Data weighting: an iterative process linking surveys, data synthesis, and population models to evaluate mis-specification

Alternative title:

- A guide to identify model misspecification and to appropriately weight data in stock assessment models
- Data weighting: Putting model specification under the microscope

James T. Thorson[1,*], Cole C. Monnahan[1], Peter-John F. Hulson[2]

[1] Resource Ecology and Fisheries Management, Alaska Fisheries Science Center, NOAA

[2] Marine Ecology and Stock Assessment, Auke Bay Laboratories, Alaska Fisheries Science Center

* Corresponding author: James.Thorson@noaa.gov

Abstract:

Integrated stock assessments specify a distribution for multiple data types, and these distributions control the relative leverage assigned to each datum. A decade of research has demonstrated that (1) proper data weighting is necessary to avoid bias resulting from overweighting noisy age- and length-composition data; (2) sampling data can be pre-processed to estimate the likely sampling variance for composition data; and (3) using random effects to estimate time-varying parameters can improve the fit to data while also changing statistical leverage, and thereby serve a similar role to reweighting data. However, there are also unresolved questions including: (A) Is it more appropriate to model age and length data as proportions-at-age and as an index for the total, or as a series of indices-at-age? (B) Are correlated residuals appropriately addressed via data weighting or do they require additional model changes (i.e., time-varying parameters)? (C) How to efficiently communicate information about sampling imprecision and model errors between sampling and stock-assessment teams? (D) how does model-based expansion of sampling data affect data weighting? And (E) how to address alternative hypotheses about factors driving poor fit to data? Here, we argue that stock assessment errors can be classified using four categories: sampling bias (e.g., changes in survey coverage), sampling imprecision (e.g., finite sample sizes), assessment model bias (e.g., incorrect demographic assumptions) and assessment model imprecision (e.g., random effects). This categorization has several implications with resulting practical recommendations. For example, we define Percent Excess Variance (PEV) from the ratio of input sample size (the measured variance of sampling imprecision) and effective sample sizes (the variance of assessment-model residuals). We propose calculating PEV as standardized diagnostic measuring the net effect of survey bias and assessment model bias and imprecision. We demonstrate PEV in a simulation experiment fitted using the Woods Hole Assessment Model

(WHAM) conditioned upon Gulf of Alaska walleye pollock, where unacknowledged fishery selectivity results in a PEV of 77% and this is eliminated when correctly specifying a time-varying estimation model. We also argue that model-based expansion of data inputs using auxiliary information can mitigate sampling bias, while also measuring sampling imprecision for spatially unrepresentative surveys. Similarly, including random effects can similarly mitigate model bias while increasing model imprecision when the demographic model has little explanatory power. Finally, we observe that down-weighting compositional data for a given fleet fails to propagate information about model residuals when interpreting abundance indices or reference points for that same fleet. When PEV is large for important fleets, we therefore encourage focused research to explain the sources of these errors rather than simply downweighting without propagating information about residuals. However, we acknowledge a continuing role for automated data weighting for less important fleets, although we recommend explicit hypotheses about potential sources of errors in those cases.

Keywords: Data weighting; stock assessment; state-space model; random effects; data standardization;

1. Integrated assessment models, and weighting data in fleets

High-quality stock assessments are one important component of effective fisheries management (Hilborn et al., 2020). In the US for example, stock assessments are central to the system of accountability measures ensuring that regional fisheries management councils do not set fishing levels above those associated with long-term policy objectives (Methot et al., 2014). For stock assessments to provide accurate management advice, their observation components (data likelihoods) need to appropriately reflect the information content in the data. However, this continues to be a major challenge despite decades of research.

Modern "integrated" stock assessments typically incorporate many different types of information (Maunder and Punt, 2013). To do so, they typically require specifying one or more "fleets," where each fleet can then be associated with common types of data:

1. *Removals*: Some fleets have a measurement of total landings, discards, or both for year *t* (c_t). Surveys are sometimes assumed to have negligible removals, although catches in a bottom trawl survey for recovering stocks can sometimes represent a substantial fraction of fishing mortality;
2. *Index of abundance*: Additionally, some fleets will provide records of catch and effort at a fine scale, allowing design- or model-based estimators to be applied to estimate an index of abundance (b_t);
3. *Age/length/sex composition*: Finally, some fleets will have catches that are subsampled, where these subsamples are then measured for age, length, and/or sex. These records can then be expanded to estimate the proportion of the population (or fleet removals) within a given age/length/sex category *a* ($p_{a,t}$), and we refer to these as composition data in the following.

Other types of data are also widespread including (but not limited to) conventional tags, weight-at-age matrices, and maturity-at-age ogives, but we focus on these three in subsequent discussions. We also note that some assessment models (e.g., Stock Synthesis: Methot and Wetzel, 2013) are designed to fit removals (c_t) and abundance indices (b_t) separately from compositions ($p_{a,t}$), while others (e.g., SAM: Berg and Nielsen, 2016) are fitted to data that represent a combination of these types, either via fitting to removals at age ($c_{a,t} \equiv c_t p_{a,t}$) or indices-at-age ($b_{c,t} \equiv b_t p_{a,t}$).

Importantly, most fleets will have two or more of these data types simultaneously. For example, many fisheries are sampled to provide a measure of removals as well as composition data, and many surveys are conducted to measure an index of abundance and age/length/sex composition. In these examples, respectively, the composition data helps to interpret the removals or abundance index by providing an estimate of fishery or survey selectivity. However, composition data will also be informative about the relative size of different cohorts as well as total mortality rates, in particular when selectivity-at-age for that fleet is relatively constant over time. In these cases, composition data plays a dual role of informing fleet selectivity (a measurement process for that specific fleet) as well as tracking cohorts through the population (an aspect of population dynamics for the stock as a whole).

Even for stocks with a well-funded monitoring program, abundance indices typically have a coefficient of variation of 5% or greater, and this is then fitted using a lognormal distribution. By contrast, the same monitoring program might sample 100s-1000s of fishes for age, and 1000-10,000s for length each year, and these are often fitted using a multinomial distribution. The integrated model then identifies parameter estimates by maximizing a joint log-likelihood, which is calculated as the sum of log-likelihoods for each fleet and data type individually. In this case, if the multinomial distribution is specified for age or length-composition data using a sample size of 100s or 1000s and selectivity-at-age is constant over time, then the statistical leverage for composition data on estimates of cohort size (and resulting trends in abundance) will typically be much greater than the leverage for abundance indices or other data types. Therefore small mis-specification of the processes affecting age/length/sex composition data can override the information arising from abundance indices.

A well-known series of papers have reviewed these topics previously (Francis, 2017, 2014, 2011), and have advocated for various methods for "tuning" the multinomial sample size associated with age/length/sex composition data. However, two major developments have also occurred since these reviews, namely: (1) increased use of age-structured state-space models fitted to indices-at-age, and (2) increased use of standardization models to pre-process data inputs to mitigate bias arising from climate-driven or logistically-constrained sampling issues. In particular, an assessment model might allow for time-varying selectivity, which decreases the statistical leverage of composition data on estimates of abundance trends and in some sense replaces the action of tuning sample sizes (Xu et al., 2020). Similarly, improved standardization of input data might improve model fit and thereby reduce the need to downweight available data (Thorson and Haltuch, 2018). These developments provide new options to deal with poor fit and high leverage for composition data, and can accomplish a similar role as tuning input sample sizes. However, we will follow past papers in using the term "data-weighting" for procedures that explicitly tune (or estimate weights) for composition data.

These two developments have therefore given new importance to the following five questions:

1. Is it more appropriate to model age and length sampling data as proportions-at-age and use a separate index for the total index of abundance or removals (i.e., similar to Stock Synthesis), or should these be combined in a series of indices-at-age (i.e., similar to SAM)?
2. Are correlated residuals appropriately addressed via data weighting or do they require additional model changes (i.e., time-varying parameters)?
3. How can survey and analytical teams efficiently communicate information about sampling imprecision for routine use in stock assessments?
4. How does model-based expansion of sampling data affect the process or interpretation of data weighting?
5. How should assessment scientists address alternative hypotheses about mechanisms that give rise to poor fit (and associated low weighting) for data?

To provide a foundation for addressing these new questions, we discuss both the processes by which removals, abundance indices, or composition data are sampled as well as how they are processed prior to inclusion in a stock assessment model. We then outline what this implies about data-weighting (which we note was conspicuously absent from prior discussions of data-weighting).

We therefore organize the paper as follows. We first review how abundance indices and compositional data arise in nature, how they are processed to generate stock-assessment inputs, and what this implies about their statistical distribution. We then expand previous efforts to partition errors into different interpretable processes, and review which might be similar across fleets. Finally, we use the preceding discussions to propose eight recommendations for applying data-weighting in real-world assessments.

2. How are samples expanded to create abundance indices and composition data

To begin, we briefly review how design-based estimators are used to expand survey data to generate abundance indices and composition data. We describe a case involving a survey with a stratified random sampling design used to generate a biomass index. We also envision that the survey has many subsamples of length but a smaller number of subsampled ages, such that proportion-at-length or proportion-at-age can be calculated. Subsampling designs vary between regions (e.g., using length-stratified or random subsampling for age-length specimens used to estimate an age-length-key), and these design decisions will then affect the design-based estimator and associated variance estimators (e.g., Hulson et al., 2023). Given these nuanced differences, we intended to provide only a broad overview involving a simplified case and introduce only the notation that is central to our argument.

