NOAA Technical Memorandum NMFS

SEPTEMBER 2013

KLAMATH-TRINITY BASIN FALL RUN CHINOOK SALMON SCALE AGE ANALYSIS EVALUATION

William H. Satterthwaite
Michael R. O'Farrell
Michael S. Mohr

National Marine Fisheries Service
Southwest Fisheries Science Center
110 Shaffer Road
Santa Cruz, CA 95060

NOAA-TM-NMFS-SWFSC-522

U. S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Southwest Fisheries Science Center

The National Oceanic and Atmospheric Administration (NOAA), organized in 1970, has evolved into an agency that establishes national policies and manages and conserves our oceanic, coastal, and atmospheric resources. An organizational element within NOAA, the Office of Fisheries, is responsible for fisheries policy and the direction of the National Marine Fisheries Service (NMFS).

In addition to its formal publications, the NMFS uses the NOAA Technical Memorandum series to issue informal scientific and technical publications when complete formal review and editorial processing are not appropriate or feasible. Documents within this series, however, reflect sound professional work and may be referenced in the formal scientific and technical literature.

NOAA Technical Memorandum NMFS

This TM series is used for documentation and timely communication of preliminary results, interim reports, or special purpose information. The TMs have not received complete formal review, editorial control, or detailed editing.

SEPTEMBER 2013

KLAMATH-TRINITY BASIN FALL RUN CHINOOK SALMON SCALE AGE ANALYSIS EVALUATION

William H. Satterthwaite
Michael R. O'Farrell
Michael S. Mohr

National Marine Fisheries Service
Southwest Fisheries Science Center
110 Shaffer Road
Santa Cruz, CA 95060

NOAA-TM-NMFS-SWFSC-522

U. S. DEPARTMENT OF COMMERCE
Penny Pritzker, Secretary of Commerce
National Oceanic and Atmospheric Administration
Dr. Kathryn D. Sullivan, Acting Administrator
National Marine Fisheries Service
Samuel D. Rauch III, Assistant Administrator for Fisheries

1 Executive Summary

An evaluation of appropriate levels of sampling needed to develop age composition estimates was made by appealing to multinomial sampling theory. Random sample sizes of at least 600–800 read scales are recommended for each stratum where age composition is of interest. This will suffice to assure confidence intervals on proportions in each age class no wider than 0.05, provide acceptable relative precision for reasonably small proportions, and have a high probability of detecting rare age classes when present. Sample sizes have tended to be larger than the recommended levels at Iron Gate Hatchery, Trinity River Hatchery, the Yurok tribal fishery, and in some cases, the total river recreational fishery and the Hoopa Valley tribal fishery. Spawners in some sectors (e.g., Scott River, Bogus Creek, Willow Creek Weir) have generally been sampled adequately for scales, while others (e.g., Salmon River, Shasta River, lower Trinity mainstem) have frequently been undersampled.

The potential to use coded-wire tag (CWT) recovery data alone to determine age composition of hatchery returns was considered. Assuming that all fish returning to hatcheries are of hatchery origin, and thus marked and tagged at an approximately 25 percent rate, the number of CWT recoveries would generally provide adequate sample sizes for age composition determination. However, using CWT data alone would only be appropriate if the contribution of natural-origin fish to hatchery returns are trivial, since CWT'd fish provide no information on the age composition of natural-origin fish. Until an evaluation of the natural-origin fish contribution to hatchery returns is undertaken, and results demonstrate that natural origin fish make up a trivial fraction of hatchery returns, the current practice of scale collection and reading from a randomly drawn sample should continue.

Scale aging validation matrices are constructed to correct for biases introduced by scale reading errors. The persistent problem of few (or zero) known age-5 fish present in annually constructed scale aging validation matrices was considered. Sample sizes for known age-5 fish used in validation matrices have almost always been inadequate (i.e., less than approximately 20 scales per true age). Establishing a library of known age-5 scales, archived across years, would allow for supplementing sample sizes when constructing validation matrices. Errors introduced by combining known age-5 fish data across years are likely to be small relative to the sampling error associated with the current reality of small annual age-5 sample sizes.

The practice of using multiple readers to age all scales is costly relative to a system where separate sets of scales are read by separate readers. It is unclear whether the benefit of using multiple readers, in terms of accuracy, is sufficient to justify the cost. A potential approach to evaluate the difference in scale reading accuracy between the multiple reader versus single reader approach is described.

Reader-specific validation matrices may be necessary if a change is made from the status quo procedure of using multiple readers to age a set of scales. If single readers are used to age separate sets of scales, the construction of reader-specific validation matrices is recommended.

Consideration was given to the appropriate spatial and temporal scales for validation matrices. Past estimates suggest that scale aging accuracy is lower when scales are taken from carcasses relative to harvested fish sampled earlier in the run and lower in the river, likely owing to differing degrees of scale resorption. The suitability of applying the same validation matrix to different strata (e.g., harvest versus carcass surveys) should be evaluated annually. If substantial differences in scale-age error rates exist between survey strata, separate validation matrices should be considered. To ensure adequate sample sizes for stratum-specific validation matrices, pooling scales across years may be warranted.

2 Introduction

At the request of Ernest Clarke (US Fish and Wildlife Service, Trinity River Restoration Program, Science Program Coordinator), we initiated an evaluation of the methodology used to generate estimates of the age composition of Klamath-Trinity Basin fall Chinook salmon. Materials reviewed included a fiscal year 2014 proposal for scale age analysis in the basin (Logan et al. 2013), the most recent report from the Klamath River Technical Team (KRTT) on age composition estimation in the Klamath-Trinity Basin (KRTT 2013), equivalent reports from earlier years, a compilation of Excel spreadsheets containing the tables from earlier such reports for run years 2003–2012 (KRTT file archive maintained by Mohr and O'Farrell), and a recent review of the adult salmonid monitoring programs of the Trinity River Restoration Program (Bradford and Hankin 2012).

