Canadian Journal of
Fisheries and
Aquatic Sciences

OPEN ACCESS | Article

The estimated impact of changes to otolith field-sampling and ageing effort on stock assessment inputs, outputs, and catch advice

Matthew R. Siskey[a,e], **André E. Punt**[a], **Peter-John F. Hulson**[b], **Meaghan D. Bryan**[c], **James N. Ianelli**[c], and **James T. Thorson**[d]

[a]School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA; [b]Auke Bay Laboratories, Alaska Fisheries Science Center, NMFS, NOAA, Seattle, WA, USA; [c]Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, NMFS, NOAA, Seattle, WA, USA; [d]Habitat and Ecological Processes Research Program, Alaska Fisheries Science Center, NMFS, NOAA, Seattle, WA, USA; [e]Washington Department of Fish & Wildlife, Olympia, WA, USA

Corresponding author: **Matthew R. Siskey** (email: Matthew.Siskey@dfw.wa.gov)

Abstract

Generating accurate data for stock assessments is resource-demanding, necessitating periodic evaluation of survey sampling designs and potential impacts on stock assessments. We developed a framework for bootstrapped resampling of survey age data and calculation of input sample sizes as a function of among-bootstrap variance in age compositions. Data from this bootstrap estimator were then used to evaluate the influence of alternative sampling rates and methods on uncertainty in estimates of the overfishing limit (OFL) calculated using stock assessment models. For dusky rockfish (*Sebastes variabilis*) and Pacific ocean perch (*Sebastes alutus*), a 10% decrease in the number of tows sampled upon led to a predicted 5%–6% increase in the CV of OFL (log–log slope = −0.576 to −0.486), which was greater than the 0%–2% increase from a 10% decrease in otoliths-per-tow (log–log slope = −0.238 to −0.029). Application of this approach across all stocks monitored in the survey of interest is required to identify which stocks (*i*) benefit the most from increased sampling of ageing structures or (*ii*) cost the least in terms of OFL uncertainty owing to reduced sampling.

Key words: age composition, data-weighting, stock assessment, sample size, otoliths

Introduction

Sustainable fisheries management is contingent upon sound advice informed by the results of stock assessments, which describe the population dynamics of the resource and provide recommendations for future catch limits (Methot and Wetzel 2013). Stock assessments rely on data collected from the fisheries harvesting the resource (fisheries-dependent data) and fisheries-independent surveys (e.g., bottom trawl surveys). Fisheries-independent surveys are often weighted highly in stock assessments owing to appropriate selection of the gear employed, consistency in where, when, and how sampling occur, and high availability of the stock to the survey (Board 2000; Chen et al. 2003; Pennino et al. 2016).

However, the costs associated with conducting fishery-independent surveys and processing the resulting biological samples can be substantial given the frequency and the broad spatial scale over which such surveys operate. In addition, the funding available to conduct fisheries-independent surveys is often fixed, creating a constraint on the amount of ship time available, tows that can be performed, and samples that can be collected and processed. Therefore, it is important to consider potential avenues for optimizing sampling efficiency across species of interest or areas of survey coverage. For example, surveys conducted in the Gulf of Alaska (GOA) and eastern Bering Sea (EBS) by National Oceanic and Atmospheric Administration (NOAA) Fisheries' Alaska Fisheries Science Center (AFSC) often involve multiple chartered vessels whereby tows are conducted across the survey area under a stratified random or systematic design (von Szalay and Raring 2018; Lauth et al. 2019). A wealth of information on catch, effort, biology, and life history are collected during these surveys, in addition to sampling of biological structures for ageing (e.g., otoliths).

A major question that is often asked in survey science is how to optimize the distribution of biological sampling efforts across species without increasing survey effort (number of ships, number of tows conducted, number of working days, etc.). For example, previous research on sampling protocols for Atlantic cod (*Gadus morhua*) on the North Sea International Bottom Trawl Survey conducted by the International Council for the Exploration of the Sea (ICES) has shown that the number of fish sampled for ageing could be reduced by at least 50% without a substantial loss of precision in the estimation of abundance-at-age (Jourdain et al. 2020). However, it is also important to quantify any potential changes in catch recommendation uncertainty that might result from a change in sampling effort, as well as the associated change in cost across alternative sampling treatments. Catch recommendations (e.g., for the acceptable biological catch, "ABC") are sometimes related to the perceived extent of uncertainty in estimates from stock assessments (Shertzer et al. 2008; NPFMC 2020). This is typically accomplished with the application of harvest control rules, which takes projected catch (i.e., the overfishing limit, "OFL") and applies a buffer that ensures the true OFL will not be exceeded given uncertainty in the assessment process. Thus, given information on the costs associated with processing biological samples, it is possible to summarize the effect of changes to sampling rates on costs and uncertainty in OFLs estimated across model fits employing replicate simulated data sets.

Aside from costs, changes in sampling effort may affect the uncertainty of data products and how those data are weighted in stock assessment models. Data products used in stock assessments have an influence on the outputs generated. It is therefore important to correctly weight the information content of those data to ensure that a good match exists between the variance of the data and that implied by the model (Francis 2017). Stewart and Hamel (2014) reviewed several approaches for setting the "sample size" of age-/length-composition data (the weight assigned in the stock assessment to each annual age-composition), which range from (1) using the number of fish sampled each year (nominal sample size); (2) setting a fixed value across all years; (3) taking the square root of the product of the number of samples and hauls; or (4) bootstrapping the data, calculating the variance among bootstrap replicates, and approximating this variance via the equivalent sample size for a multinomial distribution. Alternative estimators have been proposed, including stratified estimators (Miller and Skalski 2006), a normal approximation (Thorson 2014), and model-based expansion of compositional data (Thorson and Haltuch 2019), where the latter is analogous to propagating the joint imprecision in models that use a normal distribution for abundance-at-age (Berg and Nielsen 2016). Regardless of the method used for generating input sample sizes, these input sample sizes are typically either retained or iteratively tuned while fitting the model when a multinomial likelihood is employed for compositional data. In the tuning process, the estimated sample size that results based on the fit of the composition data (effective sample size) replaces the initial input sample size (Maunder 2011; Hulson et al. 2012). A variation of this tuning process (Francis re-weighting) involves calculating a weight based on expected and observed mean proportions-at-age, applying this to the original suite of input sample size, and re-fitting the model with this "re-weighted" set of values as the input sample size (eq. TA1.8 in Francis 2011). This re-weighting process is important for ensuring that the composition data are appropriately weighted (i.e., there is a good match of the variance between data and that implied by the model) relative to other data sets used for parameter estimation.

The main objectives of this study were to (1) develop a standardized framework (bootstrap estimator), which allows the user to explore the effect of changes in sampling rate on input sample size calculations and catch recommendation uncertainty and (2) associate a monetary cost to changes in otolith ageing efforts. The methods described below were applied to survey data for GOA stocks of walleye pollock (*Gadus chalcogrammus*), Pacific ocean perch (POP; *Sebastes alutus*), and dusky rockfish (*Sebastes variabilis*), which are surveyed by the AFSC using bottom trawl surveys and exploited by fisheries in the region. However, this framework could also, in principle, be applied to data from fishery-dependent monitoring.

Methods

A bootstrap estimator was developed to quantify the effect of changes to otolith sampling rates on input sample sizes. Input sample sizes and age-compositions generated using this estimator were then used in stock assessments to determine the influence that changes in sampling rates (based on changing the number of tows from which age samples are collected or the number of otoliths sampled per tow) have on uncertainty in catch recommendations. Finally, costs associated with processing and reading otoliths were then applied to the number of otoliths generated across sampling rate treatments to determine an average annual age-reading cost (in units US\$·year^{-1}) for each treatment (see Fig. 1 for a flow chart summarizing this framework).

In summary, this study design involved three species and two sampling methods, where each combination of species and method was analyzed using five sample-size "treatments". Each treatment was replicated 200 times (hereinafter termed "specimen bootstraps"), and each replicate involved 100 bootstraps of the expansion process (hereinafter termed "expansion bootstraps"), where the expansion bootstraps were used to calculate an input sample size for each specimen bootstrap replicate. This design therefore involved a total of 600 000 calculations of abundance-at-age, resulting in 6000 replicate assessments across species, sampling methods, and treatments.

Bootstrap estimator

We first explored how changes in otolith collection and age-reading effort affect the calculation of input sample sizes, which are used to quantify the variance in age-composition information. The bootstrap estimator used to calculate annual input sample sizes was based on an existing method for expanding survey length and age information developed by the AFSC. The expansion process uses annual, stratum-specific length information, survey catch-per-unit-effort, and stratum area to calculate annual stratum-specific numbers-at-length ("first-stage expansion"; see von Szalay and Raring 2018; Lauth et al. 2019 for details). For each year of data, the calculated numbers-at-length are summed across strata and converted into numbers-at-age using year-specific age–length keys constructed from specimens that have both age and length information recorded (second-stage expansion).

The bootstrap estimator developed for this study had two levels of bootstrapping (Fig. 2). The first bootstrapping process conducted (the "specimen bootstrap") was used to compute the uncertainty in the input samples sizes and the

Fig. 1. Flow chart depicting the approach taken to generate input sample sizes and age compositions from the bootstrap estimator, how these were used to conduct stock assessments and generate catch recommendations, and the associated catch recommendation uncertainty with ageing costs across sampling rate treatments.