To construct a design-based abundance index under this design, note that each sample i yields a measurement of density calculated as weight (or numbers) per area swept $D_i = W_i/A_i$. Given that inclusion probabilities are assumed constant in a given sample stratum x, average density for each stratum $\overline{D}_x$ is first calculated as the average of density for samples in that stratum. Stratum average densities are then expanded to the area of each stratum, and these are summed across strata within a broader region to get the index, $b = \sum_{x=1}^{n_x} A_x \overline{D}_x$. Similarly, the variance can be calculated as the area-expanded sum of the variance among samples for each stratum, $\widehat{Var}(b) = \sum_{x=1}^{n_x} A_x^2 \widehat{Var}(\overline{D}_x)$.

By contrast, constructing a design-based proportion-at-length involves more steps. Each sample i is measured for total mass W_i (as described previously when expanding an abundance index) and the design typically dictates that some portion w_i is subsampled, where each individual in this subsample is measured for length. Tabulating the lengths in bins yields a vector of subsampled abundance-at-length which is then expanded by $\lambda_i = W_i/w_i$ to predict abundance-at-length for the entire tow. This tow-level abundance-at-length is then again summed across tows in a given stratum, expanded by stratum area or auxiliary information about stock abundance in that stratum, and summed across strata to estimate total abundance-at-length. This total abundance-at-length is then sometimes converted to a proportion-at-length by dividing by the sum across lengths To develop abundance- or proportion-at-age, a further step might be involved, where a set of paired ages and length measurements is collected and analyzed to estimate a forward age-length key (Ailloud and Hoenig, 2019). Abundance-at-length can then be multiplied by this age-length key to predict abundance-at-age, and this in turn converted to proportion-at-age.

From these two descriptions we see that:

1. Each sample used to calculate proportions-at-length or –at-age involves a subsample of some size w_i that is measured for length, and hence yields a subsampled "proportion-at-length" (i.e., a vector $p_{i,c}$ that has a sum of 1 across lengths c). However, the expansion process involves multiplying this proportion by the random variable W_i (the total captured in that sample). This product $W_i p_{i,c}$ is obviously not a proportion;
2. Abundance-at-length is calculated from a multi-level sampling process that involves many potential sources of sampling variance, including the subsampled lengths/ages within each sample and the sampled abundance within each stratum. Therefore, the resulting abundance-at-length estimator is likely to have higher variance than an abundance index. Similarly, the abundance-at-age involves an estimate of the forward age-length-key, which accumulates additional variance;
3. Abundance indices can all result in measurements of zero, whenever zero animals are counted for a given year. This occurs more frequently when sampling abundance-at-age

or abundance-at-length (particularly for age/size classes that have a low numerical density), and any model must be suited to deal with these;

Additionally, the imprecision for the abundance index arises from a single source (among-sample variance within each stratum), and is straightforward to calculate. By contrast, the imprecision of proportions-at-age arises potentially from the number of individuals that are measured for age and length, the properties of the age-length-key, and many other sources.

Several different estimators have been proposed to calculate the imprecision of age and length composition data:

1. *Bootstrap estimators*: Research has proposed to resample with replacement from the set of sampling occasions (survey tows, fishing trips) and/or the specimens that are individually measured for age and length, calculate the variance among resampled replicates, and calculate the variance directly from these bootstrap samples (Crone and Sampson, 1997; Stewart and Hamel, 2014);
2. *Model-based estimators*: Alternatively, papers have proposed to fit a model to available data, calculate the standard errors for the estimated proportion, and use that directly as estimate of sampling variance (Berg and Nielsen, 2016; Thorson, 2014; Thorson and Haltuch, 2018);
3. *Design-based estimators*: As a third alternative, researchers have generalized design-based estimators to calculate the covariance resulting from a multi-level sampling design (Miller and Skalski, 2006);

In general, these estimators combine information about the multi-level sampling design, sample sizes, and the variation among samples to calculate the variance of the estimated proportions.

3. Partitioning error into different processes

We next discuss how these data are fitted in integrated stock assessment models such as Stock Synthesis (Methot and Wetzel, 2013). In the case of expanded age-composition data, for example, the expansion algorithm yields an expanded abundance-at-age, $n_{a,y}$. This can then be fitted to the assessment-model prediction of abundance-at-age, or alternatively $n_{a,y}$ can be converted to expanded proportion-at-age and fitted to the assessment-model prediction of proportion-at-age $\pi_{a,y}$. Fitting this model using maximum likelihood requires specifying a probability distribution for the data conditional upon parameters, where the log-likelihood is minimized to identify parameter estimates. Historically, a multinomial distribution was often used for age-composition data:

$$\mathbf{n}_y^* \sim Multinomial(\boldsymbol{\pi}_y, n_{input}) \quad (1)$$

where the fitted abundance-at-age $\mathbf{n}_y^*$ is a vector of $n_{a,y}^*$, calculated by taking the expanded abundance, rescaling to a proportion, and then multiplying it by an input sample size n_{input}, $n_{a,y}^* = n_{input} \frac{n_{a,y}}{\sum_{a'=1}^{A} n_{a',y}}$. This input sample size n_{input} then represents the number of idealized multinomial samples from a given fleet that would have the same approximate variances as the hierarchical sampling that occurred in nature. In the absence of a bootstrap, model-based, or design-based estimator for n_{input}, analysts have often used “rules of thumb” to define this value, or have reweighted this value as explained in a later section.

However, stock assessment models will never fit perfectly to age and length composition data. Historically, analysts would often calculate a Pearson residual as:

$$r_{a,y} = \frac{\frac{n_{a,y}}{\sum_{a'=1}^{A} n_{a',y}} - \pi_{a,y}}{\sqrt{\frac{\pi_{a,y}(1-\pi_{a,y})}{n_{input}}}} \tag{2}$$

where the numerator is the difference in proportion-at-age and the denominator is the standard deviation expected under a multinomial distribution with sample size n_{input}. More recently, these have been improved using one-step-ahead (OSA) residuals that account for the distribution of random effects as well as non-normal error distributions (Trijoulet et al., 2023). Many studies have observed that residuals have positive or negative streaks for a sequence of ages in a given year ("age-correlations"), for a sequence of years for a given age ("time-correlations"), for a sequence of ages and years for a given cohort ("cohort correlations"), and have larger magnitude than a standard normal distribution ("overdispersion").

Fitting a model where Pearson or OSA residuals have larger magnitude than a standard normal distribution has been called "overweighting" the composition data. Many studies have used simulation or case-study experiments to show that overweighting is likely to result in biased estimates of population dynamics, and that decreasing the weight in these cases will often improve assessment-model performance (Fisch et al., 2022, 2021; Punt, In press; Stewart and Monnahan, 2017; Xu et al., 2020). Similarly, patterns in residuals among ages or years is a widely used diagnostic for model mis-specification.

We attribute the lack-of-fit to stock assessment data to four different processes (summarized in Table 1). To describe these we distinguish three different properties of an estimator: (A) imprecision measures the variance around the mean of an estimator; (B) bias measures the difference between the mean of an estimator and a true value; (C) inconsistency arises when bias and imprecision do not decrease as sample sizes increase. For simplicity, we will emphasize the difference between imprecision (A) and both bias and inconsistency (B/C). We also categorize mechanisms causing imprecision or bias/inconsistency based on whether they arise during the sampling (1) or modelling (2) process.

To make this description more precise, let us assume that there is some true but unknown data-generating process $Z \sim DGP(.)$ that results in all state-variables Z associated with a given stock assessment, and we define a distribution $p(Z = z)$ for the value z that in reality arose over the spatial and temporal domain of an assessment. We also assume that there is some process resulting in data $X \sim f(Z, n_X)$ conditional upon that data-generating process and sample size n_X, where we define the distribution of data $p(X = x|z, n_X)$ conditional upon the realized state-variables. Finally, we define observable quantities $Y(Z)$ with value $y(z)$ given the realization z of state-variables, where these might include biological reference points (biomass at maximum sustainable yield, B_{msy}) and stock trends (biomass B_t). We can estimate these observables conditional upon an assumed model M and data X, where the model M is sometimes explicit (i.e., a population-dynamics model used to estimate mortality rates) and other times implicit (i.e., assumptions about the sampling frame when computing a design-based estimator). Given a realized sample x, we can apply an estimator $\hat{Y}(x, M)$ for an observable $Y(Z)$, where this estimator then has a distribution $\hat{Y}(p(X = x|z, n_X), z, M)$. We define:

- the mean for an estimator as $\mu_x \equiv \mathbb{E}_x\left(\hat{Y}(x, M)\right) = \int \hat{Y}(x, M) p(X = x|z, n_X)\, \mathrm{d}x$;
- the expected imprecision as $V = \mathbb{V}_x\left(\hat{Y}(z, M)\right) = \int \left(\hat{Y}(x, M) - \mu_x\right)^2 p(X = x|z, n_X)\, \mathrm{d}x$;
- the expected bias as $B = \mu_x - y(z)$
- the expected squared-error as $E^2 = B^2 + V$

Subsequently, we will further decompose squared-error into components arising from sampling processes vs. assessment modelling. For presentation, we'll assume that these four processes occur independently:

$$E^2 = V_{sample} + B_{sample}^2 + V_{model} + B_{model}^2 \quad (3)$$

such that expected squared-error arises as the sum of these different processes (see Table 1 for an overview). This decomposition is possible for any observable quantity $Y(Z)$, but in the following we will specifically emphasize fits to abundance-at-age data for a given fleet, and later discuss complications arising from fitting to data from multiple fleets.

3.1 Finite sample sizes causing "sampling imprecision"

We define "sampling imprecision" as imprecision arising from "taking a sample rather than a census" (Maunder and Piner, 2017). Although called "measurement error" by Francis (2011), we use the term "sampling imprecision" to indicate that additional sampling (e.g., full coverage of fishery observers resulting in a census) can sometimes eliminate this error entirely. We therefore know that sampling imprecision results in variance V_{sample}, and this variance decreases with increased sample sizes n_X or an efficient sampling design.