Our charge was to provide 1) an "assessment/evaluation of the level of sampling needed to develop the post-season age composition of the Klamath-Basin run" and 2) input on "if hatchery CWT recoveries, with the 25% ad-clip/CWT marking, could be used for hatchery age composition rather than collecting scale samples at the hatchery", along with "any other recommendations to make this project more efficient while still providing the necessary data for stock assessment and harvest management activities". Additional topics of interest we identified early in our review were 3) how to deal with the rarity of known age-5 fish when constructing scale aging validation matrices, 4) the need for multiple readers to analyze the same set of scales, 5) the need for reader-specific bias correction, and 6) the appropriate temporal and spatial scale for the construction and application of scale aging validation matrices.

All recommendations provided herein are based upon statistical power analyses and our understanding of the sampling programs in the basin. We are well aware that samplers face on-the-ground challenges that we are either unaware of or not able to fully appreciate, and that there may be reasons why sampling rates in some strata must be lower than is recommended here, or that additional information needs may merit more intensive sampling in some strata than is recommended here.

3 Level of sampling needed

The sample size requirements for any sampling program depend on the specific goals and questions the sampling is intended to address, as well as situation-specific constraints. At some level, decisions must be made about the level of precision desired, and these decisions must at times be somewhat arbitrary. In the development of scale-aging plans for California's Central Valley, sample size advice (Mohr, unpublished) was provided on the basis of a desire that 95% confidence intervals on the proportions estimated for each age-class contributing substantially to the river returns be no wider than 0.05. This seems like a reasonable standard to apply to the Klamath-Trinity Basin as well, although we note that proportions of age-2 (0.07 in 2012) and age-5 (0.01 in 2012) may be small enough that an error of 0.05 is large in relative terms.

Bromaghin (1993) describes sample size requirements for estimating proportions under a multinomial random sampling model, when more than two proportions are estimated simultaneously from the same population using simple random sampling with replacement. In practice, the sampling for scales within a stratum is further stratified temporally or conducted systematically over the course of the run and it is done without replacement, but the simple random sampling with replacement model serves as a good approximation when the sampled fraction is small, and will be "conservative" (from a statistical rather than budgetary perspective) when the sampled fraction is large. Under this sampling model, a sample size of 618 is required to assure simultaneous 95% coverage of confidence intervals no wider than 0.05 on four proportions (i.e., age-2 through age-5), assuming no uncorrected aging error. Note that simultaneous coverage for all four intervals requires that individual intervals be narrower, requiring almost 99% coverage for individual intervals when four proportions are considered. For 95% confidence intervals on individual proportions to meet the same standard would require a sample size of 381 in this case. Figure 1 shows how the required sample size varies as a function of the desired confidence interval width. Note that much larger sample sizes are needed to reduce the confidence interval width below 0.02.

As noted earlier, a constant confidence interval width of 0.05 may be wide relative to the proportions for rarer age classes (i.e., age-2 and especially age-5), but at the same time the calculations

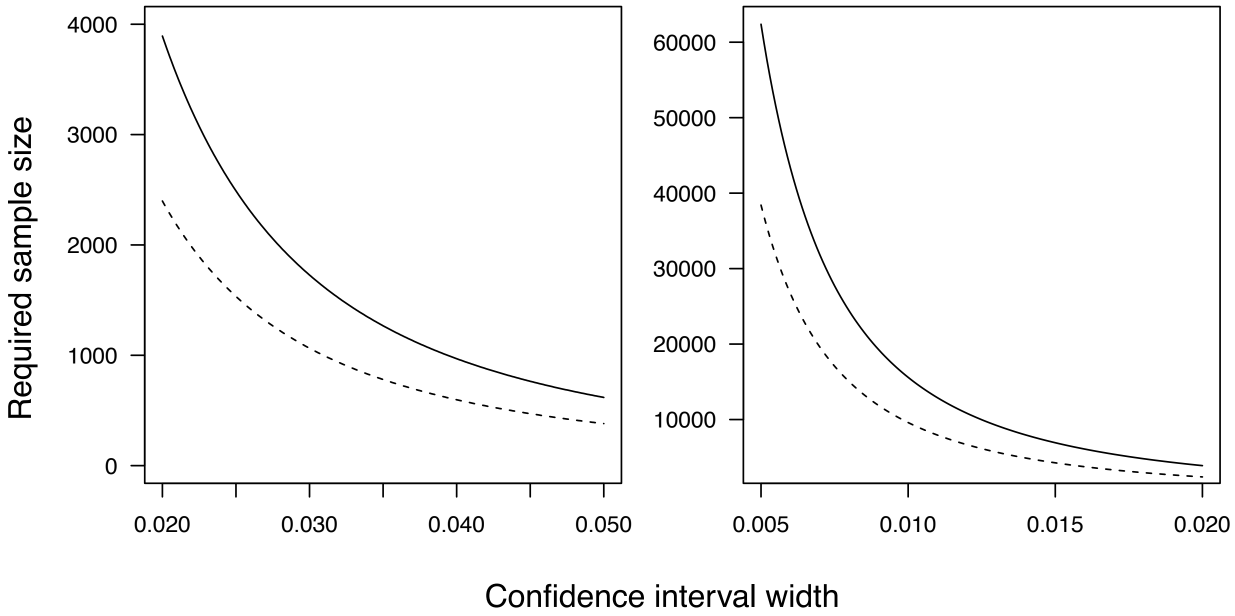

Figure 1. Required sample size given multinomial random sampling model with four age classes and desired 95% confidence interval width on proportions as specified on the x-axis. The solid line indicates the sample size required such that the confidence interval coverage is for all proportions simultaneously, while the dashed line is for a single proportion.

