Regional Information Report 5J12-26

Western Alaska Salmon Stock Identification Program Technical Document 26: Reporting Measures of Central Tendency

by

Christopher Habicht,

Andrew R. Munro,

and

William D. Templin,

November 2012

Alaska Department of Fish and Game **Divisions of Sport Fish and Commercial Fisheries**

Symbols and Abbreviations

The following symbols and abbreviations, and others approved for the Système International d'Unités (SI), are used without definition in the following reports by the Divisions of Sport Fish and of Commercial Fisheries: Fishery Manuscripts, Fishery Data Series Reports, Fishery Management Reports, Special Publications and the Division of Commercial Fisheries Regional Reports. All others, including deviations from definitions listed below, are noted in the text at first mention, as well as in the titles or footnotes of tables, and in figure or figure captions.

Weights and measures (metric)

centimeter	cm
deciliter	dL
gram	g
hectare	ha
kilogram	kg
kilometer	km
liter	L
meter	m
milliliter	mL
millimeter	mm

Weights and measures (English)

cubic feet per second	ft^3/s
foot	ft
gallon	gal
inch	in
mile	mi
nautical mile	nmi
ounce	oz
pound	lb
quart	qt
yard	yd

Time and temperature

day	d
degrees Celsius	°C
degrees Fahrenheit	°F
degrees kelvin	K
hour	h
minute	min
second	s

Physics and chemistry

all atomic symbols	
alternating current	AC
ampere	A
calorie	cal
direct current	DC
hertz	Hz
horsepower	hp
hydrogen ion activity (negative log of)	pH
parts per million	ppm
parts per thousand	ppt, ‰
volts	V
watts	W

General

Alaska Administrative Code	AAC
all commonly accepted abbreviations	e.g., Mr., Mrs., AM, PM, etc.
all commonly accepted professional titles	e.g., Dr., Ph.D., R.N., etc.
at	@
compass directions:	
east	E
north	N
south	S
west	W
copyright	©
corporate suffixes:	
Company	Co.
Corporation	Corp.
Incorporated	Inc.
Limited	Ltd.
District of Columbia	D.C.
et alii (and others)	et al.
et cetera (and so forth)	etc.
exempli gratia (for example)	e.g.
Federal Information Code	FIC
id est (that is)	i.e.
latitude or longitude	lat. or long.
monetary symbols (U.S.)	$, ¢
months (tables and figures): first three letters	Jan,...,Dec
registered trademark	®
trademark	™
United States (adjective)	U.S.
United States of America (noun)	USA
U.S.C.	United States Code
U.S. state	use two-letter abbreviations (e.g., AK, WA)

Mathematics, statistics

all standard mathematical signs, symbols and abbreviations	
alternate hypothesis	H_A
base of natural logarithm	*e*
catch per unit effort	CPUE
coefficient of variation	CV
common test statistics	(F, t, χ^2, etc.)
confidence interval	CI
correlation coefficient (multiple)	R
correlation coefficient (simple)	r
covariance	cov
degree (angular)	°
degrees of freedom	df
expected value	*E*
greater than	>
greater than or equal to	≥
harvest per unit effort	HPUE
less than	<
less than or equal to	≤
logarithm (natural)	ln
logarithm (base 10)	log
logarithm (specify base)	log_2, etc.
minute (angular)	'
not significant	NS
null hypothesis	H_O
percent	%
probability	P
probability of a type I error (rejection of the null hypothesis when true)	α
probability of a type II error (acceptance of the null hypothesis when false)	β
second (angular)	"
standard deviation	SD
standard error	SE
variance	
population	Var
sample	var

REGIONAL INFORMATION REPORT 5J12-26

WESTERN ALASKA SALMON STOCK IDENTFICIATION PROGRAM TECHNICAL DOCUMENT 26: REPORTING MEASURES OF CENTRAL TENDENCY

by
Christopher Habicht, Andrew R. Munro, William D. Templin, and James R. Jasper
Alaska Department of Fish and Game, Division of Commercial Fisheries, Gene Conservation Laboratory, Anchorage

Alaska Department of Fish and Game
Division of Sport Fish, Research and Technical Services
333 Raspberry Road, Anchorage, Alaska, 99518-1565

November 2012

Christopher Habicht, Andrew R. Munro, William D. Templin, and James R. Jasper
Alaska Department of Fish and Game, Division of Commercial Fisheries,
333 Raspberry Road, Anchorage, Alaska, 99518-1565 USA

This document should be cited as:
Habicht, C., A. R. Munro, and W. D. Templin. 2012. Western Alaska Salmon Stock Identification Program Technical Document 26: Reporting measures of central tendency. Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 5J12-26, Anchorage.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF APPENDICES