3.2 Mis-specified sampling design causing "sampling bias and inconsistency"

Similarly, sampling designs typically involve defining a sampling frame, which ideally has a perfect correspondence to the management unit ("stock") about which we seek inference (Cochran, 1977). Furthermore, many sampling designs use probability sampling, where each "sampling unit" (i.e., survey station) within this sampling frame is assigned a probability of inclusion. When the sampling frame does not correspond to a target population, even a perfect census will still result in error ("sampling inconsistency"). Similarly, when some sampling units are sampled above their intended inclusion probability, then a sample will overrepresent some components of the population and the survey may be biased for low sample sizes or inconsistent even for extremely large sample sizes. We call this "sampling bias" B_{sample}, acknowledging that it is conditional upon the specified sample size n_x and therefore is a combination of bias and inconsistency. The magnitude of sampling bias will increase due to poor assumptions about the sampling frame and logistical challenges in sampling. For example, with partial observer coverage, if fishing behavior differs between boats with and without an observer, then expanding observed trips on boats with observers will be a biased measure of fleetwide removals for any randomized allocation of observers, but this source of bias would be eliminated under complete coverage.

3.3 Parametric model mis-specification causing "model inconsistency"

Next, we note that stock assessment models typically make strong assumptions about population demography. For example, assessments typically ignore immigration/emigration from outside of a defined geographic area, and hence specify a survival function such that abundance for a given cohort can only decrease:

$$\log(N_{a+1,y+1}) = \log(N_{a,y}) - M_{a,y} - F_{a,y} \quad (4)$$

where this is identifiable because analysts typically specify some structure on natural mortality (e.g., constant mortality $M_{a,y} = M$), such that changes in cohort abundance $N_{a,y}$ over time is informative about fishing mortality rates $F_{a,y}$. Even as new data are progressively added to such a model, the parametric assumption that abundance declines for a cohort can never be overcome and will result in both bias and inconsistency when immigration, for example, results in increasing abundance-at-age for some cohorts. We see that this "model mis-specification" results in some bias B_{model}, and that the expected magnitude of this bias increases when the parametric model is based on ecological assumptions that have a poor match to the true data-generating process.

3.4 Semi-parametric model specification and "model imprecision"

Finally, hierarchical (a.k.a. state-space or mixed-effects) models specify a probability distribution for coefficients representing variation in some process over space, time, or among animals. They then estimate parameters defining this distribution jointly with other model parameters (Thorson and Minto, 2015). Estimated variability in these coefficients $\boldsymbol{\varepsilon}$ then approximates variation in growth, survival, mortality, or movement resulting from otherwise unmodeled processes (Ives, 2022). We here claim that random effects can be used to account for model misspecification in a way that translates "model bias/inconsistency" into "model imprecision" (Thorson et al., 2014).

Estimation proceeds by assuming that coefficients are "exchangeable," for example assuming that they following a multivariate normal distribution, $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma_{RE}^2 \mathbf{R})$, where $\mathbf{R}$ is the correlation among random effects and σ_{RE}^2 is the variance of random effects that can be estimated from data. These coefficients $\boldsymbol{\varepsilon}$ are "integrated out" from the marginal likelihood, such that increased sampling leads to increased information about hyperparameters θ and/or predicted values for random effects. There is ongoing research exploring different distributions for the optimal distribution for random effects to approximate different time-varying processes, often specifying random, autocorrelated, or other distributional forms for correlation $\mathbf{R}$ (Xu et al., 2019), although we do not have space to fully discuss these differences here.

For example, a state-space age-structured model (Gudmundsson, 1994; Nielsen and Berg, 2014; Stock et al., 2021) might instead specify as the survival function:

$$\log\left(N_{a+1,y+1}\right) = \log\left(N_{a,y}\right) - M_{a,y} - F_{a,y} + \varepsilon_{a,y} \quad (5)$$

where $\varepsilon_{a,y} \sim Normal(0, \sigma_\varepsilon^2)$ in this case represents the assumption that residual variation in the survival function is independent and homoscedastic. In this case, if sampling data are unbiased ($B_{sample} = 0$) and sampling errors decrease asymptotically with increased effort ($V_{sample} \to 0$), then $N_{a+1,y+1}$ and $N_{a,y}$ could both approach their true values even given immigration or other unmodeled processes. This can be seen as a corollary of the Bayesian Central Limit Theorem (a.k.a. Bernstein von-Mises theorem, (Doob, 1949)), where the specified distribution for random effects has decreasing importance as the data increase asymptotically. We therefore see that random effects will typically result in additional variance; in this example, the variance of $\varepsilon_{a,y}$ causes additional variance in $\log\left(N_{a,y}\right)$, and we call the resulting imprecision V_{model}. This imprecision V_{model} typically increases with increasing variance σ_{RE}^2 of process errors. Similarly, this imprecision V_{model} will typically decrease as more data become available, because the predicted random effects will typically have a lower standard error (Xu et al., 2019).

Including random effects can decrease the errors B_{model} that would otherwise arise when the data-generating process is not nested within the specified demographic model (Thorson et al., 2014). In other cases, a model might include random effects but include them in the wrong part of the model such that it still does not include the true data-generating process as a nested submodel. For example, an analyst might instead specify a random effect for fishery selectivity (Xu et al., 2019):

$$\log(N_{a+1,y+1}) = \log(N_{a,y}) - M_{a,y} - F_{a,y}e^{\varepsilon_{a,y}} \quad (6)$$

where, for example, $\varepsilon_{a,y}$ follows a two-dimensional smoother across years and ages. In this case, the model is more flexible but still specifies $N_{a+1,y+1} \leq N_{a,y}$. If true abundance then increases for a given cohort due to immigration, the Bayesian central limit theorem does not apply, and model mis-specification (in this case, ignoring immigration) will result in an inconsistent estimate (i.e., increasing B_{model}) rather representing additional imprecision (i.e., increasing V_{model}).

3.5 Measuring the variance of four errors

Past research (Francis, 2011; Miller and Skalski, 2006; Thorson et al., 2020) has noted that we can identify an estimator for sampling variance, $\hat{V}_{sample}(t)$ in each year t, using the bootstrap, model, or design-based estimators outlined previously. These are calculated directly from raw sampling data, and do not require any specific knowledge about the assessment model itself (although a difference between the population being sampled vs. modeled will result in model inconsistency as noted previously). These estimates of sampling variance $\hat{V}_{sample}(t)$ themselves have a standard error (Kotwicki and Ono, 2019), but for simplicity of presentation we do not further discuss the implications of the standard error of this or other variance terms.

Similarly, past research (Francis, 2014, 2011; Pennington and Godø, 1995) has used the squared Pearson residuals from the fit to a stock-assessment model as an estimator of the total squared errors, $\hat{E}^2$, and presumably this can be generalized via proper transformation of OSA residuals. We briefly note that these residuals are calculated as the difference between observations and predictions, and predictions for a given fleet are leveraged by data from that and other fleets in multi-fleet assessment models. In the following, we assume that these cross-fleet correlations in residuals are negligible, and we encourage further research regarding variance decompositions that account for multi-fleet leverage in calculating residuals.

Estimators for sampling imprecision $\hat{V}_{sample}(t)$ and total squared-errors $\hat{E}^2$ then result in an estimable decomposition of stock-assessment errors:

$$\hat{E}^2 = \hat{V}_{sample} + \underbrace{B_{sample}^2 + V_{model} + B_{model}^2}_{\text{residual error}} \tag{7}$$

where the variance arising from mis-specified sampling designs, parametric, and semi-parametric model errors are all captured in the residual "residual error" term.

3.6 Implications of error partitioning

Before proceeding further, we note that this decomposition extends previously published studies in several important ways:

1. *Revised law of conflicting data*: Maunder and Piner (2017) define the "Law of conflicting data" as "since data are facts, conflicting data implies model misspecification, but must be interpreted in the context of random sampling error". However, our presentation emphasizes that fisheries data such as fishery catch, abundance indices, and age/length compositions are typically expanded from raw observations. We agree that these raw observations are "fixed" with respect to an annual assessment modelling process[1], and any failure to fit fixed data

[1] In reality, even tow-level data are not strictly "fixed" and instead typically arise from a process of prior analysis. For example, the area-swept in a bottom trawl survey is often calculated from reconstructing a transect from a series of GPS records of a vessel during net deployment, with time-on-bottom reconstructed from assumptions about how to extrapolate newer net sensors to predict bottom contract from vessel speed and tow depth. In these and other cases, "fixed" tow-level data are subject to updates from improved process research. However, we agree that these updates to sample-level data usually occur via a slower scientific process than an operational stock assessment, and tow-level data can be considered "fixed" with respect to a given stock assessment.

implies model mis-specification. However, alternative expansion estimators will result in different sampling imprecision V_{sample} and sampling bias/inconsistency B_{sample}. For example, it is feasible to expand bottom trawl survey data while either ignoring or using auxiliary data to correct for the emigration of fishes outside of the spatial domain of the primary survey (O'Leary et al., 2020). Using auxiliary and spatially unbalanced data to estimate abundance across an expanded spatial footprint may simultaneously increase sampling imprecision V_{sample} and decrease sampling inconsistency B_{sample}. We therefore propose a Revised Law of conflicting data:

"Data are facts but are often pre-processed (using a design- or model-based estimator) prior to being fitted in a stock assessment model. Therefore, conflicting data implies model misspecification in either or both the assessment model, sampling design, or pre-processing analysis."