presented in Figure 1 assume a worst case scenario in which the proportion of one age-class is equal to 0.5 (i.e., confidence intervals on smaller proportions would be narrower given the same sample size if all proportions were far away from 0.5). To explore the sample size requirements necessary to satisfy a proportional error standard, we solved equation 15 of Bromaghin (1993) substituting various proportions for π_i and using a specified multiple of those proportions for d_i. The results are plotted in Figure 2. Very small proportions may be impossible to estimate with precision, however it might be desirable to estimate proportions ≥ 0.03 with a confidence interval no wider than half the estimated proportion, roughly equivalent to ±25% (but note that confidence intervals are not necessarily symmetric, especially for small proportions). This would require a sample size of 846 to achieve 95% coverage for four age classes simultaneously. However the required sample size decreases to 521 if coverage is evaluated for only a single confidence interval (along with 792 for 0.02 and 1606 for 0.01), which may be appropriate when evaluating proportional errors if only one age class is expected to be particularly rare. For proportions smaller than 0.02, the required sample size scales almost linearly with the inverse of the proportion, i.e. estimating a proportion half as small requires twice the sample size (Figure 2).

A final consideration is the ability to detect a rare age-class, given that it is present. This is a particular concern because the Kimura and Chikuni (1987) algorithm, often used by the KRTT to correct for scale reader bias, will always converge to a solution of zero age-5 individuals if there are zero read age-5 fish in the sample, even if the validation matrix states that age-5 fish are highly likely to be misread as age-4 fish. Allen-Moran et al. (2013) describe minimum sample sizes needed to detect rare stocks/ages. For example, there is a 99.9% chance of detecting a stock/age-class with a proportion as small as 0.01 if the sample size is 688 (but note that the proportion read as age-5 may be smaller than the proportion that is age-5). Halving the target proportion requires approximately doubling the sample size.

The considerations above suggest a sample size on the order of 600–800 would be adequate to estimate the age composition of a single population using a simple random sample, although errors in the smallest estimated proportions could be large on a relative scale. The Klamath-Trinity Basin is of course subdivided and a stratified sampling scheme is employed. To estimate the age composition for the basin as a whole, a basin-wide total sample size on the same order would suffice if the stratified sampling scheme was fully representative and proportional (i.e., each stratum received a fraction of the total sampling effort in proportion to its contribution to the total river return) both spatially and temporally. However, such a scheme would be exceedingly difficult if not impossible to implement, and would not yield stratum-specific age composition estimates which are used to derive a number of quantities important to assessment (e.g, fishery-age-specific harvest selectivity parameters, and area-age-specific spawning escapements), so we assumed that stratum-specific estimates are desired, with strata defined at the level of subdivision reported in, e.g., KRTT (2013, Table 5).

Each stratum should have sample sizes on the order of 600–800 to yield acceptable precision in stratum-specific estimates. The exception to this rule would be strata for which the stratum-specific total number of fish is so small that a sample size this large requires sampling a large fraction of the total number, or for which the total number is less than 600 fish. For small strata, the sampling with