Appendix **Page**

ABSTRACT

Uncertainty about the magnitude, frequency, location, and timing of nonlocal harvest of sockeye and chum salmon in Western Alaska fisheries was the impetus for the Western Alaska Salmon Stock Identification Program (WASSIP). The program was designed to use genetic data in mixed stock analysis to reduce this uncertainty, to identify stock contributions with the greatest resolution possible and extend these results to stock specific harvest rates where practicable. During joint meetings of the WASSIP Advisory Panel and Technical Committee, we examined different methods to reduce bias in the estimation of proportions for reporting groups because small deviations in estimation of small proportions can lead to high relative biases. This report documents the results of the WASSIP consensus-driven decision making process, presenting 1) input from the Technical Committee, 2) demonstration of the issue of cumulative bias in estimated harvest when a stock is not actually present, 3) proposed changes to reporting methods, and 4) proposals on how to best format all results to work well in the WASSIP framework and in future studies. In a simulated example of known composition, reporting medians and using an informed prior both reduced the bias in the estimated contribution of reporting groups that were not truly present in the mixture. Extending the stock composition estimates to simulated harvest numbers (coefficient of variation [CV] of 5% and assumed harvest of 100,000 fish) resulted in reduced deviations from the correct harvest using the medians compared to the means, with the exception of Togiak where the error in estimated harvest using the median was larger than using the mean. Based on these results and input from the Technical Committee and Advisory Panel, we proposed reporting the median and 90% credibility intervals as the best estimates in the reports. Means and standard deviations are also important measures to report, despite the problems associated with using them for this study and will be more useful than medians when these data are used in future studies. The medians, 90% credibility intervals, means, and standard deviations will be reported in the appendices. We will also report the probability that the catch from a given reporting group in a particular fishery is in fact zero. To reduce the confusion and maintain consistency both within and across WASSIP reports, harvest medians will be normalized so the sum of the median harvest numbers will sum to the total reported fishery harvest and the sum of the median proportions will sum to 1. These decisions will work best in the WASSIP framework, where the tabled data will be used to inform policy decisions.

Key words: Western Alaska Salmon Stock Identification Program, WASSIP, mixed stock analysis, MSA, stock composition estimates, cumulative bias, harvest numbers

INTRODUCTION

During the joint Advisory Panel (AP)/Technical Committee (TC) meeting held in Anchorage on May 16, 2012, Gene Conservation Laboratory presented issues associated with biases in estimation of proportions for reporting groups that were not present in artificial known mixtures. These biases, although low, when multiplied by the harvest numbers result in high harvest numbers relative to the escapement for reporting groups with small escapement. Estimating harvest throughout the Western Alaska Salmon Stock Identification Program (WASSIP) fisheries for reporting groups with small escapements is one of the interests for the AP. Small deviations in estimation of small proportions can lead to high relative biases. An example demonstrating this issue was shown in the meeting (Flat prior example in Table 1).

At the meeting, various options to reduce these biases were discussed, including trimmed means, censuring posterior distributions for outliers, and changing the measure of central tendency from the mean to the median. Both Dr. Adkison (at the meeting) and Dr. Weir (on a teleconference call prior to the meeting) expressed concerns with manipulating the posterior distributions, because these manipulations would be *ad hoc* and would likely suffer poor peer review. However, both TC members supported using other measures of central tendency, such as the median, that are not as sensitive to extreme values in the posterior distributions (Appendices A and B).

At the meeting, the department raised concerns that if the median is used as the measure for central tendency, the sum of the proportions within each mixture will not add to 1. This issue will be perpetuated when these proportions are multiplied by the harvest because the harvest numbers will not sum to the harvest numbers reported in the sampling report.

As a side note, the harvest numbers reported in the stock composition reports will not be identical to those reported in the sampling report because the true harvest numbers are not known with certainty and an estimate will be derived from a distribution based on coefficients of variation (CVs) agreed to by the AP and TC at the May 16, 2012 meeting.

Dr. Adkison had 3 proposals/recommendations. First he suggested using an informative prior to reduce this bias caused by non-zero estimates when contributions truly are zero (Appendix A). Although the example provided at the meeting used flat priors, this recommendation is already being applied as we are using informative priors for the WASSIP mixtures (e.g., Jasper et al. 2012).

Second, Dr. Adkison pointed out: "the multinomial likelihood constrains the contributions to be zero or greater. If the contribution is in fact zero, any uncertainty must lead to some inclusion of probability for contributions greater than zero, automatically biasing the averaged contribution upwards. It might be possible to replace the likelihood with one that allowed for the possibility of negative contributions (rescaling a multivariate Gaussian?), but the canned software probably doesn't exist and ADF&G staff probably couldn't develop and implement this modification of the Pella-Masuda Method this quickly" (Appendix A). We agree that this method would be worth pursuing, but also agree that we do not think we can develop the methods within the WASSIP timeframe.

Finally, Dr. Adkison suggested reporting the probability that the catch in a particular fishery is in fact zero (Appendix A). This information could be used to report a summed catch which excludes portions from reporting groups for which the probability that the catch was zero was greater than 50%, or 1/3, or some other sensible probability. This will illustrate the influence of highly skewed posterior distributions of stock compositions where the estimate of the mean is biased high.

Given the TC input and further consideration by the department, we propose reporting the median (and 90% credibility intervals [CIs]) in the reports accompanied by a footnote stating that they do not sum to 1. Means and standard deviations are also important measures to report, despite the problems associated with using them for this study and will be more useful than medians when using these data in future studies. Therefore, regardless of the central tendency value released in tables for this study, the medians, 90% CIs, means, and standard deviations will be reported in the appendices. We also report the probability that the catch from a given reporting group in a particular fishery is in fact zero. Here we provide the example demonstrating the issue shown in the meeting and we continue with the same example to examine differences in reporting stock composition results using the proposed changes in reporting methods.

METHODS

DEMONSTRATION OF THE ISSUE

A 400-fish mixture composed of sockeye salmon from 14 of the 24 reporting groups was assembled (Table 1). These fish were subsampled from the same sets of collections used in the baseline escapement tests (Dann et al. 2012). These samples were not in the baseline, but were taken from freshwater sites within the rivers and are thought to represent fish spawning within their respective reporting groups. The proportions of the mixture were based on 2 factors: 1) what might be expected in an Alaska Peninsula fishery, and 2) the availability of inriver samples.