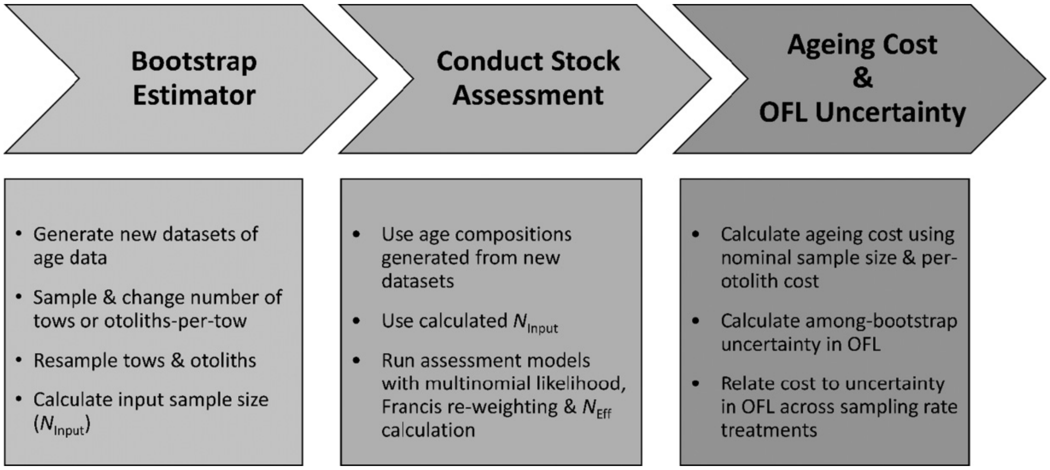
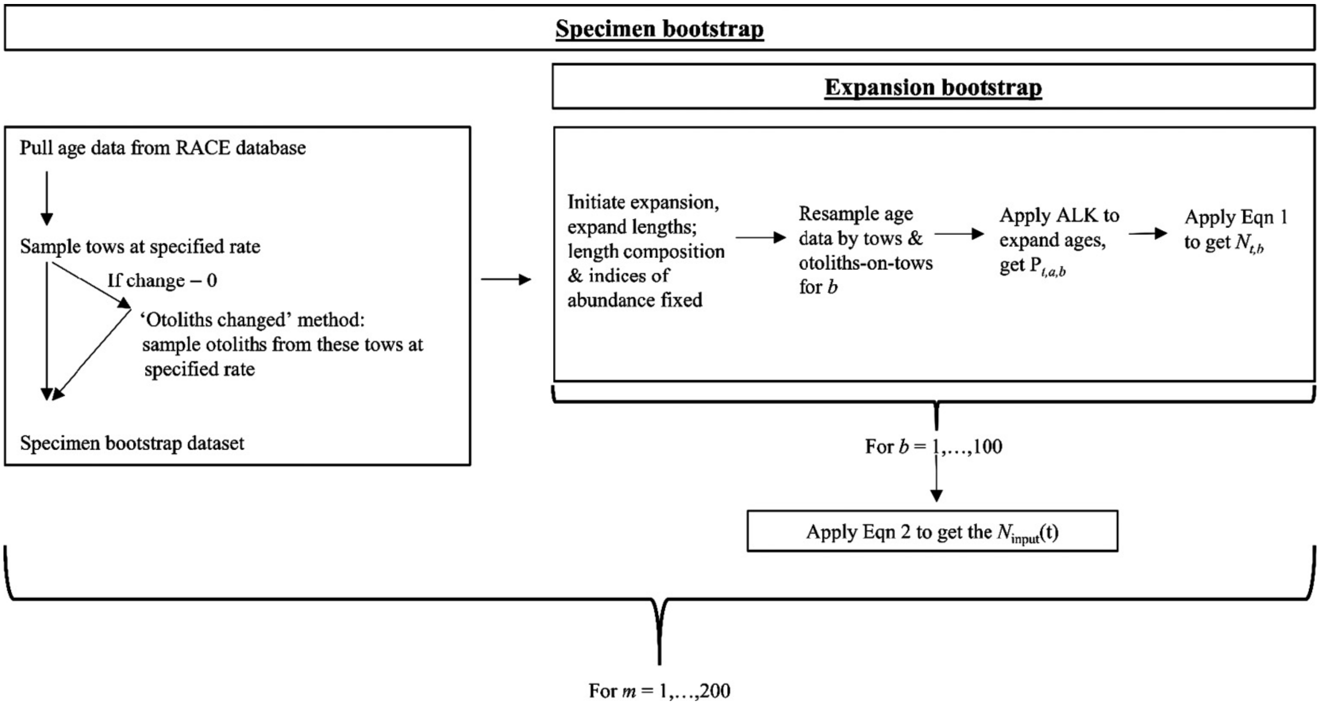

Bootstrap Estimator
Conduct Stock Assessment
Ageing Cost & OFL Uncertainty
• Generate new datasets of age data
• Sample & change number of tows or otoliths-per-tow
• Resample tows & otoliths
• Calculate input sample size (N_{Input})
• Use age compositions generated from new datasets
• Use calculated N_{Input}
• Run assessment models with multinomial likelihood, Francis re-weighting & N_{Eff} calculation
• Calculate ageing cost using nominal sample size & per-otolith cost
• Calculate among-bootstrap uncertainty in OFL
• Relate cost to uncertainty in OFL across sampling rate treatments

Fig. 2. Flow chart describing a single iteration of the bootstrap estimator, including the different processes of the "specimen bootstrap" and the "expansion bootstrap". The bootstrap estimator was used to generate 200 data sets for each sampling treatment. RACE refers to the Resource Assessment and Conservation Engineering Division of the Alaska Fisheries Science Center (NOAA Fisheries).


Specimen bootstrap
Expansion bootstrap
Pull age data from RACE database
Sample tows at specified rate
If change – 0
'Otoliths changed' method: sample otoliths from these tows at specified rate
Specimen bootstrap dataset
Initiate expansion, expand lengths; length composition & indices of abundance fixed
Resample age data by tows & otoliths-on-tows for b
Apply ALK to expand ages, get $P_{t,a,b}$
Apply Eqn 1 to get $N_{t,b}$
For $b = 1,\ldots,100$
Apply Eqn 2 to get the $N_{\mathrm{input}}(\mathrm{t})$
For $m = 1,\ldots,200$

survey age-compositions. The second bootstrapping process conducted (the "expansion bootstrap") was used to compute annual input sizes given a data set of length-composition, resampled specimen data on age and lengths generated from the "specimen bootstrap" (i.e., age–length keys), and the resulting age compositions constructed for each "expansion bootstrap" replicate. The survey abundance indices and length-compositions were not varied in either of these bootstrapping steps so their uncertainty is not accounted for.

The "specimen bootstrap" facilitated an evaluation of how changes to sampling rates impacted the calculated annual input sample sizes. This bootstrapping process involved sampling the original specimen age data while making two potential changes (termed "sampling methods"):

1. "Tows changed"—Increasing or decreasing the number of tows that otoliths were sampled on, corresponding to a hypothetical increase or decrease in funding for total sur-

vey days at sea (e.g., owing to less time spent sorting the tow catch or less time performing shipboard otolith dissections). This bootstrapping method addressed between-tow variability; or

2. "Otoliths changed"—Increasing or decreasing the number of sampled otoliths per tow, corresponding to a hypothetical increase or decrease in funding or speed for processing otoliths in the lab after the survey is completed. This bootstrapping method addressed within-tow variability.

For reference, otolith collections on the AFSC bottom trawl survey in the GOA follow a random sampling design except for some species not included in this analysis. Across the western, central, and eastern regions of the stock area, up to 10 dusky rockfish and five Pacific ocean perch are collected per tow. For walleye pollock, none are taken if fewer than 10 fish are caught, five fish are collected if 10–500 fish are caught, and 10 fish are taken if over 500 are caught (RACE 2021).

For each species of interest and sampling method, five sampling treatments were employed to change the level of otolith sampling by 0%, ±33%, and ±67%. For each sampling method, the specimen bootstrapping process cycled through every year of survey data, recorded the unique tow identifiers for each year, and sampled from them. Under the "tows changed" method, the tow IDs were re-sampled with replacement at the designated rate, and the otolith-derived age information associated with the selected tow IDs was included in the new data set. For the "otoliths changed" method, the number of tows sampling otoliths was not changed, but the tow IDs were resampled with replacement. Under this method, an additional step was conducted whereby every tow ID from the resampled selection was cycled through, and the number of otoliths on those tows was changed by the designated rate. This "specimen bootstrap" was conducted 200 times to create 200 new realizations of age data sets for each treatment, method, and species (denoted *m* in Fig. 2).

The "expansion bootstrap" involved resampling each of the 200 bootstrapped specimen data sets described above to generate 100 age–length keys per "specimen bootstrap" replicate, and then applying those age–length keys to the length-compositions to create 100 unique matrices of survey proportions-at-age (i.e., age compositions) per "specimen bootstrap" replicate. The purpose of this bootstrap was to generate replicated compositions to use in the calculation of N_{Input}. This bootstrap involved resampling tows from the available set of tows and otoliths within tows while retaining the sample sizes of tows or otoliths within tows. The expanded age-composition information generated during each of the 100 "expansion bootstrap" replicates was converted into proportions-at-age (i.e., 100 proportions-at-age matrices). The input sample size for a survey year *t*, $N_{\text{Input}}(t)$, was calculated as a function of (1) the resampled proportions-at-age by "expansion bootstrap" replicate ($P_{t,a,b}$; where *b* denotes the expansion bootstrap and *a* denotes age) compared to the mean proportions-at-age across the 100 "expansion bootstrap" replicates ($\widehat{P}_{t,a}$; eq. 1) and (2) the harmonic mean across bootstrap replicates (eq. 2), which has been found to be relatively unbiased (McAllister and Ianelli 1997; Stewart and Hamel 2014):

$$(1)\qquad N_{t,b} = \frac{\sum_{a=1} P_{t,a,b} \times \left(1 - P_{t,a,b}\right)}{\sum_{a=1} \left(P_{t,a,b} - \widehat{P}_{t,a}\right)^{2}}$$

$$(2)\qquad N_{\text{Input}}\left(t\right) = \left(\frac{\sum_{b=1} N_{t,b}^{-1}}{100}\right)^{-1}$$

The calculated input sample sizes, total number of otoliths sampled (i.e., nominal sample size), and number of tows conducted were recorded across bootstrapping methods and sampling rate treatments. In summary, this process generated 200 vectors of annual input sample sizes and 200 age-composition matrices per treatment, method, and species; together these data products were used in replicate assessment model runs (described below). To justify the number of "specimen bootstrap" replicates conducted, we ran additional "specimen bootstrap" replicates for dusky rockfish under the "tows changed" scenario for each sampling rate treatment and calculated N_{Input} when 100, 200, 300, 400, and 500 "specimen bootstrap" replicates were included. As seen in Table A1, calculated N_{Input} was stable across trials that used an increasing number of specimen bootstrap replicates. In addition, to justify the number of "expansion bootstrap" replicates we increased the number of "expansion bootstrap" replicates from 100 to 1000 for a single "specimen bootstrap" replicate for dusky rockfish across the five sampling rate treatments under the "tows changed" sampling scenario. The results of calculating N_{Input} when utilizing 100 (original), 500, and 1000 replicates of age composition (Table A2) show that N_{Input} was stable across trials, justifying the use of only 100 expansion bootstrap replicates for each replicate of specimen bootstraps.