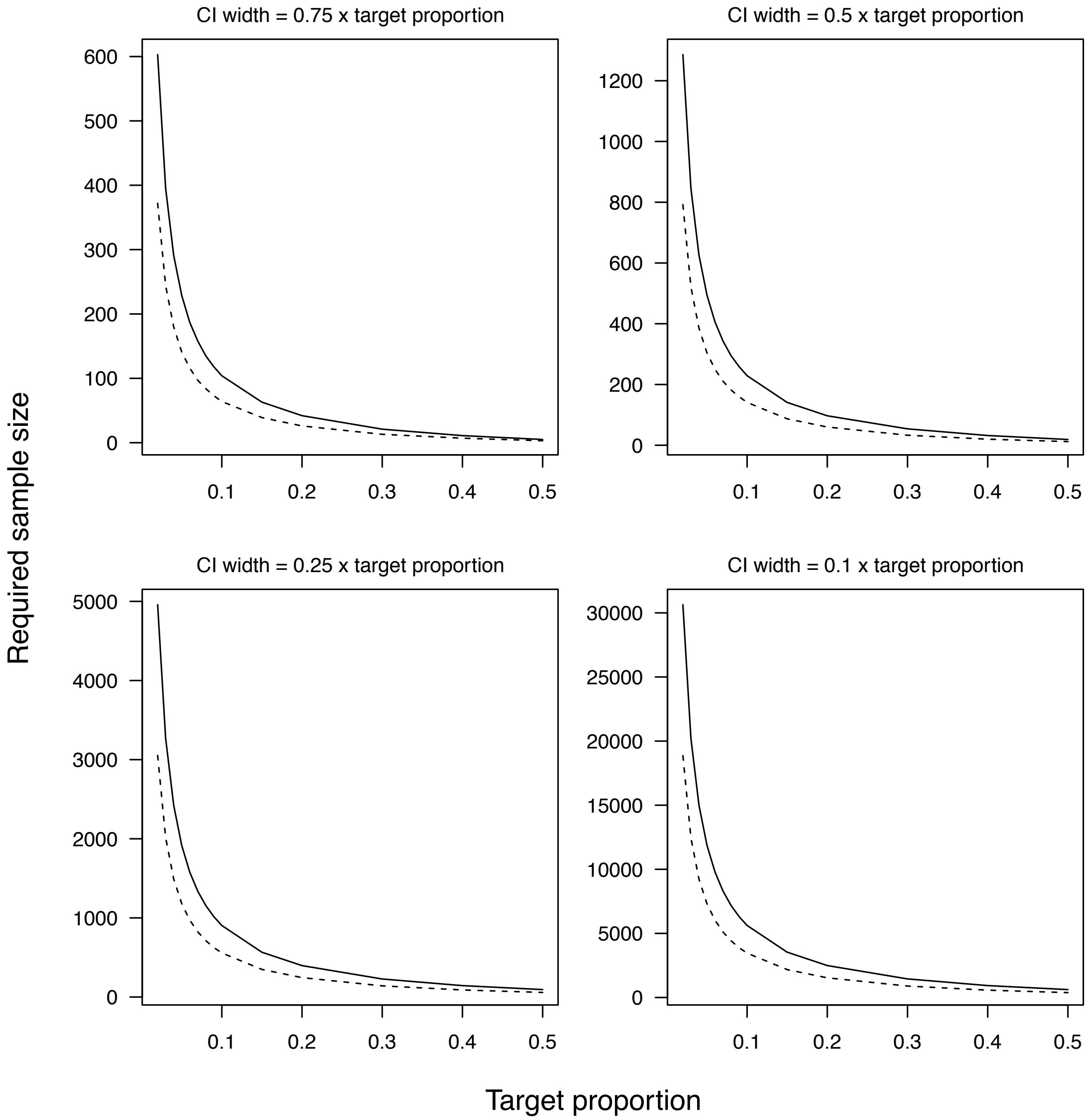

Figure 2. Required sample size given multinomial random sampling model with four age classes in which the desired 95% confidence interval width scales with target proportions as specified in panel titles, with target proportions as specified on the x-axis. The solid line indicates the sample size required such that the confidence interval coverage is for all proportions simultaneously, while the dashed line is for a single proportion.

Table 1. Number of scale-aged fish in each year-stratum, 2003–2012.

	Salmon	Scott	Shasta	Bogus	Klamath	Upper	Blue	Willow	Lower	Lower
	River	River	River	Creek	River	Klamath	Creek	Creek	Trinity	Trinity
Year	Carcass	Carcass	Carcass	Weir	Carcass	Tribs	Snorkel	Weir	Carcass	Tribs
2012	398	1,475	610	1,086	843	95	19	1,213	19	5
2011	423	1,751	280	1,099	590	144	134	805	49	100
2010	137	496	25	856	444	35	5	629	9	21
2009	224	757	372	1,052	1,154	215	70	574	45	18
2008	297	1,107	204	689	888	0	47	920	0	5
2007	133	1,299	133	808	1,102	0	19	396	0	1
2006	159	1,162	487	637	531	0	0	444	29	10
2005	54	151	404	986	214	5	0	699	25	1
2004	61	107	284	880	502	0	0	1,040	67	78
2003	792	1,209	380	326	485	55	0	570	48	54

replacement approximation no longer applies, and the required sample size could be determined using a model based on the multivariate hypergeometric sampling model. However, results from Allen-Moran et al. (2013) comparing the hypergeometric and binomial sampling models suggest that substantially smaller sample sizes will only suffice when the total number of fish is so low that a large fraction (e.g., 50% or more) would be sampled. Thus, high sampling rates would be required for acceptable precision in estimates made for small strata.

The number of unknown-age scales read at Iron Gate Hatchery has ranged from 969–1,541 between 2003 and 2012 while numbers for Trinity River Hatchery have ranged from 797–3,680. Including known-age (CWT) fish, the total sample sizes are even larger. Thus scale reading rates for hatchery spawners appear to be more than adequate and could be reduced in years of high returns. See also Section 4 for a discussion of whether the hatchery age composition could be properly estimated from the CWT, known-age, recovery data alone.

Table 1 reports the year-stratum-specific sample sizes (both tagged and untagged fish) for natural area spawners and the Willow Creek Weir, 2003–2012, and Table 2 reports the corresponding sampling fractions. In some years, the Scott River, Bogus Creek, and Klamath River Mainstem sample sizes are slightly larger than required, but in general sample sizes for these strata seem sufficient while not excessive. Recent sample sizes at the Willow Creek Weir have been adequate, although sample sizes were low and represented small proportions of the run in 2006–2007. Sam-

Table 2. Sampling fraction of scale-aged fish in each year-stratum, 2003–2012.

	Salmon	Scott	Shasta	Bogus	Klamath	Upper	Blue	Willow	Lower	Lower
	River	River	River	Creek	River	Klamath	Creek	Creek	Trinity	Trinity
Year	Carcass	Carcass	Carcass	Weir	Carcass	Tribs	Snorkel	Weir	Carcass	Tribs
2012	.09	.16	.02	.09	.05	.02	.02	.02	.03	.01
2011	.08	.32	.02	.21	.08	.02	.09	.01	.05	.16
2010	.05	.20	.02	.25	.11	.02	.01	.02	.19	.10
2009	.08	.34	.06	.18	.14	.07	.07	.03	.06	.07
2008	.12	.24	.03	.15	.13	.00	.09	.05	.00	.01
2007	.09	.29	.07	.17	.16	.00	.04	.01	.00	.00
2006	.08	.23	.22	.15	.10	.00	.00	.02	.15	.05
2005	.11	.20	.20	.18	.05	.01	.00	.05	.14	.01
2004	.10	.23	.30	.23	.10	.00	.00	.07	.07	.23
2003	.23	.10	.09	.02	.03	.03	.00	.02	.17	.08

ple sizes on the Salmon River, Shasta River, Upper Klamath River Tributaries, Blue Creek, and Lower Trinity River Mainstem and Tributaries have often been inadequate for stratum-specific estimates (with the possible exception of the Shasta River in 2012). While logistic constraints may limit sample sizes in some of these strata, and large sample sizes are impossible when the total number of fish are low (e.g., Lower Trinity River), these strata do require increased sampling intensity to yield acceptable precision of stratum-specific estimates.