Stock composition estimation followed the methods described in Dann et al. (2012) where the software BAYES was used to create a posterior distribution of 100,000 stock composition estimates from the last 20,000 iterations from each of 5 chains. Two posterior distributions were created, the first set of estimates were produced using a flat prior (all reporting groups at 1/24) and a second set of estimates were produced using an informative prior (all unrepresented reporting groups at 1% each and the remaining percentage divided among the represented reporting groups in proportion to the known proportion). The mean, median, and the upper and lower bounds of the 90% credibility intervals were calculated. We also calculated the probability that the catch from a given reporting group in a particular fishery is in fact zero by calculating the proportion of stock composition estimates in the posterior distribution for that reporting group that were less than 0.0000005, which in a harvest of 1 million fish is less than 1 fish.

Translating stock proportions into estimates of stock-specific harvest was done by multiplying each of the 100,000 stock composition iterations in the posterior by an estimate of the number of fish harvested drawn from a log-normal distribution assuming a mean harvest of 100,000 and 5% CV. Summary statistics were calculated for this distribution in the same manner as the posterior stock composition estimates.

The cumulative effect of bias in estimated harvest when a stock was not actually present was investigated using the Norton Sound reporting group estimates from the previous simulations. The proportional stock composition estimate from each of the two simulations was applied across all fisheries using the harvest numbers within all fishery and temporal strata for sockeye salmon. Estimated escapements for Norton Sound 2006 to 2008 were taken from previously reported escapement estimates (Eggers et al. 2012).

As detailed in the introduction, the department was concerned that both the sum of the median proportion would not sum to 1 and the sum of the median harvest numbers would not sum to the total reported fishery harvest. For reporting purposes, we propose to normalize the median using the following equation:

$$\tilde{p}_i' = \frac{\tilde{p}_i}{\sum \tilde{p}_i}$$

Where $\tilde{p}_i$ is the median proportion of the i^{th} reporting group and $\tilde{p}_i'$ is the normalized median for the i^{th} reporting group.

Similarly, we propose to normalize the harvest median using the following equation applied after the uncertainty of the harvest numbers has been applied to the uncertainty of the stock composition estimates:

$$\tilde{h}_i' = H \times \left(\frac{\tilde{h}_i}{\sum \tilde{h}_i}\right)$$

Where $\tilde{h}_i$ is the median harvest of the i^{th} reporting group, H is the total reported fishery harvest, and $\tilde{h}_i'$ is the normalized median harvest for the i^{th} reporting group.

Normalized median proportions and harvest will only be calculated at the end for tabulation in the reports and these numbers will not be used for calculation of the harvest rates. Harvest rate statistics will be calculated using raw posterior distributions of stock composition, harvest, and escapement estimates.

PROPOSED REPORTING

We propose to report the normalized median and 5% and 95% credibility intervals in the tables of the stock composition estimate report. We propose to report the normalized median, 5% and 95% credibility intervals, the mean, the standard deviation of the mean, and the proportion of posterior probability estimates what were less than 0.0000005 ($P = 0$) in the appendices. The solution proposed here was based on comments from the TC (see Introduction and Appendix A) and the approval from the AP at the May 16, 2012 joint AP/TC WASSIP meeting to pursue appropriate methods to reduce the influence of extreme posterior values on the stock composition estimates.

RESULTS

Reporting medians and using an informed prior both reduced the bias in the estimated contribution of reporting groups that were not truly present in the mixture (Tables 1 and 2; Figure 1). Of these, the median was the least biased. For all but one reporting group that were truly represented in the mixture, the choice of central tendency measure (mean or median) and type of prior (flat or informative), had little effect on the estimates. One reporting group (Togiak) performed worse when using the median than the mean regardless of the prior. The sum of the means added to 1.00, but the sum of the medians added to 0.97, showing a negative bias for this measure of central tendency (Table 1).

Extending the stock composition estimates to harvest numbers with a CV of 5% (assumed harvest of 100,000 fish) resulted in deviations from the correct harvest for both the summed means and the medians (Table 2). These deviations were smaller for the means (+19 to +24 fish) than the medians (–2,897 to –2,943 fish). Relative deviations from correct values paralleled those observed from the proportional stock composition estimates (Table 1).

Proposed formats for report tables and appendices are shown in Tables 3 and 4. The report tables can provide statistics for up to 4 area/time strata, while the appendix format will provide statistics for 2 area/time strata. As a result, each table will correspond to 1 or 2 appendices. A footnote was added explaining why the median was chosen as the measure of central tendency. A footnote will also be added to the harvest number tables and appendices explaining that the sum of the harvest may not add up to the harvest totals previously reported because of uncertainty applied to the harvest numbers.

DISCUSSION

Assuming that the example mixture provides similar results for other WASSIP sockeye salmon fisheries and that these other fisheries also do not contain any Norton Sound fish, using the mean, we would incorrectly calculate that a large proportion of the return was harvested. Total harvest of WASSIP fisheries south of the Yukon River was between 31 million and 35 million for the WASSIP years (2006–2008; Eggers et al. 2011). The proportions of incorrectly allocated Norton Sound fish in the example were 0.00079 for the flat prior and 0.00023 for the informative prior (Table 1). Expanding the total harvest of WASSIP fisheries south of the Yukon River by these proportions, we come up with 7,000 to 28,000 fish incorrectly allocated. This compares to an estimated escapement for Norton Sound of 22,000 to 64,000 fish. Using the median, on the other hand, would not allocate any fish (regardless of the prior) to Norton Sound (Table 1).

The negative bias in the total harvest introduced by using medians was expected and it presents a problem when summing up harvest numbers. The total harvest numbers will sum to less than the harvest numbers reported in the sampling report (Eggers et al. 2011, *Results from Sampling the 2006–2009 Commercial and Subsistence Fisheries in the Western Alaska Salmon Stock Identification Project*). We propose normalizing the proportions so that they sum to 1 and the harvests so that they sum to the total reported fishery harvest. The goal of normalizing is to reduce confusion. This procedure is *ad hoc* but not subjective.