We then calculated the slope of a log–log relationship between (*i*) input sample size and the number of otoliths (nominal sample size, "N_{Nominal}") or (*ii*) input sample size and the number of tows (Tows) using linear mixed models using the glmmTMB package in R (version 1.1.2; Brooks et al. 2017). Mixed-effects models were used for this analysis owing to the nature of the data, whereby multiple measurements of sample size taken over time are considered, suggesting some degree of non-independence. These slopes were interpreted as the expected percent change in N_{Input} per percent change in nominal sample size or number of tows for a given sampling method (i.e., "tows changed" vs. "otoliths changed"). The relationship between input sample size and nominal sample size was defined as

$$(3)\qquad \log\left[N_{\text{input}}\left(t,m\right)\right] = \alpha\left(t\right) + \beta\left(t\right) \times \log\left[N_{\text{nominal}}\left(t,m\right)\right] + \varepsilon\left(t,m\right)$$

where $\varepsilon\left(t,m\right) \sim \text{Normal}\left(0,\sigma^2_{\text{residual}}\right)$ is the residual for year *t* and specimen bootstrap *m* with variance $\sigma^2_{\text{residual}}$. The intercept for each year, $\alpha(t)$, is treated as a random effect, with mean α_0 and among-year variance σ^2_{α}, i.e., $\alpha\left(t\right) \sim \text{Normal}\left(\alpha_0, \sigma^2_{\alpha}\right)$. The slope for each year, $\beta(t)$, is also treated as a random effect with mean β_0 and among-year variance σ_β^2, i.e., $\beta(t) \sim \text{Normal}\left(\beta_0, \sigma_\beta^2\right)$. For each species, we then recorded the average slope β_0 as well as shrunk estimates of slope for each year $\beta(t)$ to calculate the log–log sensitivity to changes in the nominal sample size. We then repeated this using $\log(N_{\text{tows}}(t,m))$ in place of $\log(N_{\text{nominal}}(t,m))$ and recorded the log–log sensitivity to changes in the number of tows. Additional models were fit to the data from each year separately, and the results of these model fits are reported below.

Estimation of catch recommendation uncertainty

We next explored how changing input sample sizes affects data weighting. Effective sample sizes (N_{Eff}) were computed assuming that the age-composition data followed a multinomial distribution and using the Francis approach to data weighting (Francis 2011):

$$N_{\text{Eff}}(t) = N_{\text{input}}(t) \times w \tag{4}$$

where w was the weight calculated using eq. TA1.8 from Francis (2011). The process of fitting the assessment model followed by application of eq. 4 was repeated three times.

Stock assessments were conducted using the proportions-at-age and input sample sizes from each of the 200 "specimen bootstrap" replicates. The three stocks included in this study are Tier 3 stocks under the North Pacific Fishery Management Council's fishery management plan for the GOA stock area. This means that reliable estimates of the spawner–recruit relationship are not available, but proxies for B_{MSY} and F_{MSY} (i.e., $B_{40\%}$, $F_{35\%}$, and $F_{40\%}$) can be estimated (NPFMC 2020). Consequently, the fishing mortality used to compute the OFL is set according to the following equation:

$$F_{\text{OFL},y} = \left\{\begin{array}{ll} 0 & \text{if } B_y/B_{40\%} < 0.05 \\ F_{35\%}\dfrac{(B_y/B_{40\%} - 0.05)}{(1-0.05)} & \text{if } 0.05 \le B_y/B_{40\%} < 1 \\ F_{35\%} & \text{if } B_y/B_{40\%} > 1 \end{array}\right\} \tag{5}$$

where $F_{35\%}$ is the fully selected fishing mortality corresponding to a 65% reduction in spawning biomass-per-recruit, $B_{40\%}$ is 40% of the estimate of unfished spawning biomass, and B_y is the spawning biomass in year y. The OFL for year y is then set using eq. 6:

$$\text{OFL}_y = \sum_a W_a \frac{F_{\text{OFL},y*}\, S_{y*,a}}{Z_{y*,a}} N_{y*,a}\left(1 - e^{-Z_{y*,a}}\right) \tag{6}$$

where $N_{y*,a}$ is the estimate of the number of fish of age a at the start of the year for which an OFL is needed ($y*$), $Z_{y*,a}$ is the total mortality-at-age (which depends on $F_{\text{OFL},y*}$), W_a is the weight-at-age, and $S_{y*,a}$ is selectivity-at-age for year y.

Following the model-fitting procedure, distributions of OFL were generated from the model output. The standard deviation and coefficient of variation of the OFL estimates (SD OFL and CV OFL) were calculated across replicates using the mean OFL from the 0% treatment to capture the actual change in uncertainty that might result from the various sampling rate treatments. Calculating SD OFL and CV OFL in this manner allowed for the among-bootstrap variance in OFL to be determined for each sampling rate treatment relative to the status quo OFL, which in turn allowed quantification of the costs and benefits of reducing or increasing data in the currency of uncertainty instead of the currency of uncertainty relative to a shifting estimate of OFL. US Fishery Management Councils have adopted formal methods to account for scientific uncertainty when making catch recommendations. One common method is the "*P*-star (*P*∗) approach" (Shertzer et al. 2008; Ralston et al. 2011), whereby (*i*) a level of uncertainty in the OFL or terminal year biomass is estimated or specified (e.g., a "sigma" value), (*ii*) a *P*∗ value is specified to represent the acceptable probability of exceeding an OFL due to scientific uncertainty, and (*iii*) a form is assumed for a probability density function to capture the scientific uncertainty in the OFL (e.g., lognormal). Together, these assumptions are used to develop a buffer between the OFL and the ABC. Therefore, while our method for calculating OFL uncertainty based on variance across bootstrap replicates is specific to this study, it could be used to inform the specified sigma value when model-estimated uncertainty is unavailable.

Levene's test for homogeneity of variance (Levene 1960) was used to test for significant differences in the variance of OFL distributions between sequential pairs of sample-size treatments (e.g., variance of OFL distributions between the −67% vs. −33% treatment, −33% vs. 0% treatment, etc.), as this test is less sensitive to departures from normality than Bartlett's test (Bartlett 1937). A significant difference in the variance of OFL distributions between two treatments would suggest that the change in otolith sampling led to input sample sizes (i.e., initial weightings) and age compositions that had a statistically significant effect on the uncertainty related to OFL estimates.

In addition, we calculated the slope of a log–log relationship between the CV of OFL and the number of otoliths (nominal sample size, N_{Nominal}) using linear mixed models similar to eq. 3 above, but replacing CV OFL for N_{Input}. Here, owing to the nature of CV OFL estimates (i.e., a single value per treatment instead of annualized estimates), year was removed from the analysis. These slopes were interpreted as the expected percent change in CV OFL per percent change in nominal sample size for a given sampling method (i.e., "tows changed" vs. "otoliths changed"). This analysis was conducted to summarize the effect of changes in otolith sampling on uncertainty in catch recommendations.

Analysis of ageing cost

Costs and efforts associated with ageing species surveyed by the AFSC are summarized by Lambert et al. (2017). In the current study, this information was used in conjunction with AFSC full-time employee (FTE) salary rates calculated per 8 h day (including indirect costs) to determine the cost associated with ageing the number of otoliths sampled across the designated sampling rate treatments. Based on Lambert et al. (2017), an age reader can age 8.1 dusky rockfish, 12.5 Pacific ocean perch, or 27.2 walleye pollock otoliths per day on average. This equates to 1.0, 1.6, and 3.4 otoliths per hour for dusky rockfish, Pacific ocean perch, and walleye pollock. Dividing FTE salary-per-day by otoliths-per-day resulted in a cost of $62.50, $40.50, and $18.61 per otolith, respectively. A limitation of this cost function is that it does not incorporate costs associated with ship time (e.g., fuel, charter days, personnel time, or overtime for the ship crew and science crew, etc.). We therefore assume that the underlying costs of the survey are constant given that the total number of survey tows must remain the same. However, a scenario where per-tow otolith collections are reduced and otoliths are collected on more of the tows conducted could be evaluated if needed.

A cost relationship was created for each species based on the salary of AFSC FTEs and the number of otoliths that can be aged each day on average from 2008 to 2016. To calculate the number of otoliths that the per-otolith cost would be applied to, for each "expansion bootstrap" replicate, the total nominal sample size was first summed across years the survey was conducted (survey years) and then divided by the number of survey years (i.e., number of otoliths collected per survey year). Then, these values were averaged across "specimen bootstraps" and multiplied by the per-otolith cost. These costs were calculated for each sampling rate treatment and sampling method.

Example applications

The methods described above were applied to GOA stocks of walleye pollock, Pacific ocean perch, and dusky rockfish, which are surveyed by the AFSC using bottom trawl surveys and exploited by fisheries in the region. The three example stocks, respectively, capture a relative range of data-rich (walleye pollock, $N = 18\,353$ otolith ages from 1990 to 2019; ABC = 105 770 t, CV ABC = 0.21), to data-moderate (Pacific ocean perch, $N = 16\,984$ from 1990 to 2019; ABC = 36 164 t, CV ABC = 0.35), and data-poor stocks (dusky rockfish, $N = 9752$ otolith ages from 1984 to 2019; ABC = 7097 t, CV ABC = 0.31). However, we acknowledge that all of these stocks may be considered data-rich if compared to all stocks worldwide. Here, estimates of ABC and CV ABC provided were taken from recent stock assessments (2020) to characterize the level of uncertainty in catch recommendations. These otolith sample numbers are often reduced in the process used to expand age information for survey area and catch-per-unit-effort owing to mismatches in stratum-specific age–length keys (i.e., paired age and length data for individuals collected for otolith extraction in each stratum) employed to convert expanded length information to expanded age information (2%–30% reduction across all survey years dependent upon stock). This is largely due to the nature of data collection on each tow, where length frequency information is collected from one group of individuals, otoliths are dissected from a different group of individuals, and length data are not merged between the two sources.

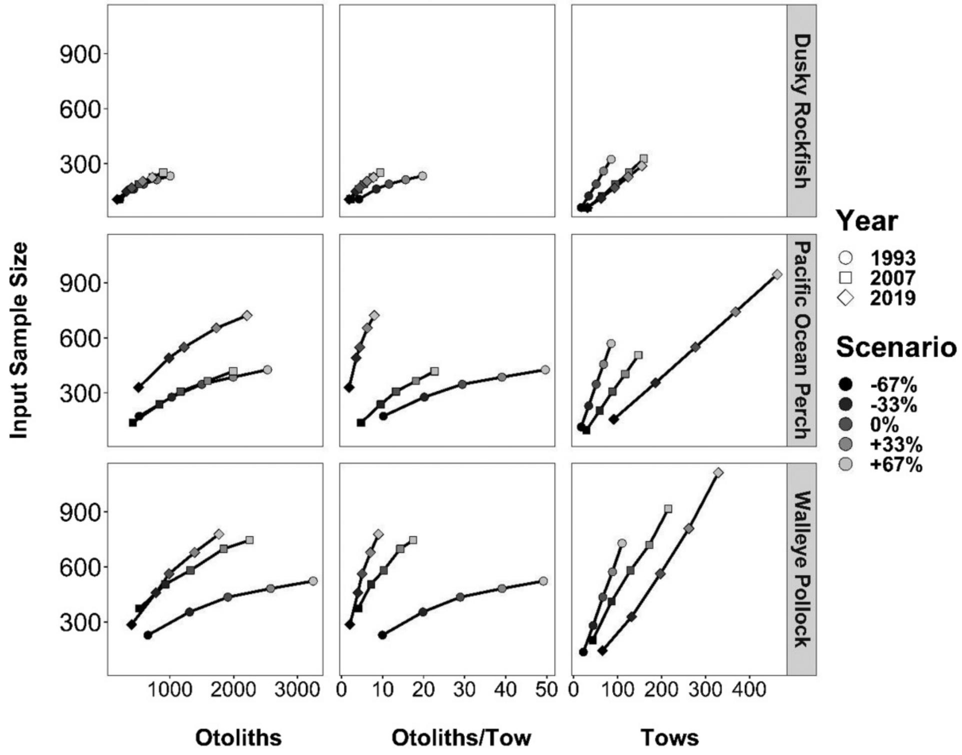
GOA stock assessments for the three species (Dorn et al. 2020; Fenske et al. 2020; Hulson et al. 2020 for dusky rockfish, Pacific ocean perch, and walleye pollock, respectively) are conducted using a statistical age-structured model developed in AD Model Builder (Fournier et al. 2012), which consists of (*i*) a population dynamics model fit to survey and fishery data (Table A3) that calculates population estimates across time and (*ii*) a projection model that predicts future population estimates based on results from the population model, resulting in biological reference points (e.g., ABC, OFL).