2. *Model imprecision vs. inconsistency*: Francis (2011) decomposes total error into process and measurement errors, and Francis (2017) notes that state-space models further decompose "process errors" into time-varying parameters, errors in fixing parameters, or specifying the wrong mathematical form. We formalize this latter decomposition by separating model inconsistency (i.e., mis-specification of fixed parameters or mathematical expressions that will result in error regardless of the quantity of data) from model imprecision (i.e., variation within the specified distribution of the random effect, but where increasing data will allow random effects to converge on the true value). The Bayesian Central Limit Theorem implies that the distribution assigned to random effects has decreasing importance as the quantity of

data increases. As a result, estimates of stock dynamics for a data-rich assessment with suitable random effects can therefore approach the true dynamics even given mis-specification of the population dynamics assumptions (e.g. Thorson et al., 2014), and the distinction between model inconsistency and imprecision is particularly relevant for data-rich assessments.

3. *Calculating excess variance as diagnostic for model mis-specification:* Using the multinomial distribution (Eq. 1), analysts often calculate a "sample size" as proportional to the reciprocal of each variance term. This arises because the multinomial distribution $\mathbf{n} \sim Multinomial(\boldsymbol{\pi}, N)$ for a proportion $p_a = n_a / \sum_{a'=1}^{A} n_{a'}$ has variance that is inversely related to sample size, $\mathrm{Var}(p_a) = \frac{\pi_a(1-\pi_a)}{N}$. We can therefore calculate the variance from expanding composition data $\mathrm{Var}(p_a)$ and convert this to an equivalent sample size $N_a = \frac{\pi_a(1-\pi_a)}{Var(p_a)}$ and define input sample size n_{input} as the harmonic mean across ages. Similarly, we can calculate the sample variance from residuals as an estimator of total squared-errors, and convert this to an effective sample size $n_{effective}$. Plugging these into Eq. 3 and re-arranging, we see that:

$$PEV = 1 - \frac{n_{effective}}{n_{input}} = \frac{B_{sample}^2 + V_{model} + B_{model}^2}{E^2} \quad (8)$$

e.g., where we define the "proportion excess variance" PEV as the proportion of squared assessment-model residuals that results from survey bias as well as bias and imprecision in the assessment model itself. PEV is then a measurable and interpretable diagnostic (ranging from 0 to 1) for the magnitude of error in those processes. Although *PEV* becomes harder to interpret in multi-fleet models (given that $n_{effective}$ is affected by fits to other fleets), we still believe that simplified and high-level statistics can elucidate theory and complement more complicated diagnostics such as OSA residuals.

For these three reasons, we believe that it is warranted to decompose error into imprecision and bias/inconsistency arising for both the sampling design/expansion and stock-assessment model.

3.7 Case study demonstration

We next provide a simple demonstration of the potential use of percent excess variance (PEV) to diagnose assessment model mis-specification or bias in the available data (see Appendix A for details). To do so, we develop a state-space age-structured assessment model using the Woods Hole Assessment Model (Stock and Miller, 2021) for Gulf of Alaska walleye pollock that closely matches the 2021 stock assessment (Monnahan et al., 2021a). This involves setting an input-sample size N_{input} for age-composition data for each of five fleets. We use a bootstrap estimator to calculate N_{input} for the NMFS bottom trawl survey (Hulson et al., 2023), fix N_{input} as number of midwater trawls for the two acoustic surveys, but do not have software to estimate the value for the fishery or the Alaska Department of Fish and Game (ADF&G) bottom trawl survey. We therefore fix a value for the fishery larger than the survey (i.e., $N_{input} = 1000$), and simulate data conditional upon this known true value. We condition our simulation upon estimates of process errors from the fit to real-world data, specifically time-varying fishery selectivity and time-varying catchability for abundance indices, so that the model represents observed dynamics for this stock.

We then fit a single replicate from this simulation using two alternative models:

1. *Mis-specified*: We first fit a model that assumes fishery selectivity and survey catchabilities are constant over time. This then represents a known source of mis-specification, given that the simulation model includes these time-varying processes.
2. *Correctly specified*: We also fit the same model but with time-varying fishery selectivity and survey catchabilities matching the structure of the simulation model (but estimating the magnitude of process errors).

$N_{effective}$ was estimated jointly with the model using the linear version of the Dirichlet-multinomial likelihood (Thorson et al., 2017). The estimated PEV (Eq. 8) for the fishery was 77.1% when fitted with a model that did not include time-varying fishery selectivity (Table S1), and this PEV was substantially larger than for any other fleet. When refitting with a model that included time-varying fishery selectivity, PEV was reduced to 0.0%. We compared estimates of the variance (0.256) and autocorrelation (0.989) for time-varying fishery selectivity between the simulation and correctly specified estimation model. The confidence interval in untransformed space for the estimated variance contained the true value (0.275), but not for the estimated autocorrelation correlation (0.898). We therefore conclude that PEV was able to identify which fleet was subject to some mis-specification, and also that the process-error variance could be usefully estimated in part due to the implicit upper bound provided by the input sample size.

4. Practical recommendations for applied stock assessments

Having categorized errors into four potential sources, we next discuss implications of this categorization (Table 2) while also proposing specific recommendations for stock-assessment practices (Table 3).

4.1 Fit proportions-at-age separately from total abundance or catch

As noted, state-space models such as SAM (Nielsen and Berg, 2014) are sometimes fitted to abundance-at-age $n_{a,y}$, which can be thought of as a product of an abundance index and proportions-at-age $n_y p_{a,y}$. However, the variance of total abundance is often lower variance the sum of variances for each abundance-at-age, i.e., $\mathrm{Var}(n_y) < \sum_{a=1}^{A} \mathrm{Var}(n_{a,y})$. Presumably such an outcome can be approximated via covariances among ages in a specified measurement covariance matrix (Berg and Nielsen, 2016). However, state-space models are sometimes fitted using a lognormal distribution for abundance-at-age (Nielsen and Berg, 2014). In this case, there is no linear combination of variances and covariances for log-abundance-at-age that will match the sampling variance of the total abundance index.

To illustrate this in more detail, imagine a fishery with nearly perfect observer coverage, but where observers can only measure length for a subsample of individuals. In this case, the overall removals c_t might be known (almost) exactly, and this corresponds to small variance in management performance (i.e., whether the fishery is catching above or below its catch quota). However, the removals-at-age $c_{a,y}$ will still have a substantial variance due to finite sample sizes for subsampled lengths. If fitting to log-removals-at-age, then a series of positive or negative residuals across ages could result in predicted removals-at-age that differ greatly from the (close-to-) known total removals when summed across ages. Even if a measurement covariance matrix with negative correlations results in small variance for $\mathrm{Var}\left(\sum_{a=1}^{A} \log(c_{a,y})\right)$, this ensures that the estimate $\sum_{a=1}^{A} \log(c_{a,y})$ approaches the measurement $\log(c_t)$ but it gives equal weight to residuals in $\log(c_{a,y})$ for ages with small and large removals. In other cases, both removals-at-age $c_{a,y}$ and total removals c_t are both imprecisely measured. In these cases, it might result in better fit to model removals at age rather than separately modelling proportions and totals (e.g., Albertsen, 2018 see Section 3.3.2.1). We note that both options are available in SAM, and empirical analyses with commercial fisheries have shown mixed support for these where North Sea cod and Northeast Arctic haddock were best fitted by abundance-at-age while Northern Shelf haddock and blue whiting were fitted better by modelling proportions-at-age (Albertsen et al., 2017). To address this:

Recommendation #1: We recommend that assessment models include options to specify a vector for abundance indices or removals across years, and a separate matrix for proportions-at-age across years, as alternative to fitting directly to the product of two. This ensures that a small variance in measurements of total removals or total abundance is appropriately propagated even when proportions are less precise.

4.2: Calculate sampling imprecision and inconsistency as starting point to interpret fit

We previously decomposed total error into components due to imprecision or inconsistency in either the field sampling or assessment model (Eq. 3). We then clarified that the variance arising from model imprecision and both data and model inconsistency are not estimable without auxiliary data. It is widely understood (but still not widely used in practice) that the imprecision of field-sampling data $\hat{V}_{sample}$ can be estimated using bootstrap, model, or design-based estimators (Berg and Nielsen, 2016; Miller and Skalski, 2006; Stewart and Hamel, 2014; Thorson and Haltuch, 2018). The length and age subsampling for commercial fisheries are often not available outside of national laboratories. In these cases, it might be necessary in multinational jurisdictions (i.e., ICES) to standardize analytical methods that can then be done independently on confidential data, such that the estimated imprecision $\hat{V}_{sample}$ can be shared even when the raw data cannot.

Equally important but less commonly understood is the fact that auxiliary data can in some cases be used to define an explicit lower bound on the unknown variance of sampling inconsistency, $B_{sample} \geq \hat{B}_{lower}$, where $\hat{B}_{lower}$ is then estimated externally from auxiliary information. As discussed previously, sampling inconsistency arises when the sampling frame for a fishery or survey does not contain the entire fishery or stock that is intended. In some cases, auxiliary data can be used to measure what portion of the stock is outside of the sampling frame, and hence estimate the sampling inconsistency resulting from that process. For example:

- Vertical survey availability: A bottom trawl survey will often miss the portion of a stock that is above the effective fishing height, and this portion can be estimated using auxiliary acoustic and midwater sampling information (Monnahan et al., 2021b);
- Horizontal survey availability: Similarly, stocks can migrate into or emigrate beyond the spatial footprint of the surveys that have been defined previously, and the portion outside can be identified in some cases using data from adjacent surveys (O'Leary et al., 2022);

In these and other cases, we can use auxiliary sampling data (e.g., from nearby surveys, tags, etc.) to measure some components of the bias $\hat{B}_{lower}$ arising from survey availability, knowing that B_{sample} must be greater than that bias.