The better performance of the median relative to the mean for the reporting groups that were not represented in the simulated mixture and with the poorer relative performance for the Togiak reporting group demonstrates that neither measure is perfect. The choice of the best measure of central tendency depends on which error/bias is more important to the stakeholders. In the WASSIP framework, where the tabled data will be used to inform policy decisions, the median may be more useful because of the reduced bias at low stock proportions. For other applications, where the results may be used in conjunction with other data (i.e., building brood tables) the mean may be a better measure. The probability of harvest from a specific reporting group being zero may also be useful for making policy decisions and this number will be provided in the appendices. Because both the median and mean will be reported (median in the tables and mean and median in the appendices), our recommended reporting strategy should meet both types of needs.

ACKNOWLEDGMENTS

The Technical Document series served as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and the Western Alaska Salmon Stock Identification Program (WASSIP) Technical Committee during the implementation of the program. The authors would like to thank the WASSIP Technical Committee and Advisory Panel for their constructive input on each of the documents throughout the project. The authors would also like to thank Erica Chenoweth who coordinated and prepared the Technical Document series for publication and Publication Specialists Amy Carroll and Joanne MacClellan for implementing the series into Regional Information Reports.

REFERENCES CITED

Dann, T. H., C. Habicht, J. R. Jasper, E. K. C. Fox, H. A. Hoyt, H. L. Liller, E. S. Lardizabal, P. A. Kuriscak, Z. D. Grauvogel, and W. D. Templin. 2012. Sockeye salmon baseline for the Western Alaska Salmon Stock Identification Project. Alaska Department of Fish and Game, Special Publication No. 12-12, Anchorage.

Eggers, D. M., A. R. Munro, and E .C. Volk. 2012. Estimating escapement of Western Alaskan sockeye salmon for Western Alaska Salmon Stock Identification Program reporting groups, 2006 to 2008. Alaska Department of Fish and Game, Special Publication No. 12-15, Anchorage.

Eggers, D. M., M. J. Witteveen, T. T. Baker, D. F. Evenson, J. M. Berger, H. A. Hoyt, H. L. Hildebrand, W. D. Templin, C. Habicht, and E. C. Volk. 2011. Results from sampling the 2006—2009 commercial and subsistence fisheries in the Western Alaska Salmon Stock Identification Project. Alaska Department of Fish and Game, Special Publication No. 11-10, Anchorage.

Jasper, J., S. Turner, and C. Habicht. 2012. Western Alaska Salmon Stock Identification Program Technical Document 13: Selection of a prior for mixed stock analysis. Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 5J12-20, Anchorage.

QUESTIONS FOR TECHNICAL COMMITTEE

1) Do the tradeoffs in benefits between medians and means support using the median as the measure of central tendency in the report tables? Is there another measure that we should consider?
2) Should the mean and CV be provided in the appendices, as proposed?
3) Should the probability that the estimate is in fact zero be included in the tables and/or the appendices? Is there a threshold probability that should be applied to determine when an estimate is zero?
4) Should we normalize the median proportions to make them sum to 1 within each area/time fishery estimate as proposed?

Technical committee comments are included in the appendices to this document.

TABLES

Table 1.–Estimated mean, median, standard deviation (SD), and upper and lower level 90% credibility intervals (CI.05 and CI.95) of stock-specific proportions in a mixture of known composition (TRUE) using a flat prior (all reporting groups at 1/24) and an informative prior (reporting groups not present at 1% and the remaining prior distributed across the other reporting groups in proportion to the TRUE proportions). The proportion of posterior probability estimates that were less than 0.0000005 (*P* = 0) is included. The known composition mixture was assembled from fish collected from inriver assessment projects.