Input sample sizes are currently used in these stock assessments assuming a multinomial likelihood for length- and age-compositions (referred to as "multinomial input sample sizes" herein). For dusky rockfish, these annual input sample sizes are calculated in the stock assessment as the square root of the product of the number of samples collected and the number of tows conducted, scaled to a maximum of 100. For Pacific ocean perch, the annual input sample sizes are calculated as the square root of the number of samples collected. For walleye pollock, a fixed starting value of 60 is used across years and iteratively tuned based on the Francis reweighting method (eq. TA1.8 in Francis 2011).

In the current study, input sample sizes generated by the bootstrap estimator were used in assessment model replicates to set starting weights. For all three stocks, the multinomial likelihood was employed along with the Francis approach for data weighting (three iterations) to calculate $N_{\mathrm{Eff}}(t)$ (eq. 4). We note that although stock assessments often assume a multinomial likelihood for compositional data, this assumption is typically violated. To illustrate this, we conducted an analysis to compare the covariance of age compositions generated by the bootstrap estimator to the covariance of a multinomial distribution for a randomly chosen data set. We chose a single species-treatment combination (dusky rockfish and the 0% treatment) and calculated the covariance matrix for the proportion-at-age of a single year (2019) across bootstrap replicates. To calculate the covariance of a multinomial distribution, we set the sample size equal to the average N_{Input} across bootstrap replicate for 2019 ($N_{\mathrm{Input}} = 169$) and the probability distribution equal to the actual proportions-at-age. The covariance of the bootstrapped age compositions does not exactly match the multinomial covariance: the diagonals do not substantially differ, but the covariance of the bootstrapped compositions has positive values in the off-diagonals and the pattern of covariances in the off-diagonals is not smooth as it is for the multinomial covariances (Fig. A1). While this is a pertinent topic for future research, we think that it is more appropriate if we followed standard practices currently employed at the institution that conducts the surveys and stock assessments included in our study (i.e., assuming that the sampling variance is multinomial).

For comparison to the distributions of OFL, reference model runs were conducted for each species using the median annual input sample sizes over the 200 "specimen bootstraps" for the 0% sampling rate treatment (no change in tows or otoliths-per-tow), age compositions calculated from original survey data (no resampling), and the multinomial likelihood (with Francis re-weighting; eq. TA1.8 in Francis 2011). The estimates of OFL that were generated from these model runs were compared to the distributions of OFL from each "specimen bootstrap", with the goal of gauging the influence of using the $N_{\mathrm{Input}}(t)$ from this study in actual assessments.

Fig. 3. Input sample sizes calculated using the bootstrap estimator related to the number of otoliths collected (Otoliths), otoliths collected per tow (Otoliths/Tow) and tows conducted (Tows), averaged across bootstrap replicates for each sampling rate treatment. Results reported in the "Otoliths" and "Otoliths/Tow" columns correspond to the "otoliths changed" sampling process employed in the bootstrap estimator, while the results reported in the "Tows" column correspond to "tows changed" sampling process. Results shown are for the years 1993 (circles), 2007 (squares), and 2019 (diamonds).



Results

Bootstrap estimator

Input sample sizes (calculated using eqs. 1 and 2) increased as the number of otoliths, number of otoliths-within-tows, or the number of tows sampled for otoliths increased (Fig. 3, showing example years, 1993, 2007, and 2019). As designed, the number of tows did not change under the "otoliths changed" sampled method. However, otoliths-per-tow changed slightly under the "tows changed" sampling method owing to bootstrapped sampling of tows with varying numbers of otoliths. There was also variability in the calculated input sample sizes across years related to the variability in nominal sample size and number of tows conducted, which likely reflects species availability to the survey and changes to sampling protocols across time (Fig. 3).

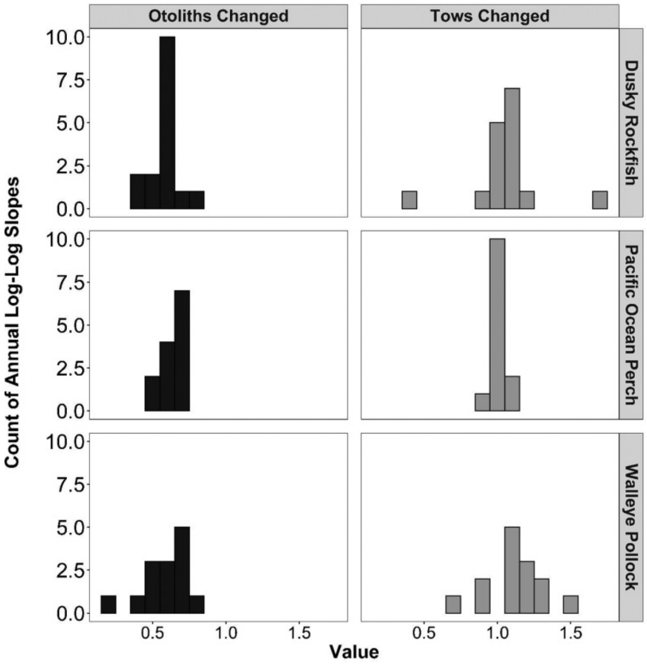
Using linear mixed models, the relationship between (*i*) input sample size and nominal sample size (NomSS) and (*ii*) input sample size and number of tows (Tows) was described on a log–log scale. The year-specific log–log slopes of these relationships (Table A4; Fig. 4) were generally higher and more variable under the "tows changed" method (mean = 1.065; range = 0.379–1.708 across species) compared to the "otoliths changed" method (mean = 0.599; range = 0.214–0.808 across species). This also held true under the summary models where year was treated as a random effect, with summary model slopes higher for the "tows changed" method (mean = 1.065; range = 1.025–1.114 across species) compared to the "otoliths changed" method (mean = 0.601; range = 0.570–0.640 across species). This suggests that there is a larger impact on calculated N_{Input} when the number of tows is changed relative to when the number of otoliths-per-tow is changed.

Estimation of catch recommendation uncertainty

Stock assessment model runs that employed the bootstrap replicates of input sample size and age composition generated distributions of OFL (Fig. 5) for each sampling rate treatment (Table 1). In general, the distributions of OFL estimates overlapped (Fig. 5), but the OFL distributions resulting from reduced sample size treatments were broader (i.e., more variability among experimental replicates) than the 0% treatment.

Overall, the mean OFL for walleye pollock declined as sampling rate was reduced from +67% to −67% across the "otoliths changed" (4650 t decline) and "tows changed" (5277 t decline) bootstrap sampling methods (Table 1), indicating some conflict among the data sources included in the assessment. However, under the "tows changed" method, most of these declines occurred when sample sizes were decreased from the 0% treatment to the −67% treatment (3350 t), while mean OFLs estimated using data from the increased sampling treatments did not change as much

Fig. 4. Slopes of the log–log relationships between (*i*) input sample size and number of otoliths sampled ("otoliths changed"; eq. 3; black, left column) and (*ii*) input sample size and number of tows conducted ("tows changed"; eq. 3; grey, right column). Log–log slope is interpreted as the expected percent change in input sample size per percent change in the number of otoliths processed or number of tows conducted.



(1310 t). Under the "otoliths changed" method, changes in mean OFL were similar regardless of the directionality of sampling rate changes (2625 t increase as sampling rate increased; 2652 t decrease as sampling rate decreased). Uncertainty in OFL estimates calculated across bootstrap replicate model runs generally declined as sampling rate was increased from the −67% treatment to the 0% treatment regardless of sampling method employed (Fig. 6). However, uncertainty in the OFL remained relatively unchanged across models runs that used data from the 0%, +33%, and +67% treatments under the "otoliths changed" sampling method. Levene's tests conducted on sequential pairs of OFL distributions suggested that significant differences in variance existed between the −67% and −33% distributions of OFL under the "otoliths changed" method and between the 0% and +33%, and the +33% and +67% distributions of OFL under the "tows changed" method (Table A5).

In comparison, mean OFL for Pacific ocean perch did not change much across sampling treatments under either sampling method (max. difference = 540 t for the "otoliths changed" method; max. difference = 1473 t for the "tows changed" method; Table 1). However, there was a clear decline in the uncertainty of the OFL as sampling rate increased under the "tows changed" method. Under the "otoliths changed" method, this pattern was not as marked but remained. Levene's test conducted on sequential pairs of OFL distributions resulted in significant differences in variance between the 0% and +33% treatments under the "otoliths changed" method and between the −67% vs. −33% and 0% vs. +33% treatments under the "tows changed" method (Table A5).

For dusky rockfish, mean OFL changed very little across sampling treatments under both sampling methods (max. difference = 59 t under the "otoliths changed" method; max. difference = 77 t under the "tows changed" method; Table 1). However, trends in OFL uncertainty were similar to those observed for Pacific ocean perch under the "tows changed" sampling method, with a clear decline in SD OFL and CV OFL as sampling rate increased (Fig. 6). Levene's test conducted on sequential pairs of OFL distributions suggested that a significant difference in variance existed between the 0% and +33% treatments under both sampling methods and between the −67% and −33% treatments under the "tows changed" sampling method (Table A5).

Using linear mixed models, the relationship between CV of OFL and nominal sample size (N_{Nominal}) was described on a log–log scale. The log–log slopes of these relationships (Table 2) were more negative under the "tows changed" method for dusky rockfish and Pacific ocean perch (range = −0.576 to −0.486 across species) compared to the "otoliths changed" method (range = −0.238 to −0.029 across species). This suggests that for these species there is a larger reduction in estimated CV of OFL as the number of tows is increased from the −67% to the +67% treatment relative to when the number of otoliths-per-tow is changed.

