimated for three readers. Table illustrates how misclassification will increase standard errors in the estimate of hatchery proportion.

	$\pi_{\mathrm{H}\mid\mathrm{H}}$	0.8			0.9			1.0		
	$\pi_{\mathrm{W}\mid\mathrm{W}}$	0.8	0.9	1.0	0.8	0.9	1.0	0.8	0.9	1.0
	p									
3 readers	0.1	0.032	0.016	0.011	0.023	0.013	0.010	0.018	0.011	0.009
(π's estimated)	0.3	0.034	0.021	0.017	0.024	0.017	0.015	0.020	0.015	0.014
	0.5	0.035	0.023	0.019	0.023	0.018	0.016	0.019	0.016	0.016
	0.7	0.034	0.024	0.020	0.021	0.017	0.015	0.017	0.015	0.014
	0.9	0.032	0.023	0.018	0.016	0.013	0.011	0.011	0.010	0.009
3 readers	0.1	0.013	0.011	0.010	0.011	0.010	0.009	0.010	0.010	0.009
(π's known)	0.3	0.018	0.016	0.015	0.017	0.015	0.015	0.015	0.015	0.014
	0.5	0.019	0.018	0.016	0.018	0.017	0.016	0.016	0.016	0.016
	0.7	0.018	0.017	0.015	0.016	0.015	0.015	0.015	0.015	0.014
	0.9	0.013	0.011	0.010	0.011	0.010	0.010	0.010	0.009	0.009
2 readers	0.1	0.015	0.013	0.010	0.013	0.011	0.010	0.011	0.010	0.009
(π's known)	0.3	0.020	0.018	0.015	0.018	0.016	0.015	0.015	0.015	0.014
	0.5	0.022	0.019	0.016	0.019	0.018	0.016	0.016	0.016	0.016
	0.7	0.020	0.018	0.015	0.018	0.016	0.015	0.015	0.015	0.014
	0.9	0.015	0.013	0.011	0.013	0.011	0.010	0.010	0.010	0.009
1 reader	0.1	0.023	0.017	0.011	0.020	0.015	0.010	0.018	0.014	0.009
(π's known)	0.3	0.026	0.021	0.017	0.022	0.019	0.016	0.020	0.017	0.014
	0.5	0.026	0.022	0.019	0.022	0.020	0.017	0.019	0.017	0.016
	0.7	0.026	0.022	0.020	0.021	0.019	0.017	0.017	0.016	0.014
	0.9	0.023	0.020	0.018	0.017	0.015	0.014	0.011	0.010	0.009

assumption to be checked because there will now be extra degrees of freedom to assess goodness-of-fit (there are 3 df, but only one parameter, p, needs to be estimated). Estimates of p can still be obtained with one reader, but there can be no check of the assumptions. Also, there can be a significant increase in uncertainty in the estimate in using only one reader.

Discussion

There are numerous classification problems in fisheries that require the judgment of trained individuals. In many of those situations no "gold standard" is available to test those judgments, 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 this 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 to presume that agreement between readings can serve as a proxy for reading accuracy. Agreement indices such as κ 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. Also, the indices in themselves do not provide inferences about the relative skill of different readers in pulling out a particular set of patterns.

Latent class models provide an 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 in 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), although 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, and Krzanowski and Marriott, 1995, for reviews).

We have not discussed the Bayesian approach to these problems in great detail, but we believe it has much to offer in that it can incorporate prior information, either in the form of expert opinion (e.g. Demissie et al., 1998) or in the form of results of earlier analyses (e.g. Viana et al., 1993). Rather than assuming that estimated accuracies are “known,” one can incorporate the uncertainty in the estimates into the prior distributions. In addition, the Bayesian approach does not rely on asymptotic results that may behave poorly with small samples. We have also not assessed the possible bias due to the lack of independence in the readings. When suitable software becomes available, this assumption should be checked.

In our examples above, misclassification error contributed relatively little to the overall uncertainty. In these applications, where estimates of hatchery contribution were used to make management decisions, the accuracy of readings were within an acceptable range. However, the criteria used to establish quality control standards in any program need to be developed in the context of how the information is to be used along with other sources of uncertainty.

