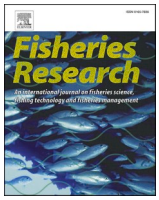
Fisheries Research 274 (2024) 107000

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



# Review of methodologies for detecting an observer effect in commercial fisheries data

Debra Duarte [a,b,*], Steven X. Cadrin [a]

[a] Department of Fisheries Oceanography, School for Marine Science and Technology, University of Massachusetts Dartmouth, 836 South Rodney French Blvd, New Bedford, MA 02744-1221,USA
[b] Northeast Fisheries Science Center, Fisheries Monitoring and Research Division, National Oceanic and Atmospheric Administration, 166 Water Street, Woods Hole, MA 02543, USA

ARTICLE INFO

Handled by A.E. Punt

*Keywords:*
Fisheries monitoring
Observer effect
Deployment effect
Bias detection
Sampling error

ABSTRACT

Observers are deployed on commercial fishing trips to collect representative samples of discard rates. However, fishers may change their fishing habits when an observer is onboard ("observer effect") or observer programs may over- or under-sample portions of the fleet ("deployment effect"). If the extent of these effects are substantial, observer data will not be representative of unobserved trips, potentially biasing the estimation of discards. This sampling bias can impact catch monitoring, stock assessments, and fishery management. The purpose of this study was to examine the power and error rate of several published methods for detecting an observer effect using a simulation of observer and deployment effects at varying sampling ratios (i.e., observer coverage) for several sample statistics. The simplest methods (t-test and F-test for difference of means and variances) provided an accurate although imprecise estimate of the observer effect size, but only when there were no deployment effects. A generalized linear mixed effects model (GLMM) was also not reliable for detecting small bias, but was not confounded by deployment effects and was relatively robust to changing coverage rates. The most complicated tests involved comparing differences in trip characteristics between subsequent trips for observed-unobserved and unobserved-unobserved pairs. These tests were able to detect smaller observer effects and were not confounded by deployment effects, but were unreliable at high coverage rates (>60%), producing both high false positive and false negative rates. Sensitivity tests also showed differing detection accuracy as the distribution of the metric of interest changed. Thus, the optimal test for detecting an observer effect will depend on the metric of interest, the coverage rate, and whether a deployment effect exists. An example from the New England groundfish fishery is provided to illustrate how conflicting results may be explained. Results should always be considered carefully when declaring that an observer effect is or is not occurring because of the sensitivity of the tests.

## 1. Introduction

Monitoring fisheries at sea helps to inform fisheries management, but representative observer coverage is expensive (Davies and Reynolds, 2002; Suuronen and Gilman, 2020). Sampling strategies are designed to meet statistical requirements but assume that samples of observed fishing trips represent those discard rates of unobserved trips (Hall, 1999). Non-random selection of trips to observe can result in "deployment effects", and alterations in fishing practices once the observer is onboard can cause "observer effects".

In fisheries with less than 100% observer coverage, total discards are typically calculated by expanding observed discards using a ratio estimator (Cochran, 1977). This can be based on effort (observed discards / observed effort x total effort) or landings (observed discards / observed total kept x total landings), both of which require a measure of total effort or landings, possibly from self-reported logbook data, dealer

*Abbreviations:* AIC, Akaike Information Criterion; ANOVA, Analysis of Variance; CV, Coefficient of variation; GARFO, Greater Atlantic Region Fisheries Office; GLMM, Generalized linear mixed effect model; K-S, Kolmogorov-Smirnov; NEFMC, New England Fishery Management Council; NEFSC, Northeast Fisheries Science Center; SBRM, Standardized Bycatch Reporting Methodology; VTR, Vessel Trip Report.