This lower bound on survey bias $\hat{B}_{lower}$ then provides an implicit upper bound on the variance that can be attributed to "assessment model imprecision". This is because we can directly measure total squared-errors $\hat{E}^2$ from model residuals, sampling imprecision

$\hat{V}_{sample}$ from expansion methods, and in this hypothetical also have a lower bound on sampling bias, $B_{sample} \geq \hat{B}_{lower}$. Plugging into Eq. 6 and re-arranging yields:

$$\underbrace{V_{model} + B_{model}^2}_{\text{assessment model errors}} \leq \hat{E}^2 - \hat{V}_{sample} - \hat{B}_{lower}^2 \tag{9}$$

This is helpful because the assessment-model imprecision V_{model} is an increasing function of the variance of random effects, σ_{RE}^2. Because the unexplained variance $\hat{E}^2 - \hat{V}_{sample} - \hat{B}_{lower}^2$ provides an explicit upper bound on assessment model errors $V_{model} + B_{model}^2$, it also provides on implicit upper bound on random-effect variances σ_{RE}^2, where this exact bound depends on how σ_{RE}^2 affects V_{model} as determined by the structure of the assessment model and the specified random effects. One way to interpret this inequality is that, as more sources of "sampling bias" are identified (i.e., $\hat{B}_{lower}^2$ increases), there is less need to invoke time-varying processes (and estimate a large variance for random effects) to explain a lack-of-fit for that data source.

In summary:

Recommendation #2: We recommend using design-, model-, or bootstrap estimators to identify the variance of all data inputs, as well as auxiliary information where available to identify the variance arising from errors in the sampling frame;

Recommendation #3: We recommend providing the variance of each data input (including the estimated imprecision of age and length compositions) to the stock assessment model 'a priori', and comparing this variance with the variance of residuals to quantify the proportion of unexplained variance. This PUV could then be used as diagnostic to identify when data should be further downweighted (or less important fleets), or additional time-varying processes considered (for more important fleets). We also recommend using auxiliary data to measure a *lower bound on the variance arising from survey bias, so that the model will not estimate a variance for random effects that results in a tighter fit to survey products than is warranted given this lower bound on survey bias. This then ensures that the variance of data inputs serves as an implicit "upper bound" on the variance of estimated random effects.*

4.3: Approximate sample size as simple currency

Despite the several studies demonstrating how to estimate the sampling variance $V_{sample}(t)$ from available data (including abundance indices over time and composition data over time and age/length/sex) we are not aware of any operational stock assessments (particularly commonly used general stock assessment packages) inputting a covariance matrix to represent sampling imprecision. By contrast, a large number of operational stock assessments specify a scalar (whether a multinomial sample size or the lognormal standard deviation) representing sampling imprecision. We therefore recommend replacing the sampling covariance among ages or lengths with input-sample size, n_{input}. This is then interpreted as an approximation that both (1) simplifies the number of inputs that must be into a stock assessment, and (2) simplifies intuition about the relative leverage of different years. This will inevitably lose information about the sampling covariance among ages or lengths, but we hypothesize that this is necessary to simplify the process sufficiently to achieve uptake in real-world assessments.

Measuring input sample size is then useful because:

1. it provides an implicit upper bound on the variance of random effects (similar to the role for $\hat{B}_{lower}$). To see this, we again inspect Eq. 9, where a decrease in input-sample-size (and resulting increase in $\hat{V}_{sample}$) causes a decrease in the upper bound on assessment model bias

and imprecision, $V_{model} + B_{model}^2$ and an in the implicit upper bound of σ_{RE}^2. These random-effect variances are often difficult to estimate, so information about their bounds is likely helpful;

2. It allows us to calculate excess variance PEV (Eq. 8) as simple diagnostic for residual forms of survey and model mis-specification.

Recommendation #4: If analysts choose not use the estimated sampling variance $\hat{V}_{survey}$ *within the stock assessment, we recommend as practical alternative that they replacing this with a single scalar quantity, "input sample size", representing the idealized multinomial sampling size with approximately similar variance. Adding additional random effects (i.e., model imprecision) will then result in smaller model residuals, and an "effective sample size" that approaches this input sample size (i.e., excess variance approaching zero). Similarly, the "input sample size" provides an implicit upper bound on the variance of random effects.*

4.4: Correct residuals via model expansion rather than data weighting

We now finally turn to the question that is central to previous discussions of "data weighting": Is there a probability distribution that we can specify for compositional data such that it eliminates problems arising from a lack of fit? We here argue that, no, using a generalized distribution that "downweights" data is likely better than using a made-up value for data weights, but that it is also better still to add additional model flexibility in other parametric ways (i.e., fix model inconsistency) or semi-parametric ways (add random effects).

To see this, we first briefly review the literature on generalized distributions or algorithms that can down-weight data (see Table 4). First, McAllister and Ianelli (1997:

Appendix 2) noted that the variance of an idealized multinomial distribution will have residual variance:

$$\left(p_{a,y} - \pi_{a,y}\right)^2 = \frac{\pi_{a,y}(1 - \pi_{a,y})}{n_{a,y}^*} \tag{10}$$

which then yields a formula for effective sample size $n_{effective} = n_y^{-1} n_a^{-1} \sum_{y=1}^{Y} \sum_{a=1}^{A} n_{a,y}^*$. Subsequently, Candy (2008) proposed using the default "saturating" parameterization of the Dirichlet-multinomial to estimate an additional parameter β representing the variance of a Dirichlet process that generates additional variance in compositional data. Thorson et al. (2017) later extended this by introducing the "linear" parameterization, where parameter $\theta = n_{input}\beta$ such that $\log(\theta) \approx \text{logit}(\frac{n_{effective}}{n_{input}})$ or equivalently $n_{effective} \approx \frac{\theta}{1+\theta} n_{input}$, such that $\frac{\theta}{1+\theta}$ results in a close-to-proportional decrease in data-weight for all compositions regardless of their assigned n_{input} (e.g., in Fig 2 of Fisch et al., 2022). This compound-distribution approach was later extended using a "multivariate-Tweedie" distribution to more closely resemble the process of expanding compositional data in a multi-level sampling design (Thorson et al., 2022).

As alternative approach, Francis (2011: Eq. TA1.8) extended Pennington and Volstad (1994) by instead modelling the variance in the average age or length for observations $\bar{p}_y$ and expectations $\bar{\pi}_y$. This "Francis method" has the stated advantage that calculating the variance of average age or length accounts for both the variance and covariance of residuals. This method was subsequently extended to conditional age-at-length data (Punt, In press).

Finally, research has also developed either the additive (Miller et al., 2016; Schnute and Haigh, 2007; Stock and Miller, 2021) or multiplicative (Cadigan, 2016) versions of a logistic-normal distribution. These two versions transform the composition data $n_{a,y} / \sum_{a'=1}^{A} n_{a',y}$ using two flavors of a multivariate inverse-logistic function, and do the same with the predicted proportions $\pi_{a,y}$, and then compute the discrepancy between these two using a multivariate normal distribution. Many papers have subsequently compared different subsets of these various methods (Cronin-Fine and Punt, 2021; Fisch et al., 2022, 2021; Hulson et al., 2012, 2011; Punt, In press; Xu et al., 2020), although results are difficult to compare among studies due to different parameterizations being used and different scenarios being tested.

As discussed extensively elsewhere, these options can be derived by assuming that there is some additional "overdispersion" process that generates variation in the observed vector $n_{a,y} / \sum_{a'=1}^{A} n_{a',y}$. Using the Dirichlet-multinomial for simplified discussion, this process involves taking a draw from a Dirichlet distribution:

$$\boldsymbol{\pi}_y^* \sim \text{Dirichlet}(\beta \boldsymbol{\pi}_y) \tag{11}$$

where β controls the variance of this process, and then using this simulated proportion $\boldsymbol{\pi}_y^*$ to fit the data using a multinomial distribution:

$$\mathbf{n}_y^* \sim \text{Multinomial}(\boldsymbol{\pi}_y^*, n^*) \tag{12}$$

By contrast, in the Francis, McAllister-Ianelli, or logistic-normal models the process generating overdispersion is implicit in the derivation (Francis, 2014, 2011; McAllister and Ianelli, 1997). However, these distributions generally differ in several ways:

1. *Fitting to zeros*: The Dirichlet-multinomial, Francis, multivariate-Tweedie, and McAllister-Ianelli methods can all be fitted to composition data that includes zeros, while the logistic-normal cannot and presumably the data must be modified to avoid zeros (e.g. combining age/length bins or adding a constant) prior to model fitting, or expanded as a zero-inflated process;

2. *One- or two-stage fits*: The Dirichlet-multinomial, multivariate-Tweedie and logistic-normal involve estimating overdispersion using parameters that can be fitted at the same time as other model parameters, while the Francis and McAllister-Ianelli methods cannot. The latter therefore require fitting a model, then adjusting the sample sizes being used, and refitting. This iterative process is sometimes called "two-stage estimation" although in practice it might require many more than two fits and there is little consistency regarding how many times to refit.
3. *Estimating residual correlations*: Dirichlet-multinomial, multivariate-Tweedie and McAllister-Ianelli methods identify overdispersion but do not calculate or use information about correlations among ages or years. By contrast, the Francis method accounts for correlations among ages when calculating the observed and expected average age, and implicitly downweights when correlations are large. Similarly, the logistic-normal can be extended to estimate the magnitude of correlations among ages. However, neither Francis not logistic-normal methods account for correlations among years.

These theoretical and practical differences presumably cause analysts to select different methods for real-world use.