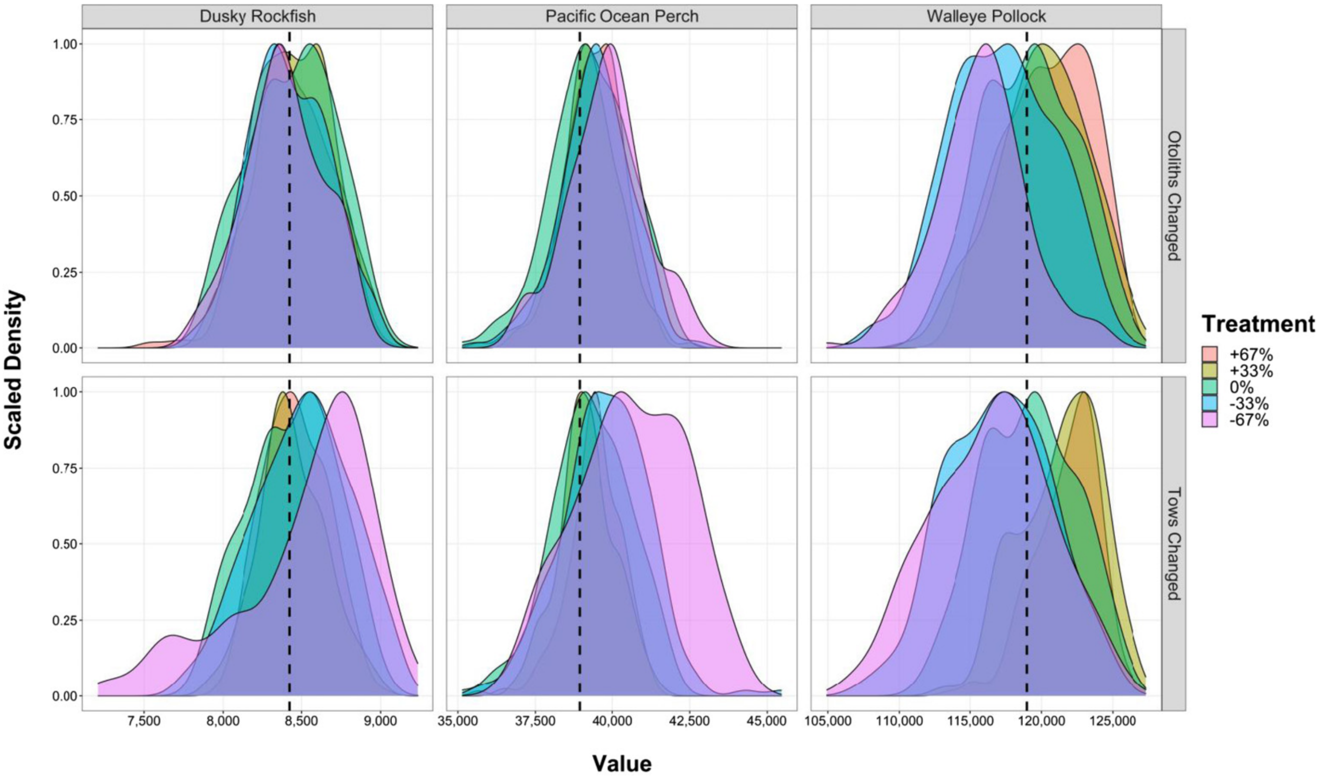
The relationship between effective sample size (after three iterations of Francis reweighting) and input sample size is reported in Table 1 and Fig. A2. This relationship captures the effect of changes to sampling on variance in bootstrapped age compositions and subsequently how the assessment model perceives the contribution of age compositions. From Fig. A2, it is clear that changes to sampling of walleye pollock affect the calculated input sample size but have no effect on the (model calculated) effective sample size. In contrast, dusky rockfish and Pacific ocean perch display a relationship of diminishing returns on information content (effective sample size) as sampling is increased under the "tows changed" sampling method.

Analysis of ageing cost

Per-otolith costs defined above were multiplied by the nominal sample size per survey year (N_{Nominal}/year) averaged across specimen bootstrap replicates to calculate total ageing costs per survey year (US$·year^{-1}) associated with each sample-size treatment. In the interest of comparison to the 0% treatment, costs were calculated as a change from the 0% treatment. These costs increased as sampling increased, but the cost per otolith varied among species based on the number of otoliths that can be aged per day (Table A6).

For dusky rockfish, the calculated change in cost ranged from US$−16 435 to US$+19 446·year^{-1} across increasing sampling rate treatments under the "otoliths changed" sampling method, and from US$−17.965 to US$+17.813·year^{-1} across treatments under the "tows changed" method. For Pacific ocean perch, the calculated change in costs ranged

Fig. 5. Distributions of overfishing limit (OFL) estimated for 2021 in assessment model runs that used data from the "otoliths changed" and "tows changed" sampling methods. Treatment refers to sets of runs that used data from each sampling rate treatment (±0%, ±33%, ±67%). The black line denotes a value of OFL calculated using the median input sample size across specimen bootstrap replicates for the 0% treatment, non-bootstrapped age compositions, and a multinomial likelihood (with Francis re-weighting).



from US$−26 158 to US$+29 930·year^{-1} across increasing sampling rate treatments under the "otoliths changed" sampling method, and from US$−27 728 to US$+27 454·year^{-1} across treatments under the "tows changed" method. For walleye pollock, the calculated change in costs ranged from US$−14 773 to US$+17 272·year^{-1} across increasing sample rate treatments under the "otoliths changed" sampling method, and from US$−16 023 to US$+15 917·year^{-1} across treatments under the "tows changed" method.

Discussion

In this study, we first demonstrated how existing compositional expansion methods can be used in a standardized framework to generate input sample sizes for use as compositional weights in stock assessments. In addition, we demonstrated how the increase of survey tows (from −67% to +67% treatments) that provided otoliths for ageing on bottom trawl surveys tends to have a larger impact than changes in the number of otoliths per tow on both (*i*) the input sample sizes and (*ii*) the uncertainty in estimates of OFL that result from using them for weighting age-composition information in stock assessments. To fully capture the uncertainty and importance of age information, this suggests that otolith collections should be distributed across as many tows as possible. This finding is consistent with previous work by Pennington et al. (2000), who recommended reduced tow durations to accommodate more stations on trawl surveys. Finally, by linking realizations of nominal sample size-per-survey year to ageing costs-per-otolith, we have also demonstrated that potential tow-based changes to otolith sampling could affect survey costs more than otolith-based changes to sampling.

In any examination of sampling efficiency where the sampling process is part of a larger scientific enterprise (i.e., sample collection, data processing, assessment model fitting, management advice, or satisfied industry partners), it is important to clearly define what an "optimal scenario" would be. However, the challenge herein is that the definition of optimal might be different for each partner to that enterprise. To a survey team collecting data, an optimal scenario might be one where sampling designs are efficient enough to satisfy logistical constraints but robust enough to ensure that the data products they provide are of high quality. To an assessment scientist, an optimal scenario might be one where uncertainty in parameter estimates and catch advice is minimized. To a fisheries manager, an optimal scenario might be one where the stock remains healthy with minimized foregone yield, while for a member of the fishing industry, an optimal scenario is likely one where revenue is maximized. Here, we have developed a generalized framework where survey and assessment scientists can explore sampling scenarios and the effect that those scenarios have on catch recom-

Table 1. Overfishing limit (OFL) and standard deviation estimated for 2021 averaged across assessment model runs that utilized input sample sizes and age compositions generated using the bootstrap estimator ($m = 200$ “specimen bootstrap” replicates per sampling treatment).

Method	Species	Sampling treatment	N_{Nominal}	N_{Input}	N_{Eff}	OFL (t)	OFL SD (t)
Otoliths changed	Dusky rockfish	−67%	163	73	10	8 392	255
		−33%	311	109	10	8 424	248
		0%	426	130	9	8 441	273
		+33%	587	153	9	8 451	240
		+67%	738	172	9	8 435	243
	Pacific ocean perch	−67%	375	152	64	39 818	1 371
		−33%	740	247	67	39 462	1 154
		0%	1 021	302	68	39 315	1 275
		+33%	1 391	356	65	39 278	1 022
		+67%	1 760	397	66	39 436	913
	Walleye pollock	−67%	493	304	6	115 801	4 663
		−33%	927	457	6	117 204	4 129
		0%	1 287	569	7	119 141	3 253
		+33%	1 778	668	7	120 019	3 317
		+67%	2 215	749	8	120 451	3 248
Tows changed	Dusky rockfish	−67%	139	44	7	8 518	443
		−33%	286	86	9	8 507	310
		0%	426	130	9	8 441	273
		+33%	565	175	10	8 441	214
		+67%	711	223	10	8 461	203
	Pacific ocean perch	−67%	337	96	52	40 626	2 192
		−33%	684	200	62	39 760	1 487
		0%	1 021	302	68	39 315	1 275
		+33%	1 351	400	69	39 153	964
		+67%	1 699	507	69	39 440	862
	Walleye pollock	−67%	426	159	6	116 489	4 905
		−33%	858	354	6	116 988	4 241
		0%	1 287	569	7	119 141	3 253
		+33%	1 710	782	8	121 319	3 546
		+67%	2 142	1 034	10	121 766	3 459

Note: Method refers to whether the number of otoliths on selected tows (“otoliths changed”) or number of tows sampled were changed (tows changed). N_{Nominal} refers to the mean nominal sample size per year across “specimen bootstraps”, while N_{Input} and N_{Eff} refer to the mean input and effective sample size across “specimen bootstraps” and years.

mendation uncertainty. This work will help these interdisciplinary teams examine trade-offs and make informed judgements on the level of sampling required to provide high-quality data products that efficiently inform assessment models on the importance of their data products, and that result in acceptable levels of uncertainty in catch advice.

We define an optimal scenario as one where sampling can be reduced but not to a degree where a significant increase in OFL uncertainty occurs. For example, in this study a reduction in sampling by 33% resulted in no significant difference in OFL uncertainty regardless of the sampling method used or stock (Fig. 6; Table A5). Therefore, a change in sample intensity of this magnitude would afford additional sampling effort (i.e., surplus sampling effort) that could be applied to a different stock that would benefit from an increase in sampling in terms of reduced OFL uncertainty (e.g., dusky rockfish or Pacific ocean perch; Fig. 6; Table A5). An analyst could use our framework to identify potential stocks that would benefit from this surplus sampling effort. However, the precision of other model inputs is important to consider in this context. Alternatively, we note that there are species with very little age data relative to those used in this study, and it could be worthwhile to allocate sampling effort gained from the results of an analysis such as this to those species. Not all of these species might need to be aged owing to relatively lower fishery importance, lower ecosystem importance, or lack of an age-structured assessment model currently in place. However, this status could change in the future, and the framework developed here would help analysts weigh the trade-offs of a new sampling endeavor.