Subregional		Flat prior						Informative prior					
reporting group	TRUE	mean	median	SD	CI.05	CI.95	*P* = 0	mean	median	SD	CI.05	CI.95	*P* = 0
Norton Sound	0.00000	0.00079	0.00000	0.00270	0.00000	0.00549	0.63445	0.00023	0.00000	0.00149	0.00000	0.00029	0.89033
Kuskokwim River	0.02000	0.01817	0.01470	0.01165	0.00563	0.04257	0.00000	0.01638	0.01333	0.01072	0.00532	0.03966	0.00000
Kanektok	0.02000	0.02063	0.01957	0.00797	0.00979	0.03488	0.00000	0.02019	0.01927	0.00748	0.00967	0.03367	0.00000
Goodnews	0.02000	0.02991	0.03065	0.01621	0.00174	0.05613	0.00001	0.02491	0.02641	0.01701	0.00072	0.05238	0.00000
Togiak	0.02000	0.00929	0.00004	0.01460	0.00000	0.04123	0.41818	0.01340	0.00013	0.01739	0.00000	0.04689	0.45222
Igushik	0.04000	0.06246	0.06154	0.02218	0.02781	0.10049	0.00260	0.05774	0.05667	0.02096	0.02534	0.09391	0.00284
Wood	0.15000	0.13341	0.13244	0.02630	0.09182	0.17809	0.00000	0.13748	0.13678	0.02599	0.09595	0.18135	0.00000
Nushagak	0.05000	0.05192	0.04995	0.02356	0.01656	0.09245	0.00612	0.05656	0.05567	0.02325	0.02093	0.09579	0.00504
Kvichak	0.10000	0.09704	0.09602	0.01954	0.06672	0.13069	0.00000	0.09747	0.09655	0.01927	0.06739	0.13072	0.00000
Alagnak	0.05000	0.02778	0.02668	0.01056	0.01257	0.04682	0.00000	0.02804	0.02686	0.01063	0.01279	0.04736	0.00000
Naknek	0.10000	0.12946	0.12862	0.02305	0.09290	0.16862	0.00000	0.12879	0.12801	0.02284	0.09243	0.16765	0.00000
Egegik	0.15000	0.17017	0.16878	0.03673	0.11279	0.23217	0.00000	0.17532	0.17475	0.03448	0.12006	0.23329	0.00000
Ugashik	0.15000	0.10558	0.10487	0.03482	0.05092	0.16363	0.00256	0.10249	0.10113	0.03208	0.05266	0.15746	0.00029
Cinder	0.00000	0.00044	0.00000	0.00166	0.00000	0.00281	0.65833	0.00011	0.00000	0.00085	0.00000	0.00008	0.90220
Meshik	0.00000	0.00028	0.00000	0.00129	0.00000	0.00143	0.68567	0.00008	0.00000	0.00076	0.00000	0.00003	0.91141
Ilnik	0.00000	0.00043	0.00000	0.00167	0.00000	0.00269	0.66016	0.00013	0.00000	0.00095	0.00000	0.00008	0.90218
Sandy	0.00000	0.00028	0.00000	0.00118	0.00000	0.00163	0.68157	0.00007	0.00000	0.00057	0.00000	0.00004	0.91254
Bear	0.00000	0.00020	0.00000	0.00092	0.00000	0.00099	0.69656	0.00005	0.00000	0.00050	0.00000	0.00002	0.91684
Nelson	0.00000	0.00011	0.00000	0.00052	0.00000	0.00054	0.71530	0.00002	0.00000	0.00024	0.00000	0.00001	0.92510
NW District	0.00000	0.00247	0.00151	0.00293	0.00000	0.00838	0.13860	0.00144	0.00004	0.00251	0.00000	0.00668	0.47177
South Peninsula	0.00000	0.00024	0.00000	0.00114	0.00000	0.00118	0.69359	0.00006	0.00000	0.00060	0.00000	0.00002	0.91486
Black Lake	0.08000	0.09186	0.09123	0.01700	0.06513	0.12073	0.00000	0.09286	0.09221	0.01696	0.06609	0.12170	0.00000
Chignik Lake	0.05000	0.04697	0.04582	0.01410	0.02590	0.07197	0.00000	0.04615	0.04502	0.01390	0.02536	0.07077	0.00000
East of WASSIP	0.00000	0.00012	0.00000	0.00062	0.00000	0.00059	0.71123	0.00003	0.00000	0.00029	0.00000	0.00001	0.92172
Sum		1.00000	0.97242					1.00000	0.97282				

Table 2.–Estimated mean, median, standard deviation (SD), and upper and lower level 90% credibility intervals (CI.05 and CI.95) of stock-specific harvest (assuming 100,000 fish harvest), a harvest CV of 5%, and using proportions from a mixture of known composition (TRUE) estimated using a flat prior (all reporting groups at 1/24) and an informative prior (reporting groups not present at 1% and the remaining prior distributed across the other reporting groups in proportion to the TRUE proportions). The proportion of posterior probability estimates what were less than 0.0000005 (*P* = 0) is included.

Subregional		Flat prior						Informative prior					
reporting group	TRUE	mean	median	SD	CI.05	CI.95	*P* = 0	mean	median	SD	CI.05	CI.95	*P* = 0
Norton Sound	0	79	0	271	0	548	0.69847	23	0	149	0	29	0.91113
Kuskokwim River	2,000	1,818	1,469	1,170	560	4274	0.00000	1,639	1,330	1,076	529	3,969	0.00000
Kanektok	2,000	2,064	1,956	805	974	3507	0.00000	2,019	1,925	755	961	3,385	0.00000
Goodnews	2,000	2,991	3,054	1,629	174	5644	0.00012	2,491	2,639	1,708	72	5,255	0.00032
Togiak	2,000	929	4	1,464	0	4136	0.46024	1,340	13	1,742	0	4,710	0.47171
Igushik	4,000	6,247	6,143	2,242	2,762	10098	0.00286	5,775	5,659	2,119	2,525	9,458	0.00299
Wood	15,000	13,343	13,225	2,715	9,097	17995	0.00000	13,750	13,654	2,691	9,503	18,348	0.00000
Nushagak	5,000	5,193	4,983	2,373	1,647	9287	0.00669	5,657	5,555	2,347	2,084	9,637	0.00558
Kvichak	10,000	9,707	9,582	2,017	6,613	13205	0.00000	9,749	9,644	1,991	6,672	13,198	0.00000
Alagnak	5,000	2,779	2,660	1,066	1,249	4706	0.00000	2,805	2,682	1,074	1,272	4,754	0.00000
Naknek	10,000	12,949	12,839	2,399	9,195	17076	0.00000	12,882	12,782	2,373	9,148	16,955	0.00000
Egegik	15,000	17,021	16,843	3,778	11,189	23477	0.00000	17,535	17,430	3,562	11,902	23,586	0.00000
Ugashik	15,000	10,560	10,463	3,526	5,059	16471	0.00286	10,251	10,089	3,255	5,223	15,883	0.00031
Cinder	0	44	0	166	0	281	0.72426	11	0	85	0	8	0.92321
Meshik	0	28	0	129	0	142	0.75479	8	0	76	0	3	0.93221
Ilnik	0	43	0	168	0	268	0.72860	12	0	95	0	8	0.92361
Sandy	0	28	0	118	0	164	0.74981	7	0	57	0	3	0.93294
Bear	0	20	0	92	0	99	0.76605	5	0	50	0	2	0.93830
Nelson	0	11	0	52	0	54	0.78680	2	0	24	0	1	0.94632
NW District	0	247	151	294	0	840	0.15348	144	4	252	0	668	0.48343
South Peninsula	0	24	0	114	0	118	0.76429	6	0	60	0	2	0.93625
Black Lake	8,000	9,189	9,109	1,767	6,447	12225	0.00000	9,288	9,205	1,761	6,536	12,301	0.00000
Chignik Lake	5,000	4,698	4,576	1,431	2,571	7244	0.00000	4,617	4,492	1,411	2,524	7,134	0.00000
East of WASSIP	0	12	0	62	0	59	0.78354	3	0	29	0	1	0.94300
Sum	100,000	100,024	97,057					100,019	97,103				

Table 3.–Proposed format for stock composition report tables (this example is for a Bristol Bay Area fishery for sockeye salmon). Normalized medians (Med.) and the lower and upper 90% credibility intervals (CI.05 and CI.95) are reported. As with previously presented tables, the date range and the number of fish successfully genotyped (N) are provided for each area/time stratum.