What has generally gone undiscussed in this extensive literature is that residuals in composition data also reflect mis-specification that affects the interpretation of other data (removals or abundance-indices) from that same fleet, as well as reference points calculated for that fleet. For example, samples of the age-composition from fishery catches might have positive correlations for older ages and negative for younger ages in a given year. If these correlations are larger than expected for a multinomial distribution, then data suggests that the fishery likely did, in fact, target older ages in that year. This could arise due to the fishery targeting a spatial component of the stock where older ages aggregate, or due to less strict restrictions on bycatch that allow targeting high-profit areas that were previously avoided. In either case, it is critical that this information about fishery removals be used to properly interpret other components of the model. In this example:

1. Higher selectivity for old individuals also likely means that a lower catch (in numbers) can explain total removals (as measured in biomass). Treating correlations as a residual process that only affects fishery comps then ignores the implications for fitting (or conditioning upon) fishery removals for that fleet;
2. Higher selectivity for old individuals also likely has large implications for calculating yield per recruit and spawning biomass per recruit. Spawning biomass per recruit is in turn typically used to calculate spawning potential ratio (SPR). Attributing residual patterns in fishery comps to a residual "observation" process likely ignores the implications for SPR target and limit calculations.

In this light:

Recommendation #5: We recommend that analysts use OSA instead of Pearson residuals, to account for the action of any random effects and also any non-normal error distributions. We similarly recommend that these residuals be visualized, where patterns among ages and years can be used to diagnose model-specification.

Recommendation #6: We recommend that model weighting be considered only as a first-pass response to overdispersion, and that assessment scientists additionally seek to attribute residual patterns to additional model processes for important fleets (fisheries with a large portion of total removals, or trusted surveys). This is necessary to ensure that overdispersion (and any *correlation among ages and years) is interpreted not just for fitting age/length compositions, but also when (1) fitting to abundance indices and removals or (2) calculating reference points and management quantities from that same fleet. For less important fleets (e.g., fisheries with a small fraction of removals), it might be less important to propagate information from age and length-composition residuals when interpreting removals and references points, so for these less-important fleets it is more defensible to use data-weighting without further investigation.*

4.5 Collect and synthesize auxiliary information that can mitigate sampling inconsistency

As we discussed previously, assessment error can be decomposed into imprecision and inconsistency resulting from both sampling and assessment-model specification. When residuals are overdispersed for the composition data of a given fleet, assessment scientists often downweight these data using one or more data-weighting algorithms. However, the past decade has also seen increased interest in model-based methods to expand sampling data. These estimators can improve statistical efficiency (decrease V_{sample}) or mitigate sampling bias (decrease B_{sample}), and we discuss these respectively here.

In some cases, model-based estimators can improve sampling efficiency and therefore reduce "sampling imprecision" (i.e., improve statistical efficiency). For example, an efficient sampling design will allocate samples in proportion to the population variance. However, some species with a patchy distribution will have a substantial fraction of total survey catch in one or a few tows (Thorson et al., 2011). In these cases, a design-based algorithm will be driven predominantly by the small number of extreme catches, and this will obscure the useful signal that otherwise justifies conducting a survey. The statistical efficiency for this fixed design can in some cases be increased using a model-based estimator (Thorson et al., 2015), and in some cases this decreased imprecision can then be seen to propagate through the assessment model and result in a higher effective sample size (Thorson and Haltuch, 2018).

More usefully, though, model-based estimators can also be designed to use auxiliary information to estimate or even reduce the magnitude of "sampling inconsistency". In these cases, model-based estimators seek to minimize bias that arises when using survey data that are not representative of the modeled stock. For example, changes in regional habitat might increase the proportion of the stock that is expected to occur outside of a given sampling design. For yellowfin sole in the eastern Bering Sea, for example, spring warmth drives the timing of movement from offshore to onshore habitats where warm temperatures increase the overlap with the summertime survey (Wilderbuer et al., 1992), and this effect can then be corroborated when fitting a temperature-dependent catchability coefficient representing survey availability in the stock assessment (Nichol et al., 2019). Rather than fitting an additional catchability-coefficient in the assessment model, however, it might be feasible to combine fishery and survey data to jointly estimate the timing of movement and the abundance that would have resulted at a standardized time in seasonal migration. A similar approach has been done, e.g., using larval otoliths to back-calculate the timing of a winter survey relative to winter spawn timing for Gulf of Alaska walleye pollock (Rogers and Dougherty, 2019).

In summary:

Recommendation #7: We recommend research to identify auxiliary data (whether combining habitat information, multiple surveys, or process research) that can be used to decrease sampling imprecision and inconsistency, which otherwise result in downweighting of composition data. This research will typically occur in parallel to an operational assessment, *and in some cases can be done by survey teams and reviewed during Methods Reviews that operate in parallel to operational stock assessment reviews.*

4.6 Provide a rationale if substantially downweighting individual data sets

As discussed previously, data are typically downweighted due to a combination of survey and model imprecision and inconsistency. However, assessment-model imprecision and inconsistency is likely to cause errors in fitting data for multiple fleets. Downweighting a single fleet while leaving another with larger weight corresponds to a hypothesis about the sources of error (presumably in that case, the error for the downweighted fleet arises from sampling inconsistency). In the context of fitting abundance indices, past studies have cautioned against taking the average of multiple indices as if it were the only potential outcome (Schnute and Hilborn, 1993; Walters and Maguire, 1996). This same intuition applies when downweighting composition data, such that the resulting assessment might be driven by only those data that are weighted more highly. Similarly, Francis (Francis, 2017, 2014, 2011) proposes a “rule of thumb” that, when abundance indices and composition data conflict, it is likely the abundance index that is trustworthy. However, this rule-of-thumb will clearly break down, e.g., when the survey is not representative of the stock but age/size structure is relatively homogenous. In this light:

Recommendation #8: We recommend that data weighting be interpreted as a data-driven and explicit hypothesis about the sources of error, including model and survey imprecision and inconsistency, and ideally that the sensitivity to these choices be presented to highlight remaining uncertainties about errors. In cases when no data are available to evaluate these *alternative hypotheses, an ensemble of models can be used to communicate resulting uncertainty, or justification provided for the decision of what data to downweight or not.*

5. Where do we go from here?

Finally, we conclude by recommending a few priorities for future development and research. These include (1) improved diagnostics and guidance for what assessment-model changes (including time-varying parameters) to explore when initial model fits suggest a substantial downweighting for data, and (2) and establishing an iterative process linking assessment-model fit to coordinated research regarding sampling inconsistency. We conclude by briefly discussing each of these.

5.1 Improved diagnostics and guidance for time-varying processes

Composition data are often re-weighted by default because no analysis has been conducted to estimate an appropriate input-sample size. Analysts should seek to fix these cases, using known methods to estimate input-sample-size (see Recommendations #2/4). Even when this is done, however, there will still be cases when data are poorly fitted and initial model-based re-weighting suggests substantial downweighting (i.e., $PEV > 0.5$). In these cases, an assessment scientist will be faced with many potential options for additional model changes to improve fit. These include adding time-varying selectivity, improving the specification of growth, using a spatially stratified model, or many other options. However, there is little practical guidance available for the steps an analyst should follow in revising their model to improve the fit such that effective sample size approaches input sample size. We therefore recommend research regarding:

1. identifying a threshold for excess variance PEV that should trigger additional exploration;
2. statistical diagnostics to identify the likely process (i.e., time-varying growth, selectivity, etc.) that can explain the lack-of-fit in a given model;
3. the consequences of mis-specifying which process is time-varying, ideally identifying a procedure that minimizes the risk of mis-specification across a wide range of states-of-nature (i.e., a minimax justification for specifying time-varying processes, see e.g., Szuwalski et al. (2018)); and
4. methods to build an ensemble of models representing alternative hypotheses about the process causing poor fit.

Studies along these lines could then contribute to a "cook-book" of potential responses when initial fits suggest a high excess variance.

5.2 Iterative process linking assessment-model fit to sampling inconsistency

In some cases, initial model fits will identify that data must be downweighted and subsequent model expansion will provide a clear avenue for revising the model and thereby decrease excess variance below an acceptable threshold. For example, the eastern Bering Sea pollock stock assessment includes a non-parametric model for time-varying survey selectivity (Ianelli et al., 2018). This improves the fit to survey age-composition data while ensuring that results are also used when interpreting the survey abundance index. However, subsequent research has sought to attribute this time-varying selectivity to the vertical distribution of pollock and their resulting availability to different bottom-trawl vs. midwater acoustic survey gears (Kotwicki et al., 2015; Monnahan et al., 2021b). This example illustrates that data-weighting can be a starting point for further coordinated research (involving stock-assessment, survey, and other scientists). In particular, this research would seek to transition from an estimated time-varying parameter in a stock-assessment model (i.e., "estimation") to an improved process for measuring the time-varying process directly in nature, and thereby provide an updated data set that accounts for that process in a more rich set of data (i.e., "monitoring"). We realize that this process is likely expensive and therefore only practical to implement for the most important stocks, but also see that it is an important goalpost for directing research and development for all stock assessments.

6. Summary and conclusions

In this paper, we provide a more formal basis for discussing "data-weighting" by decomposing lack of fit into either imprecision or bias in either field-sampling or assessment modelling steps of a stock assessment (Table1). We then discussed implications of this decomposition (Table 2) and provided several short-term recommendations (Table 3), emphasizing the importance of quantify sampling imprecision for composition data using an input-sample-size that can be routinely computed using design- and model-based methods. We concluded by outlining long-term research recommendations, including the need to establish a useful threshold for excess variance, and developing an interactive process for linking data-weighting back to improved data collection and processing. We hope that future discussions of data-weighting will recognize that data-weighting is not simply a concern for stock-assessment scientists when tuning a model, but instead provides a way to broadly organize research spanning modelling, survey, and other fisheries scientists focused on explaining the complex processes affecting ocean populations.

Acknowledgements

Thanks to J. Ianelli, M. Maunder, T. Miller, and C. Albertsen for comments on an earlier draft. We also thank M. Maunder, A. Punt, and the Food and Agriculture Organization (FAO) for hosting the Center for the Advancement of Population Assessment Methodology (CAPAM) “Stock Assessment Good Practices Workshop” October 24-28, 2022 in Rome. Finally, we thank the many individuals who attended that workshop; the text is modified from that original version to address comments received during the discussion on “data-weighting.”