Typically, 15–40% of the fisheries harvest is sampled (Logan et al. 2013) with all sampled scales read. Total recreational harvest (primarily from the Klamath River) between 2003–2012 ranged form 2,615 to 17,432 with realized sample sizes varying from 610–2,355. Thus, sample sizes for the recreational fishery have typically been adequate, and while some years the sample size may have been larger than required, it appears that the sampling rate is appropriately scaled back in years of high harvest (i.e., sample sizes varied only approximately 4-fold while harvest varied almost 7-fold). The combined Yurok and Hoopa Valley tribal harvest is typically larger than the recreational harvest and sampled at a lower rate, usually resulting in adequate sample sizes. Sample sizes for the combined harvest varied from 2,896 (2005) to 5,725. The number of scales read in the Yurok tribal harvest on the Klamath River varied from 1,128 (2006) to 2,916 below the Highway 101 bridge and from 853 (2007) to 2,252 above the Highway 101 bridge. The number of scales read in the Hoopa Valley tribal harvest on the Trinity River ranged from 373 (2004) to

1,887, and except for 2004 and 2005 (495), at least 640 scales were read in each of the remaining years between 2003 and 2012. Thus, the number of scales read from the harvest of these two tribes has typically been more than adequate, and especially for the Yurok tribal fishery could be reduced in years of high harvest while still achieving the standards for precision outlined earlier.

4 Potential to estimate hatchery age composition from CWT data only

From 2003–2012, spawner counts ranged from 11,231–40,015 at Iron Gate Hatchery and 4,549–30,391 at Trinity River Hatchery. Currently, approximately 25% of the production at these two hatcheries is adipose fin clipped and coded-wire tagged. Therefore, if all spawners counted at the hatchery were of hatchery origin, we would expect that 25% would contain a CWT, and thus there is the potential to obtain known-age fish sample sizes well over 1,000 based on the CWT recoveries alone at the hatcheries. From the perspective of sample size, this would be adequate to estimate the age composition of the fish entering the hatcheries. However, there are two important additional considerations that should be addressed before doing so.

First, not all fish entering the hatchery are hatchery-origin fish. It is not reasonable to assume that the age structure of hatchery- and natural-origin fish entering the hatcheries is identical given the known life history and maturation schedule differences between fingerling and yearling releases, as well as the potential for interannual variation in the factors determining the year class strength of natural-origin fish. Therefore, the contribution rate of hatchery- versus natural-origin fish, as well as its variability over time, would need to be evaluated to address this question. If natural-origin fish were found to make up a nontrivial fraction of fish entering the hatchery, the overall age structure could not be properly estimated from the CWT'd fish alone. Rather, it would appear necessary to continue the current practice of drawing a random sample of fish, whether marked with an adipose fin clip or not, of adequate size to estimate the age composition of the composite population from which the sample was drawn. It would then be possible to use scale-read ages from only the subset of fish without CWT, correct the age structure of those scale-read fish using the usual bias correction method, and then pool the bias-corrected number of scale-aged fish in each age class with the number of CWT-aged fish in each age class to generate an estimate of the overall age composition of the fish entering the hatchery. In this scenario, some scales from CWT'd fish likely would need to be read to construct the validation matrix (see Sections 5 for sample size suggestions and 8 for considerations regarding stratum-specific validation matrices), but reading the scales from all or even most the CWT'd fish may not be strictly necessary for the purposes of estimating the age composition. However, it is possible that a small number of ad-clipped fish may lack a CWT due to tag loss or marking errors.

Second, it might be important to read scales from some minimum number of known-age fish in the hatcheries in order to obtain adequate sample sizes to construct the scale aging validation matrix/matrices. Records of known age-5 fish are particularly sparse in previous years' validation matrices, but sample sizes for age-2 fish have also sometimes been low, especially for the Trinity River. Age-5 spawners are rare at the hatcheries and thus hatcheries may yield few age-5 fish for validation purposes. Age-2 fish are also somewhat underrepresented at hatcheries, but hatcheries may be more effective at boosting the sample size for known age-2 fish. However before advocating increased collection of fish from any particular sector for constructing validation matrices, attention must be given to the appropriate spatial and temporal scale for developing and applying validation matrices, as described in Section 8.

5 Age-5 fish in validation matrices

Scale aging validation matrices are used to correct for biases in scale reader assigned ages based on the age-specific misassignment probabilities (Kimura and Chikuni 1987) as estimated from the reading of scales of known-age CWT'd fish. Table 3 provides example validation matrices (discussed further in Section 8) that illustrate how these matrices are constructed. The number of known age-5 fish with scales read in any particular year is typically low (less than 10 in 5/10 years

Table 3. Scale aging validation matrices used for reader bias correction within the Klamath River proper in 2005 (KRTAT 2006), calculated separately for scales collected from carcasses (panels a and c) versus scales collected from the harvest (panels b and d). Panels a and b show the raw number of scales from fish of a given known-age which were assigned to the various possible read-ages, while panels c and d report the derived set of probabilities that a fish of a given known-age is assigned to the various possible read-ages.

(a) Carcass numbers.

	Known-age			
Read-age	2	3	4	5
2	9	9	1	0
3	0	318	36	0
4	0	15	260	23
5	0	1	0	9
Total	9	343	297	32

(b) Harvest numbers.

	Known-age			
Read-age	2	3	4	5
2	10	1	0	0
3	2	121	3	1
4	0	13	92	1
5	0	0	2	6
Total	12	135	97	8

(c) Carcass probabilities.

	Known-age			
Read-age	2	3	4	5
2	1.000	0.026	0.003	0.000
3	0.000	0.927	0.121	0.000
4	0.000	0.044	0.875	0.719
5	0.000	0.003	0.000	0.281
Total	1.000	1.000	1.000	1.000

(d) Harvest probabilities.