	Stratum 1 (6/12-16; N = 387)			Stratum 2 (6/18-28; N = 399)			Stratum 3 (6/28-7/10; N = 380)			Stratum 4 (7/10-7/30; N = 360)		
	Med.	CI.05	CI.95	Med.	CI.05	CI.95	Med.	CI.05	CI.95	Med.	CI.05	CI.95
Norton Sound	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Kuskokwim Bay	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Bristol Bay	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
North Peninsula	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
South Peninsula	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Chignik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
East of WASSIP	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Bristol Bay												
Togiak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Igushik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Wood	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Nushagak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Kvichak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Alagnak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Naknek	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Egegik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Ugashik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx

Note: Medians are reported because they are less susceptible to the influence of outliers than means and are therefore less biased for small stock-composition estimates.

Table 4.–Proposed format for stock composition report appendices (this example is for sockeye salmon). Summary statistics from the posterior distributions are provided as the normalized median (Med.), bounds of the 90% credibility intervals (CI.05 and CI.95), the mean (Mean), the standard deviation of the mean (SD), and the proportion of posterior probability estimates that were less than 0.0000005 ($P = 0$). As with previously presented appendices, the date range and the number of fish successfully genotyped (N) are provided for each area/time stratum. Note that only two strata are presented in each appendix so that they fit on a single page (there will be two appendices for each table).

Reporting Group	Stratum 1 (6/12-16; N = 387)						Stratum 2 (6/18-28; N = 399)					
	Med.	CI.05	CI. 95	Mean	SD	$P = 0$	Med.	CI.05	CI.95	Mean	SD	$P = 0$
Norton Sound	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Kuskokwim Bay												
Kuskokwim R.	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Kanektok	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Goodnews	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Bristol Bay												
Togiak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Igushik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Wood	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Nushagak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Kvichak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Alagnak	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Naknek	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Egegik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Ugashik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
North Peninsula												
Cinder	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Meshik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Ilnik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Sandy	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Bear	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Nelson River	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
NW Dist.-BH	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
South Peninsula												
Chignik	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Black Lake	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
Chignik Lake	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx
East of WASSIP	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx	x.xx