Works cited

Ailloud, L.E., Hoenig, J.M., 2019. A general theory of age-length keys: combining the forward and inverse keys to estimate age composition from incomplete data. ICES J. Mar. Sci. 76, 1515–1523. https://doi.org/10.1093/icesjms/fsz072

Albertsen, C.M., 2018. State-space modelling in marine science (PhD Thesis). PhD Thesis. Technical University of Denmark, National Institute of Aquatic ….

Albertsen, C.M., Nielsen, A., Thygesen, U.H., 2017. Choosing the observational likelihood in state-space stock assessment models. Can. J. Fish. Aquat. Sci. 74, 779–789. https://doi.org/10.1139/cjfas-2015-0532

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

Cadigan, N.G., 2016. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Can. J. Fish. Aquat. Sci. 73, 296–308. https://doi.org/10.1139/cjfas-2015-0047

Candy, S.G., 2008. Estimation of effective sample size for catch-at-age and catch-at-length data using simulated data from the Dirichlet-multinomial distribution. CCAMLR Sci. 15, 115–138.

Cochran, W.G., 1977. Sampling Techniques, 3rd Edition, 3rd ed. John Wiley & Sons.

Crone, P.R., Sampson, D.B., 1997. Evaluation of assumed error structure in stock assessment models that use sample estimates of age composition, in: Int. Symp. on Fishery Stock Assessment Models for the 21st Century, Anchorage, Alaska, EEUU. 8Á11 October.

Cronin-Fine, L., Punt, A.E., 2021. Modeling time-varying selectivity in size-structured assessment models. Fish. Res. 239, 105927. https://doi.org/10.1016/j.fishres.2021.105927

Doob, J.L., 1949. Application of the theory of martingales. Calc. Probab. Ses Appl. 23–27.

Fisch, N., Ahrens, R., Shertzer, K., Camp, E., 2022. An empirical comparison of alternative likelihood formulations for composition data, with application to cobia and Pacific hake. Can. J. Fish. Aquat. Sci. 79, 1745–1764. https://doi.org/10.1139/cjfas-2022-0036

Fisch, N., Camp, E., Shertzer, K., Ahrens, R., 2021. Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. Fish. Res. 243, 106069. https://doi.org/10.1016/j.fishres.2021.106069

Francis, R.I.C.C., 2017. Quantifying annual variation in catchability for commercial and research fishing. Fish. Res., Data conflict and weighting, likelihood functions, and process error 192, 5–15. https://doi.org/10.1016/j.fishres.2016.06.006

Francis, R.I.C.C., 2014. Replacing the multinomial in stock assessment models: A first step. Fish. Res. 151, 70–84. https://doi.org/10.1016/j.fishres.2013.12.015

Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138.

Gudmundsson, G., 1994. Time Series Analysis of Catch-At-Age Observations. J. R. Stat. Soc. Ser. C Appl. Stat. 43, 117–126. https://doi.org/10.2307/2986116

Hilborn, R., Amoroso, R.O., Anderson, C.M., Baum, J.K., Branch, T.A., Costello, C., de Moor, C.L., Faraj, A., Hively, D., Jensen, O.P., Kurota, H., Little, L.R., Mace, P., McClanahan, T., Melnychuk, M.C., Minto, C., Osio, G.C., Parma, A.M., Pons, M., Segurado, S., Szuwalski, C.S., Wilson, J.R., Ye, Y., 2020. Effective fisheries management instrumental in improving fish stock status. Proc. Natl. Acad. Sci. 117, 2218–2224. https://doi.org/10.1073/pnas.1909726116

Hulson, P.J.F., Hanselman, D.H., Quinn, T.J., 2012. Determining effective sample size in integrated age-structured assessment models. ICES J. Mar. Sci. J. Cons. 69, 281–292.

Hulson, P.J.F., Hanselman, D.H., Quinn, T.J., 2011. Effects of process and observation errors on effective sample size of fishery and survey age and length composition using variance ratio and likelihood methods. ICES J. Mar. Sci. J. Cons. 68, 1548–1557.

Hulson, P.-J.F., Williams, B., Bryan, M., Conner, J., Siskey, M.R., Stockhausen, W.T., McDermott, S., Long, W.C., 2023. Subsampling catches to determine sex-specific length frequency in Alaska Fisheries Science Center bottom trawl surveys (NOAA Technical Memorandum No. NMFS-AFSC-464). Alaska Fisheries Science Center.

Ianelli, J.N., Kotwicki, S., Honkalehto, T., McCarthy, A., Stienessen, S., Holsman, K., Siddon, E., Fissel, B., 2018. Assessment of the walleye pollock stock in the Eastern Bering Sea (NPFMC Bering Sea and Aleutian Islands SAFE). North Pacific Fishery Management Council, Anchorage, AK.

Ives, A.R., 2022. Random errors are neither: On the interpretation of correlated data. Methods Ecol. Evol. 13, 2092–2105. https://doi.org/10.1111/2041-210X.13971

Kotwicki, S., Horne, J.K., Punt, A.E., Ianelli, J.N., 2015. Factors affecting the availability of walleye pollock to acoustic and bottom trawl survey gear. ICES J. Mar. Sci. J. Cons. 72, 1425–1439.

Kotwicki, S., Ono, K., 2019. The effect of random and density-dependent variation in sampling efficiency on variance of abundance estimates from fishery surveys. Fish Fish. 20, 760–774. https://doi.org/10.1111/faf.12375

Maunder, M.N., Piner, K.R., 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fish. Res., Data conflict and weighting, likelihood functions, and process error 192, 16–27. https://doi.org/10.1016/j.fishres.2016.04.022

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74. https://doi.org/10.1016/j.fishres.2012.07.025

McAllister, M.K., 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., Tromble, G.R., Lambert, D.M., Greene, K.E., 2014. Implementing a science-based system for preventing overfishing and guiding sustainable fisheries in the United States. ICES J. Mar. Sci. J. Cons. 71, 183–194. https://doi.org/10.1093/icesjms/fst119

Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.

Miller, T.J., Hare, J.A., Alade, L.A., 2016. A state-space approach to incorporating environmental effects on recruitment in an age-structured assessment model with an application to southern New England yellowtail flounder. Can. J. Fish. Aquat. Sci. 73, 1261–1270.

Miller, T.J., 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.

Monnahan, C.C., Dorn, M.W., Deary, A.L., Ferriss, B.E., Fissel, B.E., Honkalehto, T., Jones, D.T., Levine, M., Rogers, L., Shotwell, S.K., 2021a. Assessment of the Walleye Pollock Stock in the Gulf of Alaska.

Monnahan, C.C., Thorson, J.T., Kotwicki, S., Lauffenburger, N., Ianelli, J.N., Punt, A.E., 2021b. Incorporating vertical distribution in index standardization accounts for spatiotemporal availability to acoustic and bottom trawl gear for semi-pelagic species. ICES J. Mar. Sci. https://doi.org/10.1093/icesjms/fsab085

Nichol, D.G., Kotwicki, S., Wilderbuer, T.K., Lauth, R.R., Ianelli, J.N., 2019. Availability of yellowfin sole Limanda aspera to the eastern Bering Sea trawl survey and its effect on estimates of survey biomass. Fish. Res. 211, 319–330. https://doi.org/10.1016/j.fishres.2018.11.017

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. 158, 96–101.

O'Leary, C.A., DeFilippo, L.B., Thorson, J.T., Kotwicki, S., Hoff, G.R., Kulik, V.V., Ianelli, J.N., Punt, A.E., 2022. Understanding transboundary stocks' availability by combining multiple fisheries-independent surveys and oceanographic conditions in spatiotemporal models. ICES J. Mar. Sci. 79, 1063–1074. https://doi.org/10.1093/icesjms/fsac046

O'Leary, C.A., Thorson, J.T., Ianelli, J.N., Kotwicki, S., 2020. Adapting to climate-driven distribution shifts using model-based indices and age composition from multiple surveys in the walleye pollock (Gadus chalcogrammus) stock assessment. Fish. Oceanogr. 29, 541–557. https://doi.org/10.1111/fog.12494

Pennington, M., Godø, O.R., 1995. Measuring the effect of changes in catchability on the variance of marine survey abundance indices. Fish. Res. 23, 301–310. https://doi.org/10.1016/0165-7836(94)00345-W

Pennington, M., Volstad, J.H., 1994. Assessing the Effect of Intra-Haul Correlation and Variable Density on Estimates of Population Characteristics from Marine Surveys. Biometrics 50, 725–732. https://doi.org/10.2307/2532786

Punt, A.E., In press. Some insights into data weighting in integrated stock assessments. Fish. Res.

Rogers, L.A., Dougherty, A.B., 2019. Effects of climate and demography on reproductive phenology of a harvested marine fish population. Glob. Change Biol. 25, 708–720. https://doi.org/10.1111/gcb.14483

Schnute, J.T., Haigh, R., 2007. Compositional analysis of catch curve data, with an application to *Sebastes maliger*. ICES J. Mar. Sci. J. Cons. 64, 218–233.

Schnute, J.T., Hilborn, R., 1993. Analysis of Contradictory Data Sources in Fish Stock Assessment. Can. J. Fish. Aquat. Sci. 50, 1916–1923. https://doi.org/10.1139/f93-214

Stewart, I.J., 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.

Stewart, I.J., Monnahan, C.C., 2017. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. Fish. Res. 192, 126–134. https://doi.org/10.1016/j.fishres.2016.06.018

Stock, B.C., 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.