	Known-age			
Read-age	2	3	4	5
2	0.833	0.007	0.000	0.000
3	0.167	0.896	0.031	0.125
4	0.000	0.096	0.948	0.125
5	0.000	0.000	0.021	0.750
Total	1.000	1.000	1.000	1.000

in the Klamath River and 9/10 years in the Trinity River 2003–2012). In some years no scales from known age-5 fish are sampled, and it is simply assumed in constructing the validation matrix that all known age-5 fish are read as age-5 fish (i.e., the age-5 column of the validation matrix is all zeroes except for a one in the age-5 row). This treatment of zeroes is clearly *ad hoc*, and estimating probabilities of correct reads from small samples is unsatisfactory both because of the restriction to discrete possible values (e.g., if only two known-age scales are read, the only possible estimates are 100% correct, 50% correct, or 0% correct) and the large amount of sampling uncertainty associated with small samples (e.g., with 1/2 correct, the standard error (SE) on the estimated probability is 0.35, with ± 2 SE yielding a confidence interval of -0.2 to 1.2). A sample size of at least 20 scales per true age is required for probabilities to be estimated in increments as small as 0.05, and also yields an SE near 0.05 for large or small probabilities (i.e., the SE associated with 19/20 is

0.049, although the SE increases for proportions closer to 0.5, with a maximum of 0.11 for 10/20). Although sufficient sample sizes may generally be available for ages 2–4, it seems unlikely that sample sizes this large (or ideally larger) will be available for known age-5 fish in most years, even with increased sampling effort targeted toward larger or more downstream fish (see also the spatial issues raised in Section 8).

An alternative approach employed in California's Central Valley (B. Kormos, CDFW, personal communication) is the establishment of an archive library of known age-5 scales across years. In the Central Valley case, digital images of scales are obtained and archived, although mounted slides could be similarly preserved and cataloged. Then if the number of known age-5 scales available in a particular year is inadequate, scales (or scale images) could be randomly drawn from the archive and pooled with the current year's samples until an adequate sample size is obtained. Alternately, since in most years the number of known age-5 scales read will be very small, scales from all fish in the archive (along with the current year's) might be used to construct the current year's validation matrix unless the estimated read probabilities differ significantly between the current year's scales and the collection of scales to date.

A potentially more efficient alternative to re-reading all scales each year would be to keep a running tally of the ages read from true age-5 scales across years, and using this tally to augment the validation matrix rather than re-reading the archived scales each year. This approach would not be able to capture the effects of changing readers (or of changes in a single reader's performance), but if the same readers are used for many years and perform consistently, or if differences among readers are small relative to the statistical power available (see Section 8), this may be a small loss.

Neither a running tally nor re-reading the archived scales will be able to capture differences in scale misreading rates caused by biological changes across years (e.g., scales from years with anomalously slow growth may be more difficult to read correctly). However, especially for age-5 fish, the practically achievable sample sizes are likely not adequate to reliably identify the relatively small changes across years that seem apparent in the misread rates for the age-3 and age-4 fish (see Section 8).

6 Need for multiple readers

Currently, two readers independently assign an age to each scale, and if they disagree on the age assigned to a particular fish they confer and reach consensus. The validation matrix is then constructed based on these consensus ages assigned to known-age fish, and used to adjust age proportions among read-age fish. Presumably, this two-reader method leads to increased accuracy in assigned ages and a more precisely estimated validation matrix. However, this comes at a cost since the same amount of reader time could be used to read twice as many scales if effort was not duplicated.

Our review of the literature did not reveal convincing demonstrations one way or the other regarding the utility of a second reader. Flain and Glova (1998) reported that "fish aging by one experienced reader was more precise and more accurate than by combined efforts of several readers", but this was based on a comparison of the accuracy of reads made by a single experienced reader aging New Zealand Chinook salmon versus the reported accuracy in the literature of multiple readers in a variety of other systems, with no direct comparisons of single versus multiple readers in the same system.

In part, the potential benefits of a second reader will depend on how often the two readers disagree, and how often the two-reader system is effectively equivalent to a single-reader system where one reader consistently defers to the judgment of the other. If the currently available records allow, we suggest performing an evaluation of the historical data to see how often the consensus age was originally assigned by one reader versus another in this system, as well as quantifying how often initial disagreements occurred and how often re-assessment of a single reader's ages would be warranted. An even more useful analysis, if possible, would be to construct independent validation matrices for each reader and apply these independent validation matrices to the read ages assigned to each known-age scale by each reader. In addition, the combined-reader validation matrix should be applied to the consensus ages read for the known age scales. Then, a comparison could be made between the actual age composition of the known-age scales and the estimates that would have been generated based on the first reader alone, the second reader alone, and the consensus ages.

Repeated across years and sets of readers, this would provide direct evidence for or against the benefits of the multiple reader approach, and allow quantifying the benefits such that an informed judgment could be made as to whether the benefits justified the increased cost. If the currently available data do not allow for such a comparison, data recording and archiving procedures should be revised to allow for it in the future, such that for every scale read, the data from that scale (including its corresponding CWT, if applicable) are uniquely identifiable and a record is available of the ages originally assigned by each reader as well as the consensus age assigned to that scale.

7 Need for reader-specific validation matrices

In a two-reader, consensus-based approach, it would seem like the only possible approach is one where the validation matrix is also based on the consensus of the same set of multiple readers, since the Kimura and Chikuni (1987) approach updates proportions rather than individual assignments, and so there would be no way to correct and then combine individual reader results. If multiple readers were used, but to age different sets of scales to maximize the total number of fish aged per reader time spent, it would seem like there would be little cost to constructing reader-specific validation matrices so long as each reader read a sufficient number of known-age scales, and this would be the recommended approach. This might require supplying each reader with supplemental known age-5 (and possibly age-2) scales to assure adequate reader-specific sample sizes. However, should this prove cumbersome, it appears that reader-specific variability in error rates may be small relative to the amount of sampling error likely given practical sample sizes (see Section 8), such that a combined validation matrix may sacrifice little accuracy. Detailed comparisons of individual reader performance would be needed to assess this rigorously.