FIGURES

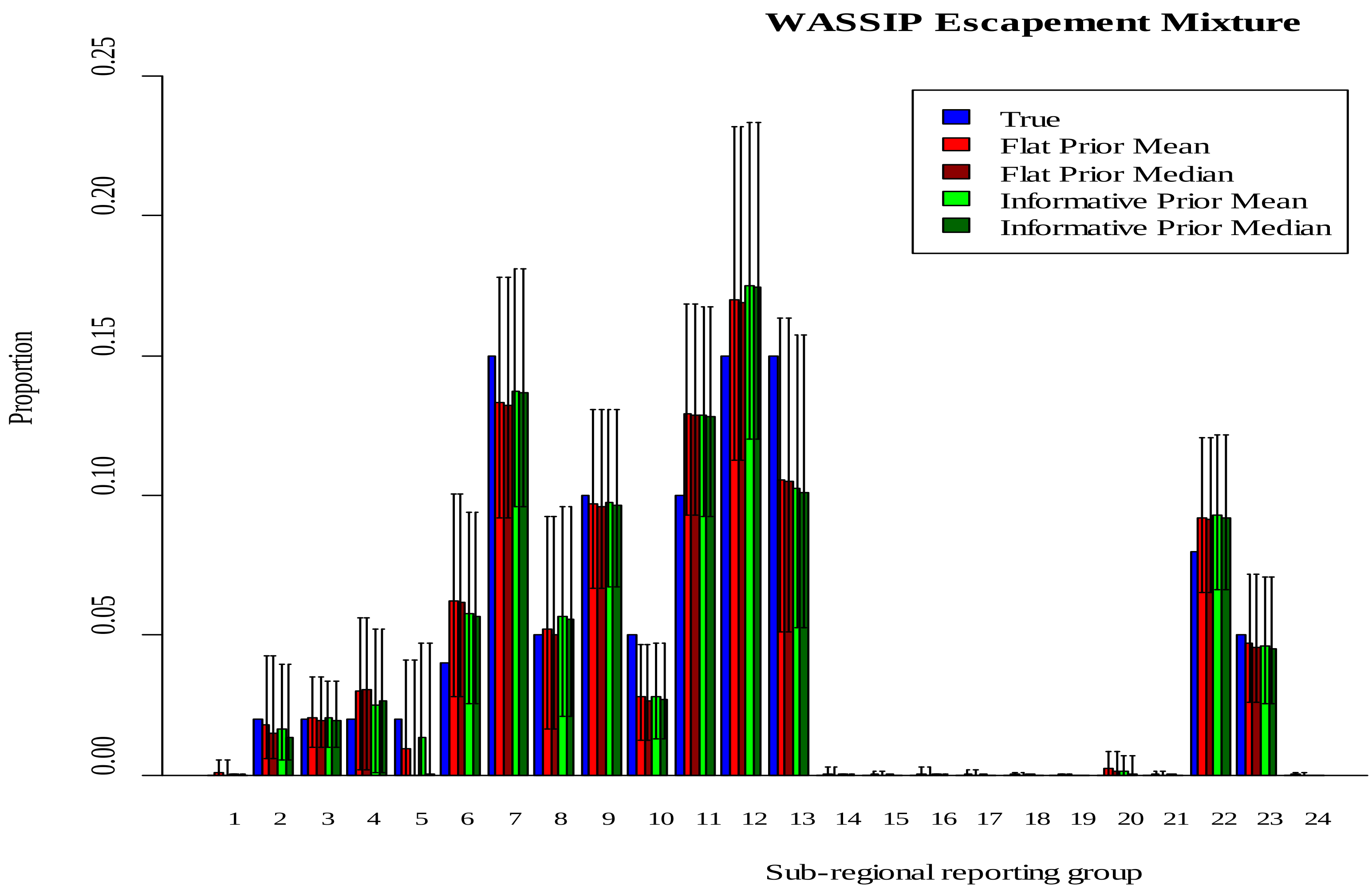

Figure 1.–Stock composition of a mixture with known composition (True) estimated from means and medians using a flat prior (all reporting groups at 1/24) and an informative prior (all reporting groups at the True proportions). Whiskers indicate the width of the 90% credibility intervals. The known composition mixture was assembled from fish taken from inriver assessment projects within reporting groups. Subregional reporting group are arranged in the same order as in Table 1.

APPENDICES

Appendix A.–E-mail message sent by Milo Adkison to TC and department staff commenting on bias in harvest estimates of small stocks on 5/17/2012.

The procedure for estimating harvests of stocks in the WASSIP fisheries has an upwards bias for stocks that are a small component of the catch. As Bill Templin pointed out, in the case of a small stock and a fishery that catches a lot of fish, this can result in an apparently large harvest rate on the stock even if the fishery is catching no fish from the stock. As Bruce Weir said, this isn't an error – ADF&G is implementing the estimation procedure correctly, but this bias is intrinsic to the estimation procedure. I'm not sure if other people have dealt with this issue in mixture estimation – WASSIP is somewhat unique in its interest in the small components of mixtures.

There are two factors that lead to estimates of significant catch from stocks that are absent or minimally present in a fishery. First, the multinomial likelihood constrains the contributions to be zero or greater. If the contribution is in fact zero, any uncertainty must lead to some inclusion of probability for contributions greater than zero, automatically biasing the averaged contribution upwards. It might be possible to replace the likelihood with one that allowed for the possibility of negative contributions (rescaling a multivariate Gaussian?), but the canned software probably doesn't exist and ADF&G staff probably couldn't develop and implement this modification of the Pella-Masuda method this quickly.

The other factor is the uninformative prior. The prior can be quite influential in cases where the genetic separation among groups is weak. Even in cases where a priori knowledge would suggest that a large harvest in a fishery is improbable because the stock is very small or very distant, the prior used will tend to push the estimated contribution to be equal to that of other stocks at the same level (reporting group or sub-group). The theoretical solution is to use an informative prior that takes in all useful information external to WASSIP (stock size, geographic proximity, distributional information from tagging, parasite studies, etc.). We discussed this earlier in the process and rejected it as the main approach because of the paucity of the information available. However, as a sensitivity analysis such priors might still be worth a look. One such prior could be proportional to the sum of terminal catch and escapement.

At this last meeting, several post-hoc modifications of the MCMC draws from the posterior were proposed to reduce the bias in the harvest estimates. There is a strong statistical foundation to the existing procedure that estimates harvest and captures the uncertainty in their estimates in a rigorous way. The MCMC approach represents the results faithfully, and any culling of the draws is an ad hoc modification of the posterior distribution. This would result in your entire process being ad hoc, and if you're going to treat the data in an ad hoc fashion there are much simpler procedures you could use. However, given the existing approach, you can be creative in extracting results from the posterior in a way that minimizes/illustrates some of the bias.

First, there is nothing wrong in reporting the median value as a measure of central tendency. This is still biased, and the sum of the medians from a fishery won't sum to the total catch, but to me that's not a big issue if the principal interest is in the reporting groups. However, when you start aggregating all of the reporting group catches across fisheries, summing up the means is still probably better.

-continued-

Second, you could also report the probability that the catch in a particular fishery is in fact zero. This calculation is pretty straightforward. I can also see that in addition to reporting the estimated catch across all fisheries, you could report a summed catch excluding that from fisheries where the probability that the catch was zero was greater than 50%, or 1/3, or some other sensible probability. This would illustrate the effect of the fisheries where the contribution might be zero but the point estimate was high because of high catches in the fishery.

Appendix B.–Excerpt of an e-mail message sent by Bruce Weir to department staff on 5/15/2012 following a teleconference.

This is in response to your query of April 10 about reporting harvest rate estimates. I have read your Technical Document 19[1], on escapement estimation in Western Alaskan chum salmon, and this has given me a new appreciation of the difficulties faced by WASSIP. I should say first that

I found the document well-reasoned and carefully written.

Estimates of quantities such as harvest rates or escapement rates are just that: estimates. Even the best sampling design and most careful data collection will lead to estimates with sampling variation and this variation should be reported. With your Bayesian approach, this is accomplished with posterior credible intervals, analogous to confidence intervals for frequentist estimates. In TD 19* you present figures for the coefficients of variation (standard deviation divided by the mean) for escapement rates and these are often around 0.5. If you adopted a crude rule of reporting an estimate plus or minus two standard deviations, you would have the wide interval from zero to twice the estimate (I realize your credible intervals are more precise and are not symmetric about the estimate).