Owing to the logistical constraints of bottom trawl surveys (i.e., set funding, set number of tows, etc.), it is easier to make changes to the number of otoliths collected and read for a species on each tow. However, marine fishes often form aggregations with similar ages and lengths such that ages and lengths are correlated for all fish in a given tow. Consequently, increasing the number of tows from which otoliths are collected for a given species would have a greater impact

Fig. 6. Standard deviation of overfishing limit (SD OFL) and coefficient of variation (CV OFL) for each species (columns) and sampling rate treatment (±%, *x*-axis). The mean OFL from the 0% treatment was used in calculation of the SD OFL and CV OFL. Whiskers show 95% confidence intervals. Results are based on data generated from the "otoliths changed" (black) and "tows changed" (grey) methods. Note: circles in the middle of a line corresponding to the 0% change treatment match exactly.

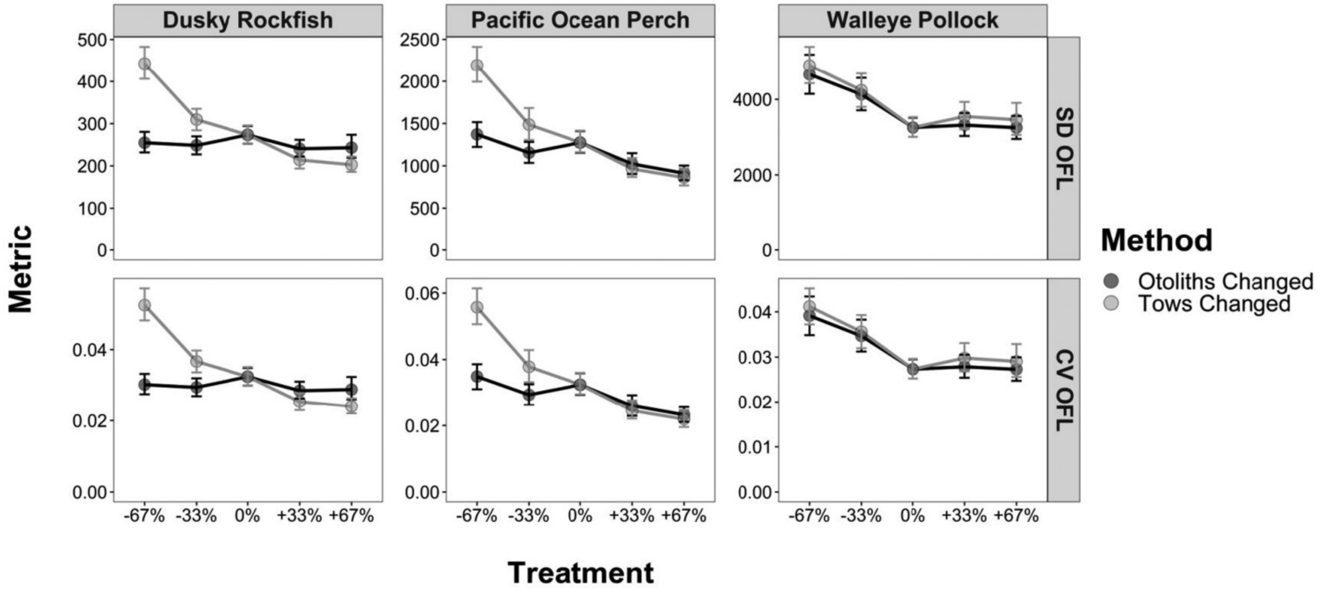


Table 2. Slope of the log–log relationship between the coefficient of variation for overfishing limit (CV OFL) and nominal sample size across sampling treatments for each species and sampling method.

Species	Method	Slope
Dusky rockfish	Otoliths changed	−0.029
	Tows changed	−0.486
Pacific ocean perch	Otoliths changed	−0.238
	Tows changed	−0.576
Walleye pollock	Otoliths changed	−0.263
	Tows changed	−0.239

Note: Method refers to whether the number of otoliths on selected tows (otoliths changed) or number of tows sampled were changed (tows changed). Values of nominal sample size used were the mean per year across bootstraps. Log–log slope is interpreted as the expected percent change in CV OFL per percent change in the nominal sample size.

on the information content (Hulson et al. 2012; Kotwicki et al. 2014). This suggests that it is possible to increase the information obtained by sampling fewer otoliths-per-tow and instead increasing the number of tows sampled. The framework we developed is generalized and could be applied to stocks from other regions; however, to showcase the utility of it, we review specific examples of how the findings herein could be applied to refine protocols for our example region (the GOA). In the context of the GOA bottom trawl survey, sampling for otoliths is currently conducted randomly within a tow. However, there are often stipulations in place based on the number of individuals of a species that are caught on a tow that dictates whether specimens are kept and how many. For example, adult walleye pollock are only sampled if more than 10 individuals are observed in a tow; if 10–500 fish are caught five are sampled, and if >500 are caught 10 are sampled (RACE 2021). Additional analysis could be performed to identify how many tows would be sampled if the threshold for sampling was increased from 10 fish to perhaps 20 or more. There are other species sampled on this survey that have similar catch stipulations for retaining specimens for otolith extraction, and a reduction in pollock sampling could allow additional tows to be sampled for those species by reducing the catch threshold for sampling. This would require including those species in the framework we have developed to identify which stocks would benefit in terms of OFL uncertainty reductions. In addition, dusky rockfish and Pacific ocean perch are currently sampled every tow they are caught on, up to 5 and 10 individuals, respectively. Given that the +33% otoliths-per-tow sampling treatment resulted in a significant decrease in OFL uncertainty (Fig. 6; Table A5), this framework could also be used to assess when an increase in the number of otoliths taken per tow could benefit the stock assessments (i.e., reducing the uncertainty in OFL).

Given the limited number of stocks in this study (i.e., three out of 23 stocks or complexes assessed by the NPFMC in the GOA stock area), there may be other stocks in this region that would benefit more or cost less (in terms of sampling costs or OFL uncertainty) as a result of changes to otolith sampling. In addition, at least at the AFSC, the same ageing laboratory conducts age readings for both the GOA and EBS stock areas, so it is feasible to propose a change in sampling across species within a stock area, or between species across stock areas. Therefore, we suggest that this analysis could be applied to more species in the GOA, as well as to species within the EBS stock area, to determine how a change in sampling protocols would achieve the most appropriate, cost-effective trade-off. Further, the three examples we provide showed relatively little response in effective sample size across changes to input sample size (Fig. A2). Since the calculation of effective sample size incorporates process error that could be the result of model misspecification, down-weighting of input sample size could reflect this misspecification instead of the explicit effects of changes to sampling rate. Therefore, in future work, it may be prudent to make comparisons between other stocks that display more closely related changes in effective and input sample sizes as a result of changes in ageing effort.

Despite our discussion of sampling, our analysis was focused on the effects of changes to ageing effort not collection effort. Therefore, if agencies wish to retain current sampling rates to provide insurance against a potential future need of more samples, they could continue collecting individuals at the current rate and make an adjustment to the number of otoliths that are read in routine annual ageing campaigns. Future needs could arise owing to structural changes to stock assessment methods (e.g., a shift from sex-aggregated to sex-specific, a change in the spatial stock structure, etc.), and we suggest that with small changes to our generalized framework (e.g., generating sex-specific age compositions and input sample sizes, or generating these data products for strata under new stock structure definitions) it could be an important tool for assessing the effect of data collection strategies on estimation of the OFL and uncertainty of the OFL.

In addition, as seen in Fig. 3, there was variation in calculated input sample size across years. This is directly related to the amount of sampling that was initially conducted and the nature of that sampling (i.e., inconsistent number of individuals sampled per tow or inconsistent number of tows sampled per year). These inconsistencies led to higher variance than would be the case for an optimal design given the bootstrapping process, which results in variation in the calculated input sample size. Unfortunately, the variation in catch that leads to differential numbers of otoliths per tow is simply the nature of fishing operations, especially in a fisheries-independent survey operation.

The framework we present is a generalized process that can be employed to address various questions related to sampling on fisheries-dependent and -independent platforms. The presented application of our framework is related to otolith sampling and ageing; however, it could easily be adapted to explore questions related to length frequency, catch-per-unit-effort, or other biological information routinely collected on these platforms. One important source of uncertainty that should be evaluated in the future involves the sampling design employed for AFSC surveys, whereby length frequency information is collected separately from specimens that are collected for ageing. This creates mismatches between the length bins for expanded length compositions and those associated with aged fish, resulting in "lost" ages when the age–length key is applied to the expanded length compositions in design-based estimators. Thus, the current analysis is appropriate for evaluating age-reading costs and trade-offs in an academic sense. However, to generate inputs used in formal assessments, analysts should likely include sampling of length frequency information at the tow level in the bootstrapping process as well (i.e., add a level of sampling in the estimator for the length expansion process). Other options for addressing this issue would be to (1) explore spatially aggregated (or constant) age–length keys that share information with a stratum that is missing a length bin necessary for expansion of ages or (2) share age–length key information from nearby tows in other strata when this mismatch occurs as Jourdain et al. (2020) did.

While we employed a random sampling design for selecting tows and otoliths-within-tows, many agencies still employ a length-stratified sampling design for otolith collections. Our framework can easily accommodate this in the sampling process for otoliths-within-tows by (1) adding the definitions for desired length bins, (2) slotting individual fish on each tow into these bins, and (3) adding another loop within the tow loop to cycle through length bins for sampling at the desired rate. In addition, the model-fitting module of this framework is also generalized, can be easily adapted for other stocks, and can employ any desired likelihood for compositional information (e.g., Francis reweighting as used here, the Dirichlet-multinomial, etc.). However, stock assessment methodologies vary across the globe, so our application may be most appropriate for US scientists. Some agencies (e.g., ICES and Fisheries and Oceans Canada) use stock assessment models that do not employ explicit data-weighting methods (e.g., state-space assessment models: Nielsen and Berg 2014; Woods Hole Assessment Model (WHAM): Stock and Miller 2021). Instead, these models conduct data-weighting implicitly via the choice of the likelihood function, correlation structure, and subsequent estimation of covariance parameters. Therefore, these assessment methods do not require input sample sizes for explicit data-weighting methods, but they could still benefit from the full "expansion" and "specimen" bootstraps as they provide a better understanding of variation in the annual sampling programs, and in the underlying populations. In addition, the "specimen bootstrap" that resamples age data at the desired sampling rate could be employed to calculate alternative age compositions that result from changes in sampling rate.