8 Appropriate temporal & spatial scale for validation matrices

Figure 3 depicts annual variation in the proportion of fish of different known ages that were scale-aged correctly. Annual variation in the estimated proportion aged correctly reflects the com-

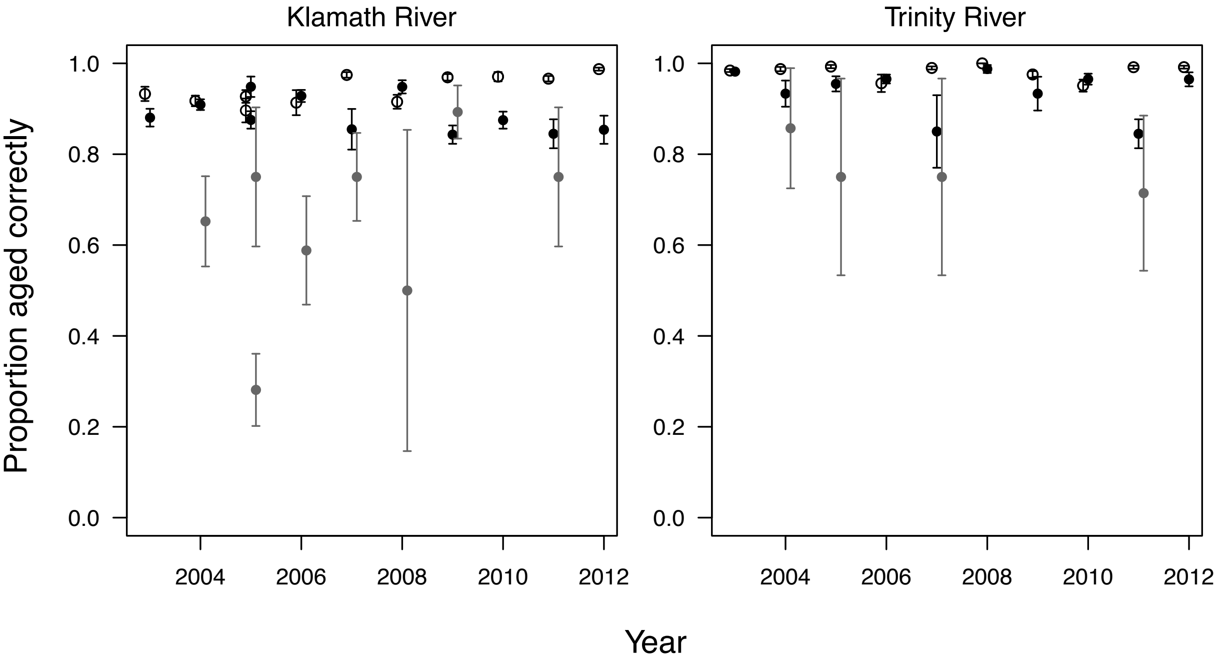

Figure 3. Proportion of scales from known age-3 (white circles), age-4 (black circles), and age-5 (grey circles) fish correctly assigned as reported in the 2003–2012 validation matrices. Error bars are ±1 SE. Age-5 results are only plotted for years in which at least two known age-5 scales were read, in addition the age-5 2003 Klamath River and 2006 and 2009 Trinity River results (2/2, 2/2 and 3/3 correct, respectively) are not plotted since a SE could not be calculated. Note that in 2005 separate estimates were made for the Klamath River harvest versus carcass data, and the accuracy was higher for age-3 in the carcass data and higher for age-4 and age-5 in the harvest data.

bined effects of variation in scale characteristics and readability due to biological differences (e.g., growth rates) across years, the effects of any changes in readers through time, and sampling error associated with finite sample sizes available each year. Note however that the years analyzed (2003–2012) do not include years with ocean conditions associated with anomalously poor growth in California salmon populations (Satterthwaite et al. 2012). For age-3, the probability of assigning correctly is typically high and sample sizes are typically large, resulting in precise estimates that vary little in most years, while those years appearing most different from the rest also have smaller sample sizes and increased uncertainty. For age-4, the rates of correct assignment are lower, sample sizes smaller, and uncertainty higher, but again few years stand out as clearly different from the rest given the uncertainty. Error rates for age-5 appear higher, along with much higher uncertainty due to small sample sizes. Together, these results suggest that annual variation in scale reading error may be small relative to the precision that typical sample sizes make possible, and attempting to quantify annual variability in age-5 error rates may be particularly difficult. Thus, we reiterate our Section 4 suggestion to consider pooling age-5 results across years or establishing an archive library of age-5 (and possibly age-2, although error rates are typically much lower) scales to supplement the available sample size when generating new validation matrices. For other ages, adequate year-specific data should generally be available and used when possible, bearing in mind the lack of "poor growth years" in the dataset used to quantify variation in the error rates to date.