In conclusion, we believe that the use of agreement measures in combination with latent class models can contribute significant information about both the proportions of interest and the quality control aspects of an otolith-marking program. Furthermore these approaches could have application to similar areas in fisheries which require judgments that are not free of error.

Acknowledgments

We thank Bob Wilbur for editorial comments and three anonymous reviewers for valuable suggestions.

Literature cited

Agresti, A.
1990. Categorical data analysis. John Wiley, New York, NY, 576 p.

Albert, P. S., L. M. McShane, and J. H. Shih.
2001. Latent class modeling approached for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. Biometrics 57:610–619.

Bartholomew, D. J.
1987. Latent variable models and factor analysis. Oxford Univ. Press, New York, NY, 427 p.

Christensen A. H., T. Gjorup, J. Hilden, C. Fenger, B. Henriksen, M. Vyberg, K. Ostergaard, and B. F. Hansen.
1992. Observer homogeneity in the histologic diagnosis of *Helicobacter pylori*: latent class analysis, kappa coefficient, and repeat frequency. Scand. J. Gastroenterol. 27:933–939.

Clogg, C. C.
1995. Latent class models. Chapter 6 *in* Handbook of statistical modeling for the social and behavioral sciences (G. Arminger, C. C. Clogg, and M. E. Sobel, eds.), p. 311–359. Plenum Press, New York, NY.

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.

Demissie, K., N. White, L. Joseph, and P. Ernst.
1998. Bayesian estimation of asthma prevalence, and comparison of exercise and questionnaire diagnostics in the absence of a gold standard. Ann. Epidemiol. 8:201–208.

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.

Everitt, B. S.
1984. An introduction to latent variable models. Chapman and Hall, London, 107 p.

Fleiss, J. L.
1981. Statistical methods for rates and proportions, 2^{nd} ed. John Wiley, New York, NY, 352 p

Formann, A. K.
1994. Measurement errors in caries diagnosis: some further latent class models. Biometrics 50:865–871.
1996. Latent class analysis in medical research. Stat. Meth. Med. Res. 5:179–211.

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

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

Joseph, L., T. Gyorkos, and L. Coupal.
1995. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am. J. Epidemiol. 141:263–72.

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, 280 p.

McCutcheon, A. L.
1987. Latent class analysis. Sage, Beverly Hills, CA, 96 p.

McLachlan, G. J., and T. Krishnan.
1997. The EM algorithm and extensions. John Wiley, New York, NY, 304 p.

Microsoft Corporation.
1993. Microsoft Excel user’s guide. Microsoft Corporation, Redmond, WA.

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. Dep. 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.
1998. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. J. Am. Stat. Assoc. 93:920–928.

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. Am. J. Epidemiology 107:71–76.

SAS Institute.
1989. SAS/STAT user's guide, version 6, 4th ed. SAS Institute, Cary, NC.

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

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

Viana, M. A. G., V. Ramakrishnan, and P. S. Levy.
1993. Bayesian analysis of prevalence from the results of small screening samples. Commun. Statist. Theory Meth. 22:575–585.

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. Am. Fish. Soc. Symp. 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 three-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 by using PROC NLIN. This computation 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 appropriate intervals. 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 for this application. The residual weighted sum of squares listed by SAS is the chi-squared goodness-of-fit-statistic. The option, OUTEST, outputs point estimates and the the estimated covariance matrix for the parameters. SAS code for the multistrata model used in the second example is also available from the authors.

```
/* SAS Code for estimating 3-reader, 1-stratum model */

data a;
   array x{8} x1-x8;
   input y;
   ntot+y;                                        /* accumulating sample size */
   if _n_=8 then call symput(‘ntot’,ntot);        /* put total into macro var */
   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 outest=est;              /* sigsq=1 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);              /* a is accuracy for H */
   e2=a1*a2*(1-a3)*p+(1-b1)*(1-b2)*b3*(1-p);              /* b is accuracy for W */
   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)*&ntot;
   bounds   0.5<=a1<=1, 0.5<=a2<=1, 0.5<=a3<=1, 0.5<=b1<=1, 0.5<=b2<=1,
            0.5<=b3<=1, 0<=p<=1;
   _weight_=1/model.y;
run;
```