Bootstrap estimators are sometimes biased, and there are defined methods for quantifying this bias (e.g., "the bootstrap estimate of bias", "the improved estimate of bias", "the jackknife estimate of bias"; Efron and Tibshirani 1994). Naturally, these estimates of bias can be used to correct an estimator for this bias. However, Efron and Tibshirani (1994) note that bias correction can be dangerous in practice owing to the possibility of a bias-corrected estimator displaying a substantially higher standard error that must be checked for. Regardless, this would be a natural next step for research seeking to improve the bootstrap estimator we developed for this study. In this vein, one would need to weigh the bias estimated for the bootstrapping process against the size of standard errors generated for a bias-corrected estimator and decide which product is safer to use. However, this would require developing a complex simulated data set that retains the hierarchical structure of the length and age data collected on research surveys that facilitates the sampling of tows, lengths- or ages-within-tows, and then the resampling of these simulated data for the calculation of N_{Input}. The estimated bias could then be applied to the estimator to correct for it, given that the standard error generated in this process does not outweigh the original bias that was estimated. Given the scope of this analysis, we think it is best pursued in a future study.

In conclusion, we recommend that science agencies interested in restructuring their sampling protocols employ the framework developed here to quantify trade-offs and develop optimal sampling strategies. We have demonstrated that there are alternatives to current sampling regimes among the species included here. Applying this process to a greater number of stocks may elucidate a more effective strategy for changes to sampling effort (i.e., which species can afford decreased otolith sampling vs. those that would benefit the most in terms of information content and OFL uncertainty). This process would benefit from a parallel effort to generate model-based estimates of input sample size, which could corroborate the design-based calculation of age compositions and bootstrap estimates of input sample size. We also recommend additional exploration regarding how changes in the number of tows sampled from would simultaneously affect the precision of indices of abundance.

Acknowledgements

We thank S. Kotwicki and S. Lowe (AFSC) and two anonymous reviewers for their helpful comments on an earlier draft.

Article information

History dates

Received: 9 March 2022
Accepted: 16 August 2022
Accepted manuscript online: 21 September 2022
Version of record online: 4 November 2022

Copyright

Data availability

No new data were collected as a result of this study. All survey and fishery data used to generate data products for stock assessments included are hosted by the National Oceanic and Atmospheric Administration and are available upon request unless bound by privacy stipulations.

Author information

Author ORCIDs

Matthew R. Siskey https://orcid.org/0000-0002-7834-4749
James T. Thorson https://orcid.org/0000-0001-7415-1010

Author notes

Present address for Matthew R. Siskey: Washington Department of Fish & Wildlife, Olympia, WA, USA.

Author contribution

Conceptualization: MRS, AEP, MDB, JNI, JTT
Data curation: P-JFH, JNI
Formal analysis: MRS
Funding acquisition: AEP, JTT
Investigation: MRS
Methodology: MRS, AEP, P-JFH, MDB, JTT
Project administration: AEP, JTT
Resources: AEP, JTT
Software: MRS
Supervision: AEP, JTT
Validation: MRS, AEP, P-JFH, JNI, JTT
Visualization: MRS, MDB
Writing – original draft: MRS
Writing – review & editing: MRS, AEP, P-JFH, MDB, JNI, JTT

Competing interests

The authors declare there are no competing interests.

Funding statement

This publication is partially funded by the Cooperative Institute for Climate, Ocean, & Ecosystem Studies (CIOCES) under NOAA Cooperative Agreement NA20OAR4320271, Contribution No. 2022–1190.

References

Bartlett, M.S. 1937. Properties of sufficiency and statistical tests. Proc. R. Stat. Soc. Ser. A, **160**: 268–282.

Berg, C.W., and Nielsen, A. 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES J. Mar. Sci. **73**(3): 1788–1797. doi:10.1093/icesjms/fsw046.

Board, O.S. 2000. Improving the collection, management, and use of marine fisheries data. National Academics Press, Washington, DC.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W. Nielsen, A., et al. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal, **9**(2): 378–400. doi:10.32614/RJ-2017-066.

Chen, Y., Chen, L., and Stergiou, K.I. 2003. Impacts of data quantity on fisheries stock assessment. Aquat. Sci. **65**: 92–98. doi:10.1007/s000270300008.

Dorn, M.W., Deary, A.L., Fissel, B.E., Jones, D.T., Lauffenburger, N.E. Palsson, W.A., et al. 2020. Assessment of the Walleye Pollock stock in the Gulf of Alaska. North Pacific Fishery Management Council, Gulf of Alaska Stock Assessment and Fishery Evaluation Report.

Efron, B., and Tibshirani, R.J. 1994. An introduction to the bootstrap(1st ed.). Chapman and Hall/CRC. Boca Raton, Fl. 10.1201/9780429246593.

Fenske, K.H., Hulson, P-J.F., Williams, B., and O'Leary, C.A. 2020. Assessment of the Dusky Rockfish stock in the Gulf of Alaska. North Pacific Fishery Management Council, Gulf of Alaska Stock Assessment and Fishery Evaluation Report.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A. Maunder, M.N., et al. 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim. Methods Softw. **27**: 233–249. doi:10.1080/10556788.2011.597854.

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. **68**: 1124–1138. doi:10.1139/F2011-025.

Francis, R.I.C.C. 2017. Revisiting data weighting in fisheries stock assessment models. Fish. Res. **192**: 5–15. doi:10.1016/j.fishres.2016.06.006.

Hulson, P-J.F., Hanselman, D.H., and Quinn, T.J. 2012. Determining effective sample size in integrated age-structured assessment models. ICES J. Mar. Sci. **69**: 281–292. doi:10.1093/icesjms/fsr189.

Hulson, P-J.F., Lunsford, C.R., Fissel, B., and Jones, D. 2020. Assessment of the Pacific ocean perch stock in the Gulf of Alaska. North Pacific

Fishery Management Council Gulf of Alaska Stock Assessment and Fishery Evaluation Report.

Jourdain, N.O.A.S., Breivik, O., Fuglebakk, E., Aanes, S., and Volstad, J.H. 2020. Evaluation of sampling strategies for age determination of cod (*Gadus morhua*) sampled at the North Sea International Bottom Trawl Survey. ICES J. Mar. Sci. **77**(3): 859–869. doi:10.1093/icesjms/fsaa013.

Kotwicki, S., Ianelli, J.N., and Punt, A.E. 2014. Correcting density-dependent effects in abundance estimates from bottom-trawl surveys. ICES J. Mar. Sci. **71**(5): 1107–1116. doi:10.1093/icesjms/fst208.

Lambert, G., Helser, T.E., Berger, A., Olsen, E., Hastie, J. O'Malley, J., et al. 2017. NOAA technical memorandum: importance of age data collection for stock assessments: a US national perspective. Report from the Otolith Sampling Size Working Group (OSSWG).

Lauth, R.R., Dawson, E.J., and Conner, J. 2019. Results of the 2017 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna. US Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-396.

Levene, H. 1960. Robust tests for equality of variances. *In* Contributions to probability and statistics: Essays in honor of Harold Hotelling. *Edited by* I. Olkin, H. Hotelling, et al. Stanford University Press, Stanford, CA. pp. 278–292.

Maunder, M.N. 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: estimating the effective sample size. Fish. Res. **109**: 311–319. doi:10.1016/j.fishres.2011.02.018.

McAllister, M.K., and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. **54**: 284–300.

Methot, R.D., and Wetzel, C.R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. **142**: 86–99. doi:10.1016/j.fishres.2012.10.012.

Miller, T.J., and Skalski, J.R. 2006. Integrating design- and model-based inference to estimate length and age composition in North Pacific longline catches. Can. J. Fish. Aquat. Sci. **63**: 1092–1114. doi:10.1139/F06-022.

Nielsen, A., and Berg, C.W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. **158**: 96–101. doi:10.1016/j.fishres.2014.01.014.

NPFMC. 2020. Fishery management plan for groundfish of the Gulf of Alaska. North Pacific Fishery Management Council, Anchorage, AK.

Pennington, M., Burmeister, L.M., and Hjellvik, V. 2000. Assessing the precision of frequency distributions estimated from trawl-survey samples. Fish. Bull. **100**(1): 74–80.

Pennino, M.G., Conesa, D., Lopez-Quilez, A., Munoz, F., Fernandez, A., and Bellido, J.M. 2016. Fishery-dependent and -independent data lead to consistent estimations of essential habitats. ICES J. Mar. Sci. **73**(9): 2302–2310. doi:10.1093/icesjms/fsw062.

RACE. 2021. Scientific operations plan: 2021 Gulf of Alaska Bottom Trawl Survey, Resource Assessment and Conservation Engineering Division, Cruise 2021-01. National Oceanic and Atmospheric Administration, Seattle, WA, USA.

Ralston, S., Punt, A.E., Hamel, O.S., Devore, J.D., and Conser, R. 2011. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. Fish. Bull. **109**: 217–231.

Shertzer, K.W., Prager, M.H., and Williams, E.H. 2008. A probability-based approach to setting annual catch levels. Fish. Bull. **106**: 225–232.

Stewart, I.J., and Hamel, O.S. 2014. Bootstrapping of sample sizes for length- or age-composition data used in stock assessments. Can. J. Fish. Aquat. Sci. **71**: 581–588. doi:10.1139/cjfas-2013-0289.

Stock, B.C., and Miller, T.J. 2021. The Woods Hole Assessment Model (WHAM): a general state-space assessment framework that incorporates time- and age-varying processes via random effects and links to environmental covariates. Fish. Res. **240**: 105967. doi:10.1016/j.fishres.2021.105967.

Thorson, J.T. 2014. Standardizing compositional data for stock assessment. ICES J. Mar. Sci. **71**(5): 1117–1128. doi:10.1093/icesjms/fst224.

Thorson, J.T., and Haltuch, M.A. 2019. Spatiotemporal analysis of compositional data: increased precision and improved workflow using model-based inputs to stock assessment. Can. J. Fish. Aquat. Sci. **76**: 401–414. doi:10.1139/cjfas-2018-0015.

von Szalay, P.G., and Raring, N.W. 2018. Data Report: 2017 Gulf of Alaska Bottom Trawl Survey. US Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-374.

Appendix A

Tables A1–A6 and Figs. A1 –A2 appear on the following pages.

Table A1. Mean input sample sizes (N_{Input}) for dusky rockfish calculated with the bootstrap estimator across "specimen bootstrap" replicates and years using $m = 100$, 200, 300, 400, or 500 "specimen bootstrap" replicates per sampling treatment.

Sampling treatment	$m = 100$	$m = 200$	$m = 300$	$m = 400$	$m = 500$
−67%	44	44	44	44	44
−33%	86	86	86	86	86
0%	129	130	130	129	129
+33%	176	175	175	175	175
+67%	222	223	223	222	222

Note: All trials were conducted under the "tows changed" sampling method (i.e., the number of tows sampled was changed).

Table A2. Input sample sizes (N_{Input}) for dusky rockfish calculated using the bootstrap estimator for a single "specimen bootstrap" replicate per sampling treatment when $b = 100$, 500, or 1000 "expansion bootstrap" replicates were included in the calculation.