We noted that in 2005, separate validation matrices were constructed (KRTAT 2006) for scales taken from carcasses versus harvest in the Klamath River proper (Table 3). Only 9/32 known age-5 carcasses were aged correctly, compared to 6/8 known age-5 scales sampled from the harvest. This is consistent with the concern raised by Logan et al. (2013) regarding "progressive resorbtion of last annulus on scales collected in terminal spawning areas, as contrasted to scales read by the same reader that may have been collected early after river entry". It thus may be inappropriate to apply the same validation matrix to scales collected in the harvest, scales collected in carcass surveys, and scales collected from live fish recovered at hatcheries or weirs which may be intermediate between the two. The criteria used by the KRTAT to assess the "significance" of the observed difference in the validation matrices was not described in their report (KRTAT 2006), and the potential for significant differences among strata is not discussed in their reports from other years. Stratum-specific results are reported for the Trinity River (but not the Klamath River) in Appendix G of the annual KRTT reports, but the realized sample sizes for age-5 fish are inadequate to allow for within-year comparisons. Pooling all years (2001–2012) together, 8/9 known age-5 fish were correctly aged from scales collected in the Hoopa Valley tribal harvest (89%), while 26/38 (68%) were aged correctly from the Trinity River Hatchery (sample sizes for other strata were clearly inadequate even after pooling across years). Although this difference is not statistically significant ($p = 0.41$, Fisher's exact test), the statistical power of this test is low (even after pooling across years) and the observed difference is consistent with expectations based on scale resorption and the 2005 Klamath River results. While sample sizes of age-2 and age-5 fish are likely inadequate for stratum-specific validation matrices in most years, pooling scales across years for these ages might allow for stratum-specific validation matrices. We recommend further comparison of harvest versus carcass versus hatchery scale aging error rates to determine whether the 2005 results are typical.

9 Acknowledgments

We thank Shanae Allen-Moran and Allen Grover for their helpful reviews of an earlier draft of this paper.

References

Allen-Moran, S. D., W. H. Satterthwaite, and M. S. Mohr (2013). Sample size recommendations for estimating stock composition using genetic stock idenfification (GSI). *U. S. Department of Commerce, NOAA Technical Memorandum NOAA-TM-NMFS-SWFSC-513*.

Bradford, M. and D. Hankin (2012). Trinity River Restoration Program adult salmonid monitoring evaluation. A review conducted for the Trinity River Restoration Program, 1313 South Main Street, Weaverville, California. March 26, 2012. 47 pgs., http://odp.trrp.net/FileDatabase/Documents/TRRP%20%282013%29%20TRRP%202012%20Annual%20Report1.pdf.

Bromaghin, J. F. (1993). Sample size determination for interval estimation of multinomial probabilities. *The American Statistician 47*(3), 203–206.

Flain, M. and G. J. Glova (1998). A test of the reliability of otolith and scale readings of Chinook salmon (*Oncorhynchus tshawytscha*). *New Zealand Journal of Marine and Freshwater Research 22*(4), 497–500.

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

KRTAT (Klamath River Technical Advisory Team) (2006). Klamath River fall Chinook age-specific escapement, river harvest, and run size estimates, 2005 run. Available from U. S. Fish and Wildlife Service, 1829 South Oregon Street, Yreka, California 96097.

KRTT (Klamath River Technical Team) (2013). Klamath River fall Chinook salmon age-specific escapement, river harvest, and run size estimates, 2012 run. http://www.pcouncil.org/wp-content/uploads/age_comp_final_27Feb2013.pdf.

Logan, E., B. Matilton, D. Williams, and G. Kautsky (2013). Klamath-Trinity River fall run Chinook scale age analysis. Proposed investigation for federal fiscal year FY2014. Available from Trinity River Restoration Program, 1313 South Main Street, Weaverville, California 96093.

Satterthwaite, W. H., M. S. Mohr, M. R. O'Farrell, and B. K. Wells (2012). A Bayesian hierarchical model of size-at-age in ocean-harvested stocks – quantifying effects of climate and temporal variability. *Canadian Journal of Fisheries and Aquatic Sciences 69*(5), 942–954.

RECENT TECHNICAL MEMORANDUMS

SWFSC Technical Memorandums are accessible online at the SWFSC web site (http://swfsc.noaa.gov). Copies are also available from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161 (http://www.ntis.gov). Recent issues of NOAA Technical Memorandums from the NMFS Southwest Fisheries Science Center are listed below:

NOAA-TM-NMFS-SWFSC-512 The Sacramento Index *(SI)*.
O'FARRELL, M. R., M. S. MOHR, M. L. PALMER-ZWAHLEN, and A. M. GROVER
(June 2013)

513 Sample size recommendations for estimating stock composition using genetic stock identification (GSI).
ALLEN, S. D., W. H. SATTERTHWAITE, and M. S. MOHR
(June 2013)

514 Sources of human-related injury and mortality for U. S. Pacific west coast marine mammal stock assessments, 2007-2011.
CARRETTA, J. V., S. M. WILKIN, M. M. MUTO, and K. WILKINSON
(July 2013)

515 Photographic guide of pelagic juvenile rockfish (*SEBASTES* SPP.) and other fishes in mid-water trawl surveys off the coast of California.
SAKUMA, K. M., A. J. AMMANN, and D. A. ROBERTS
(July 2013)

516 Form, function and pathology in the pantropical spotted dolphin (*STENELLA ATTENUATA*).
EDWARDS, E. F., N. M. KELLAR, and W. F. PERRIN
(August 2013)

517 Summary of PAMGUARD beaked whale click detectors and classifiers used during the 2012 Southern California behavioral response study.
KEATING, J. L., and J. BARLOW
(September 2013)

518 Seasonal gray whales in the Pacific northwest: an assessment of optimum sustainable population level for the Pacific Coast Feeding Group.
PUNT, A. E., and J. E. MOORE
(September 2013)

519 Documentation of a relational database for the Oregon sport groundfish onboard sampling program.
MONK, M. E., E. J. DICK, T. BUELL, L. ZUMBRUNNEN, A. DAUBLE and D. PEARSON
(September 2013)

520 A fishery-independent survey of cowcod (*SEBASTES LEVIS*) in the Southern CA bight using a remotely operated vehicle (ROV).
STIERHOFF, K. L., S. A. MAU, and D. W. MURFIN
(September 2013)

521 Abundance and biomass estimates of demersal fishes at the footprint and piggy bank from optical surveys using a remotely operated vehicle (rov).
STIERHOFF, K. L., J. L. BUTLER, S. A. MAU, and D. W. MURFIN
(September 2013)