Stock, B.C., Xu, H., Miller, T.J., Thorson, J.T., Nye, J.A., 2021. Implementing two-dimensional autocorrelation in either survival or natural mortality improves a state-space assessment model for Southern New England-Mid Atlantic yellowtail flounder. Fish. Res. 237, 105873. https://doi.org/10.1016/j.fishres.2021.105873

Szuwalski, C.S., Ianelli, J.N., Punt, A.E., 2018. Reducing retrospective patterns in stock assessment and impacts on management performance. ICES J. Mar. Sci. 75, 596–609. https://doi.org/10.1093/icesjms/fsx159

Thorson, J.T., 2014. Standardizing compositional data for stock assessment. ICES J. Mar. Sci. J. Cons. 71, 1117–1128. https://doi.org/10.1093/icesjms/fst224

Thorson, J.T., Bryan, M.D., Hulson, P.-J.F., Xu, H., Punt, A.E., 2020. Simulation testing a new multi-stage process to measure the effect of increased sampling effort on effective sample size for age and length data. ICES J. Mar. Sci. 77, 1728–1737. https://doi.org/10.1093/icesjms/fsaa036

Thorson, J.T., Haltuch, M.A., 2018. 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. https://doi.org/10.1139/cjfas-2018-0015

Thorson, J.T., Johnson, K.F., Methot, R.D., Taylor, I.G., 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fish. Res. 192, 84–93. https://doi.org/10.1016/j.fishres.2016.06.005

Thorson, J.T., Miller, T.J., Stock, B.C., 2022. The multivariate-Tweedie: a self-weighting likelihood for age and length composition data arising from hierarchical sampling designs. ICES J. Mar. Sci. fsac159. https://doi.org/10.1093/icesjms/fsac159

Thorson, J.T., Minto, C., 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES J. Mar. Sci. J. Cons. 72, 1245–1256. https://doi.org/10.1093/icesjms/fsu213

Thorson, J.T., Ono, K., Munch, S.B., 2014. A Bayesian approach to identifying and compensating for model misspecification in population models. Ecology 95, 329–341. https://doi.org/10.1890/13-0187.1

Thorson, J.T., Shelton, A.O., Ward, E.J., Skaug, H.J., 2015. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. ICES J. Mar. Sci. J. Cons. 72, 1297–1310. https://doi.org/10.1093/icesjms/fsu243

Thorson, J.T., Stewart, I.J., Punt, A.E., 2011. Accounting for fish shoals in single-and multi-species survey data using mixture distribution models. Can. J. Fish. Aquat. Sci. 68, 1681–1693.

Trijoulet, V., Albertsen, C.M., Kristensen, K., Legault, C.M., Miller, T.J., Nielsen, A., 2023. Model validation for compositional data in stock assessment models: Calculating residuals with correct properties. Fish. Res. 257, 106487. https://doi.org/10.1016/j.fishres.2022.106487

Walters, C., Maguire, J.-J., 1996. Lessons for stock assessment from the northern cod collapse. Rev. Fish Biol. Fish. 6, 125–137.

Wang, S.-P., Maunder, M.N., 2017. Is down-weighting composition data adequate for dealing with model misspecification, or do we need to fix the model? Fish. Res., Data conflict and weighting, likelihood functions, and process error 192, 41–51. https://doi.org/10.1016/j.fishres.2016.12.005

Wilderbuer, T.K., Walters, G.E., Bakkala, R.G., 1992. Yellowfin sole, Pleuronectes asper, of the eastern Bering Sea: biological characteristics, history of exploitation, and management. Mar Fish Rev 54, 1–18.

Xu, H., Thorson, J.T., Methot, R.D., 2020. Comparing the performance of three data-weighting methods when allowing for time-varying selectivity. Can. J. Fish. Aquat. Sci. 77, 247–263. https://doi.org/10.1139/cjfas-2019-0107

Xu, H., Thorson, J.T., Methot, R.D., Taylor, I.G., 2019. A new semi-parametric method for autocorrelated age- and time-varying selectivity in age-structured assessment models. Can. J. Fish. Aquat. Sci. 76, 268–285. https://doi.org/10.1139/cjfas-2017-0446

Table 1: Proposed decomposition of the mismatch between data and stock-assessment model predictions (i.e., “errors”). This involves a 2x2 factorial cross of two types of error (rows) and two stages of the stock-assessment process (columns), and each cell lists examples that would cause that type of error (see Sections 3.1 through 3.4 for details).

		Stage of stock-assessment process	
		1: Field sampling and pre-processing data products	2: Stock assessment modelling and interpretation
Type of error	A: Imprecision (decreases with more data within a given year)	**1A: Sampling imprecision** (V_{sample}) • Finite survey sample sizes • Intra-haul correlations and inter-haul variation	**2A: Model imprecision** (V_{model}) • Process errors representing interannual variation in growth, mortality, or migration (i.e., semi-parametric model mis-specification)
	B/C: Bias / Inconsistency (does not decrease with new data)	**1B/C: Sampling bias** (B_{sample}) • Mis-specified survey design Distribution shifts (horizontal, vertical, among habitats)	**2B/C: Model bias** (B_{model}) • Ignoring migration, environmentally driven survival, and fishery targeting (i.e., parametric model mis-specification)

Table 2: Implications of the proposed decomposition of errors (see Table 1 for details), listing the implication, manuscript section with further discussion, and a published example for each

Implication	Manuscript section	Published example
Input sample size n_{input} measures "sampling imprecision", so further downweighting $n_{effective}/n_{input}$ measures the total resulting from sampling bias, model bias, and model imprecision	4.3	(Thorson and Haltuch, 2018)
Model-based expansion of sampling data can transform "sampling bias" into "sampling imprecision"	3.6	(O'Leary et al., 2020)
Auxiliary data can provide a lower bound on "sampling bias"	4.2	(Monnahan et al., 2021b)
Adding additional random effects (i.e., for time-varying processes) can transform "model bias" into "model imprecision"	4.4	(Stock et al., 2021)
Model-based downweighting of data is useful either: 1. for unimportant fleets, where unexplained model bias likely has little effect; or 2. when fitting to data when the n_{input} is not measured, and hence no starting point is available without model-based weighting; or 3. for fleets where biased fit to age/length composition will not also translate to bias for fit to indices or removals.	4.4	(Wang and Maunder, 2017)

Table 3: Recommendations resulting from this summary of data expansion and error decomposition

Recommendation
We recommend that assessment models include options to specify a vector for abundance indices or removals across years, and a separate matrix for proportions-at-age across years, rather than fitting to a combination of these two. This ensures that a small variance in measurements of total removals or total abundance is appropriately propagated even when proportions are less precise
We recommend using design-, model-, or bootstrap estimators to identify the variance of all data inputs, as well as auxiliary information where available to identify the variance arising from errors in the sampling frame;
We recommend providing the variance of each data input (including measured imprecision and the magnitude of survey mis-specification measured using auxiliary data) to the stock assessment model, so that the model will not estimate a variance for random effects that results in a tighter fit to each datum than is warranted by its specified variance. This then ensures that the variance of data inputs serves as an "upper bound" on the variance of estimated random effects.
If analysts choose not use the estimated sampling variance V^_survey within the stock assessment, we recommend as practical alternative that they replacing this with a single scalar quantity, "input sample size", representing the idealized multinomial sampling size with approximately similar variance. Adding additional random effects (i.e., model imprecision) will then result in smaller model residuals, and an "effective sample size" that approaches this input sample size (i.e., excess variance approaching zero).

Similarly, the "input sample size" provides an implicit upper bound on the variance of random effects.

We recommend that analysts use OSA instead of Pearson residuals, to account for the action of any random effects and also any non-normal error distributions. We similarly recommend that these residuals be visualized, where patterns among ages and years can be used to diagnose model-specification.

We recommend that model weighting be considered only as a first-pass response to overdispersion, and that assessment scientists instead seek to attribute residual patterns to additional model processes for important fleets. This is necessary to ensure that overdispersion and correlations among ages and years are interpreted not just for fitting age/length compositions, but also when (1) fitting to abundance indices and removals or (2) calculating reference points from that same fleet.

We recommend research to identify auxiliary data (whether combining habitat information, multiple surveys, or process research) that can be used to decrease sampling imprecision and inconsistency, and thereby mitigate the errors that are otherwise combined in "assessment model imprecision" that drive the downweighting of composition data. This research will typically occur in parallel to an operational assessment, and in some cases can be done by survey teams and reviewed during Methods Reviews with associated terms of reference in a given management region.

We recommend that data weighting be interpreted as a data-driven hypothesis about the sources of error, including model and survey imprecision and inconsistency, and ideally that the sensitivity to these choices be presented to highlight remaining uncertainties about errors.

Table 4 – Summary of different distributions (including alternative parameterizations where they exist) used to fit to compositional data (i.e., proportions at age, length, sex, and stage), including an early citation for each method, whether estimation occurs jointly with other parameters ("Likelihood") or requires a post-hoc tuning as a second stage of estimation ("2-stage") and also noting that the multinomial and Dirichlet-multinomial do not integrate to one across the vector of proportions and hence model selection cannot be used to compare fit between proper and improper likelihoods, whether the distribution can be fitted to proportions that include zeros, and whether the distribution uses information about an input sample size to evaluate subsequent data-weighting.

Method name	Estimation (2-stage or likelihood)	Permits zeros (Yes or No)	Uses input sample size (Yes or no)
Multinomial	Likelihood (improper)	Yes	Yes
Dirichlet	Likelihood	No	No
Dirichlet-multinomial	Likelihood (improper)	Yes	Yes
A. Saturating (Candy, 2008)			
B. Linear (Thorson et al., 2017)			
McAllister-Ianelli (1997)	2-stage	Yes	Yes
Francis (2011)	2-stage	Yes	Yes
Logistic normal:	Likelihood	No	No

A. Additive (Schnute and Haigh, 2007)			
B. Multiplicative (Cadigan, 2016)			
Multivariate Tweedie (Thorson et al., 2022)	Likelihood	Yes	Yes