Sampling treatment	$b = 100$	$b = 500$	$b = 1000$
−67%	54	53	53
−33%	81	80	80
0%	147	146	144
+33%	155	156	156
+67%	228	224	224

Note: All trials were conducted under the "tows changed" sampling method (i.e., the number of tows sampled was changed).

Table A3. Types of data used in the stock assessments for each stock.

Species	Data type	Years
Dusky rockfish	Fishery catch	1977–2020
	Fishery age composition	2000–2006, 2008–2018 (biennial)
	Fishery length composition	1990–1999, 2007–2019 (biennial)
	NMFS Bottom Trawl Survey Biomass Index	1984–1999 (triennial), 2001–2019 (biennial)
	NMFS Bottom Trawl Survey age composition	1984–1999 (triennial), 2001–2019 (biennial)
Pacific ocean perch	Fishery catch	1961–2020
	Fishery age composition	1990, 1998–2002, 2004–2006, 2008–2018 (biennial)
	Fishery length composition	1963–1977, 1991–1997
	NMFS Bottom Trawl Survey Biomass Index	1990–1999 (triennial), 2001–2019 (biennial)
	NMFS Bottom Trawl Survey age composition	1990–1999 (triennial), 2003–2019 (biennial)
Walleye pollock	Fishery catch	1970–2020
	Fishery age composition	1975–2019
	Shelikof Strait Acoustic Survey Biomass Index	2008–2020 (2011 missing)
	NMFS Summer Acoustic Survey Biomass Index	2013–2019 (biennial)
	NMFS Summer Acoustic Survey age composition	2013–2019 (biennial)
	NMFS Summer Acoustic Survey length composition	2015–2019 (biennial)
	NMFS Bottom Trawl Survey Biomass Index	1990–1999 (triennial), 2001–2019 (biennial)
	NMFS Bottom Trawl Survey age composition	1990–1999 (triennial), 2001–2019 (biennial)
	NMFS Bottom Trawl Survey length composition	1999–2019 (biennial, 2003 missing)
	ADF&G Trawl Survey Abundance Index	1988–2020
	ADF&G age composition	1989–1997 (missing 1991 and 1995), 1998–2018 (biennial)

Note: All compositional data and catch-at-age or number-at-age data inputs are in proportions. NMFS = National Marine Fisheries Service; ADF&G = Alaska Department of Fish & Game.

Fig. A1. (A) Covariance matrix calculated from proportions-at-age for one year of data (number of age classes = 22; year = 2019) across specimen bootstrap replicates ($m = 200$). Age composition replicates used were taken from the dusky rockfish 0% change sampling rate treatment. (B) Covariance matrix calculated from a random multinomial sample, where the sample size was set to the mean input sample size across bootstrap replicates of the 0% treatment ($N_{\mathrm{Input}} = 169$) and the probability was set as the original proportions-at-age for dusky rockfish.

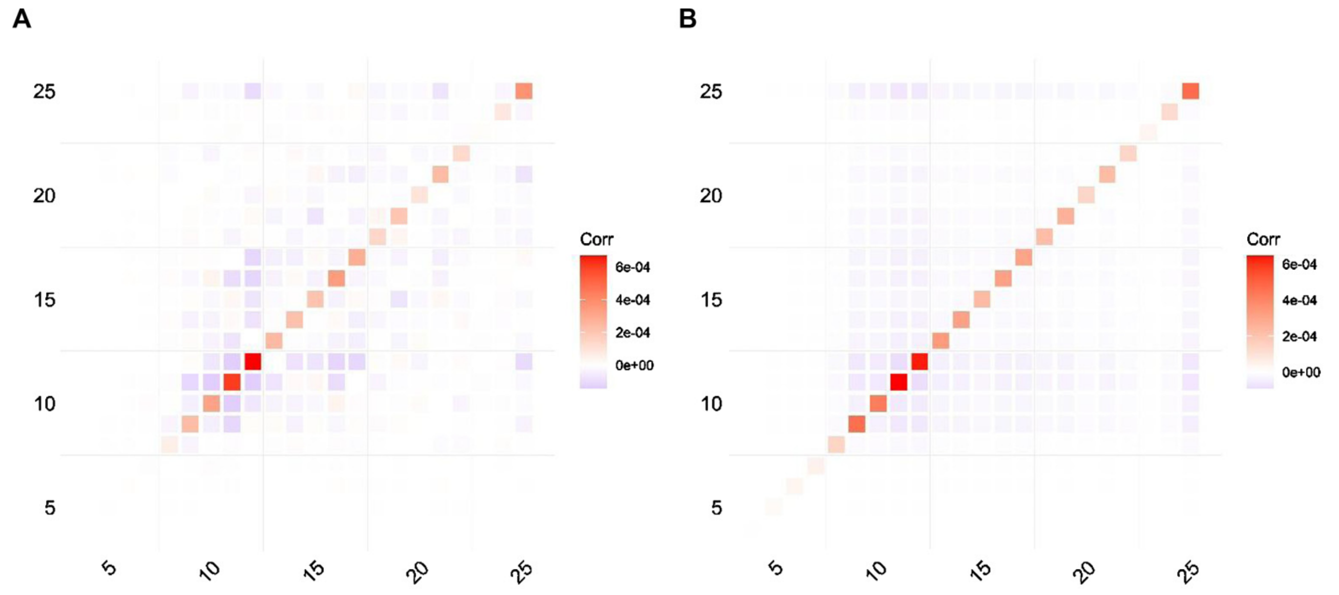


Table A4. Slope of the log–log relationships between (*i*) input sample size and number of otoliths (“otoliths changed”; eq. 3; third column) and (*ii*) input sample size and number of tows sampling otoliths (“tows changed”; eq. 3; fourth column).

Species	Year	Otoliths changed	Tows changed
Dusky rockfish	1984	0.632	0.379
	1987	0.433	0.941
	1990	0.808	1.708
	1993	0.542	1.066
	1996	0.579	1.066
	1999	0.577	1.098
	2001	0.630	1.086
	2003	0.426	1.211
	2005	0.559	0.994
	2007	0.599	1.034
	2009	0.691	1.091
	2011	0.592	1.033
	2013	0.604	1.046
	2015	0.636	1.054
	2017	0.635	1.079
	2019	0.538	0.976
Pacific ocean perch	1990	0.590	1.024
	1993	0.593	1.022
	1996	0.659	1.042
	1999	0.627	1.003
	2003	0.684	1.030
	2005	0.669	0.996
	2007	0.721	1.028
	2009	0.491	0.915
	2011	0.624	1.050
	2013	0.696	1.043
	2015	0.715	1.024
	2017	0.707	1.050
	2019	0.541	1.104
Walleye pollock	1990	0.664	1.292
	1993	0.529	1.058
	1996	0.214	0.696
	1999	0.666	1.172
	2001	0.468	1.054
	2003	0.550	1.123
	2005	0.770	1.278
	2007	0.461	0.901
	2009	0.375	0.859
	2011	0.594	1.150
	2013	0.660	1.161
	2015	0.626	1.106
	2017	0.739	1.513
	2019	0.664	1.239

Note: Log–log slope is interpreted as the expected percent change in input sample size per percent change in the number of otoliths processed or number of tows conducted.

Fig. A2. Input sample sizes calculated using the bootstrap estimator related to the effective sample size as determined through assessment model runs, averaged across bootstrap replicates for each sampling rate treatment and years.

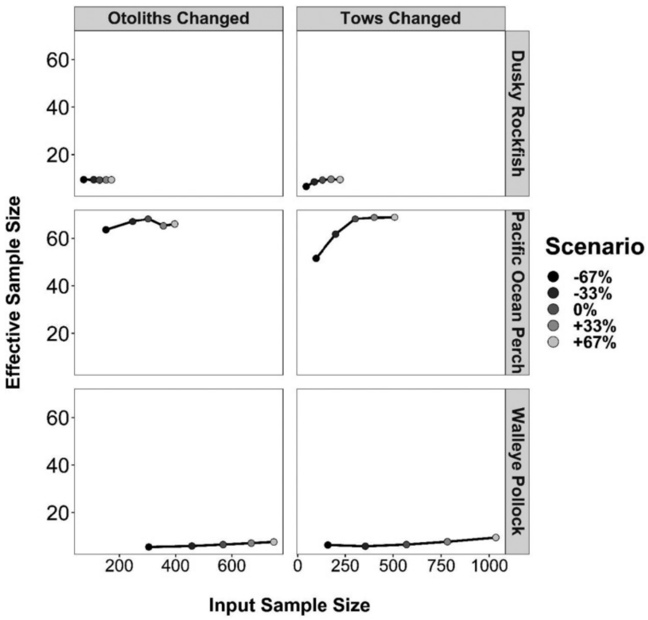


Table A5. *P*-value results from Levene's test for homogeneity of variance ran on sequential pairs of overfishing limit (OFL) distributions that resulted from stock assessment model runs using data products from each sampling rate treatment (±%) for each sampling method ("otoliths changed" vs. "tows changed") and stock.

Method	Stock	−67% vs. −33%	−33% vs. 0%	0% vs. +33%	+33% vs. +67%
Otoliths changed	Dusky rockfish	0.835	0.094	0.041	0.783
	Pacific ocean perch	0.178	0.083	0.001	0.422
	Walleye pollock	0.012	0.103	0.533	0.736
Tows changed	Dusky rockfish	0.001	0.283	0.001	0.803
	Pacific ocean perch	0.001	0.272	0.001	0.141
	Walleye pollock	0.114	0.127	0.020	0.001

Table A6. The change in the mean number of otoliths across the number of calendar years since the start of the survey (nominal sample size/year; N_{Nominal}/year) averaged across specimen bootstraps for each sampling rate treatments (Treatment) and sampling method (Method). Change in cost per year (US$·year^{-1}) associated with ageing the N_{Nominal}/year of each treatment given the per-otolith cost calculated for each stock.

Method	Species	Treatment	N_{Nominal} change (otoliths·year^{-1})	Cost change (US$·year^{-1})
Otoliths changed	Dusky rockfish	−67%	−263	−16 435
		−33%	−115	−7 193
		0%	0	0
		+33%	161	10 039
		+67%	311	19 446
	Pacific ocean perch	−67%	−646	−26 158
		−33%	−281	−11 374
		0%	0	0
		+33%	369	14 958
		+67%	739	29 930
	Walleye pollock	−67%	−794	−14 773
		−33%	−360	−6 697
		0%	0	0
		+33%	492	9 154
		+67%	928	17 272
Tows changed	Dusky rockfish	−67%	−287	−17 965
		−33%	−141	−8 795
		0%	0	0
		+33%	139	8 680
		+67%	285	17 813
	Pacific ocean perch	−67%	−685	−27 728
		−33%	−337	−13 658
		0%	0	0
		+33%	330	13 357
		+67%	678	27 454
	Walleye pollock	−67%	−861	−16 023
		−33%	−428	−7 970
		0%	0	0
		+33%	423	7 874
		+67%	855	15 917