Estimation is especially difficult when true rates are small since zero counts are likely in small samples. A frequentist rule of thumb is that a count of zero in a sample of size n has a 95% probability when the true rate is as large as 3/n (i.e. 0.01 when n=300). Moreover, even low miss-assignment rates can have a large effect on credible intervals when counts are small.

Given your methods of observation, with their inherent sampling variabilities, the only way to decrease the variation in your estimates is to increase sample size. I assume that you are constrained by resources and by actual population sizes for any particular sample, so you may be left with the option of aggregating over samples (i.e. aggregating over time strata, or area strata, or fisheries). I expect you will need empirical studies to show the effects on the widths of credible intervals for different aggregation strategies, and I would expect that these effects will be different in different regions.

Please let me know if I can be of any assistance as you move forward with your work.

---

[1] *Eggers, D., A. Munro and E. Volk. *In prep*. Estimating escapement of Western Alaskan chum salmon for Western Alaska Salmon Stock Identification Program reporting groups, 2007 to 2009. Alaska Department of Fish and Game, Special Publication, Anchorage.

Appendix C.–Excerpt of an email message from Robin Waples to Bill Templin, with cc to AP and TC members dated 6/25/12.

. . . In general I agree with the comments by Milo and Bruce and with the approach you have laid out. but I have a few comments you might consider:

Estimates for low contributors. Actually, this issue has been recognized for a long time (at least 25 years, which is when I first got involved with GSI). However, I am not aware that anyone has come up with an optimal solution.

I think it is important not to lose sight of exactly how your estimates come about. You might recall that in the TC comments provided in advance of the first meeting, we cautioned against overoptimism in predicting precision for estimating contributions of small stocks. The reason for this has more to do with the size of the sample from the fishery that with any problems in distinguishing stocks genetically. You might recall we pointed out that it is easy to calculate the best-case scenario with respect to precision, which occurs when every fish in the sample from the fishery can be assigned with 100% certainty to its population of origin. If your sample is 400 fish and the total harvest is one hundred thousand fish, then each fish in your sample gets expanded by a factor of 250. In that case, the point estimates for contributions by each stock must be multiples of 250: 0, 250, 500, 750 , ... If this fishery is a million fish, these step increases are each size 2500. This is a bit of a cartoon, of course, since you use a Bayesian analysis with informative priors based on previous contribution estimates, and some uncertainty is associated with the genetic assignments. But the point remains that whether a particular population is represented by zero or one fish in the sample from the fishery can make a huge difference in the overall estimated contribution. This intrinsic level of uncertainty cannot be reduced by anything you do later in the analysis, unless you get a larger sample from the fishery.

From this perspective it is useful to consider how you arrive at the values shown in Table 2, which show estimates of 10-50 fish for a third of the stocks, even with flat priors. My concern is that this implies a false sense of precision in the estimates which, as noted above, would be in increments of 250 fish if genetic assignments were perfect. The fact that you get estimates that are a fraction of this increment does not mean that you have somehow achieved greater precision; rather, it reflects the fact that not all populations can be reliably distinguished genetically. As a consequence, these populations attract some fractional estimates because they are genetically similar to other populations in the baseline.

Probability that an estimate is zero. Conceptually, this seems like a good idea and I would think it would be useful information for managers and stakeholders. However, again it is important to remember how this would be calculated. It seems that what you are proposing is averaging across MCMC iterations. In these iterations, the program is basically going back and forth between states that say "I am certain there are no fish in the sample from this population" and "I don't think there are any fish from this population in the sample, but I can't be sure because some of the populations are hard to tell apart." So, calling this a probability is perhaps a bit misleading. At a minimum, I think it would be useful to go back to the simple model and consider these two independent scenarios:

-continued-

1. *There are no fish in the sample from this population.* In that case, you can ask the question, "How large a fraction of the total fishery could realistically come from this population, given that we randomly sampled 400 fish and found none?" Since the sample is small compared to the total fishery, you could treat this as sampling with replacement and use the binomial distribution to generate the probabilities of getting zero hits in a sample of 400, given a range of hypothetical contributions by that stock to the overall fishery. You might pick a total contribution of that stock to the harvest that is effectively zero for all management and conservation purposes, and calculate the probability that the contribution is less than that. Or, you might pick a level that would be a concern and test whether it is plausible the contribution could have been that high.
2. There is exactly one fish in the sample from this population. In that case, you have a point estimate of 250 for the overall fishery. Again, you could use binomial sampling theory to generate a range of true contributions that are compatible with your data.

As noted above, this example is simplistic but I think it would be useful to do this exercise so you have a benchmark, best-case measure of uncertainty against which to compare the results you are getting with other methods. You should be worried if the precision you appear to get with the GSI methods is greater than you can get with this simple model.

MCMC averaging. The text on p. 3 says the posterior distribution of stock composition estimates was generated from the last 20,000 iterations of each of 5 chains. I have a couple of questions about this. First, did you really use the last 20K consecutive iterations each time? Generally, thinning is done so you skip a number of iterations between those you actually collect data from. Second, did you do any diagnostics to evaluate consistency among chains? What if different chains end up on different local peaks, of which there might be many? In that case, you might expect that variance among chains is much greater than variance among iterations within chains, and using only 5 chains might not be sufficient to characterize the likelihood surface.

Anyway, those are my thoughts on this difficult issue. Perhaps Bruce and Milo can comment on whether they think any of this makes sense.
With regard to the specific questions:

1. Overall I think it is reasonable to use the median
2. I think this is a good idea since the means sum to 1.0 and have a long history of use for comparative purposes
3. I think this is a good idea but see my comments
4. that seems reasonable; otherwise you have catch that is unaccounted for