* Corresponding author at: Northeast Fisheries Science Center, Fisheries Monitoring and Research Division, National Oceanic and Atmospheric Administration, 166 Water Street, Woods Hole, MA 02543, USA.
*E-mail addresses:* debra.duarte@noaa.gov (D. Duarte), scadrin@umassd.edu (S.X. Cadrin).

https://doi.org/10.1016/j.fishres.2024.107000
Received 1 January 2024; Received in revised form 13 March 2024; Accepted 13 March 2024

reports, and/or vessel monitoring systems. Sampling designs are often stratified so that expansions are by strata such as gear type and region (e.g., Rago et al., 2005). Confidence intervals around these estimates indicate the variability in the observed trips within each stratum and are a function of the sampling ratio (i.e., observer coverage). Non-representative sampling may lead to biased estimates, causing errors in total catch estimates, inaccurate stock assessments, and non-optimal target reference points (Rudd and Branch, 2017).

The observer effect occurs when the presence of an observer causes the fisher to behave differently than they would on an unobserved trip. For example, a captain may fish an area of lower bycatch with an observer or use a different type or size of gear. Observer effects may also be unintentional. For example, if the vessel experiences higher operating costs while carrying an observer which results in the trip being shorter than it would have been, discard rate estimates may be biased (Faunce and Barbeaux, 2011). If an observer effect is large enough, it can violate the assumption that samples from observed trips are representative of the fleet as a whole and bias the extrapolations of observer discard rates to unobserved trips (Babcock et al., 2003).

Several methods have been developed for detecting observer effects. Rago et al. (2005) compared observed and unobserved trips for average landings, trip duration, and spatial distribution of fishing effort using paired t-tests of averages and standard deviations by stratum. They described two types of bias: non-representative sampling (i.e., the observer and deployment effects), and the "statistical properties of the consistency of the estimators" (i.e., the inherent bias of ratio estimators, on the order of $1/n$ (Cochran, 1977)) and concluded there was no evidence that increasing the sampling percentage would reduce the observer effect. Faunce and Barbeaux (2011) proposed a more complex method for detecting observer effects, using generalized linear mixed effects models (GLMMs), and accounting for factors such as vessel size, gear type, and season. Vessel identity was used as a random effect to account for the differences in mean landings. If removal of the observer term (a nominal binary variable) resulted in a higher AIC (Akaike Information Criterion) value, it was considered strong evidence that an observer effect had occurred. The method developed by Benoît and Allard (2009) compared pairs of sequential fishing trips by the same vessel in the same fishery. If the difference in the metric of interest (e.g., total landings) between observed trips and their sequential unobserved trips was greater than the difference between two consecutive unobserved trips, then an observer effect was deemed to be occurring.

The multispecies groundfish fishery in New England has been primarily managed by an output-control system that allocates tradeable quota to self-organized fishing cooperatives ("sectors") since 2010. Observed discard rates are used to monitor quotas as well as for stock assessments. This system may create an economic incentive for fishers to behave differently on observed trips (Demarest, 2019). The amount of observer coverage needed is based on achieving lower than a 30% coefficient of variation (CV) for discard estimates for each of 24 managed stocks (GARFO Greater Atlantic Regional Fisheries Office, 2021). Realized observer coverage rates ranged from 14% to 32% between 2010 and 2019. The fishery management plan was amended in 2021 to improve the reliability and accountability of catch reporting by increasing the target monitoring rate (combination of human observers and electronic monitoring) to 100% (NEFMC New England Fishery Management Council, 2021).

### 1.1. Objectives

The goal of this study is to test the efficacy of published methods (Rago et al., 2005, Faunce and Barbeaux, 2011, and Benoît and Allard, 2009) for detecting an observer effect, using data from the Northeast groundfish fishery as a case study. For each method, the following questions are asked:

1. In the absence of any effects (e.g., randomly selected trips with no systematic differences between observed and unobserved trips), how often would we expect to see false positive results indicating an observer effect?
   a. If there are false positives, what factors may be influencing them (e.g., observer coverage rate, within-trip variability)?
2. In the presence of an observer effect (i.e., systematic differences introduced on observed trips), how often would we expect to see true positive results?
   b. If there are false negatives, what factors may be influencing them (e.g., observer coverage rate, within-trip variability)?
   c. What is the minimum detectable effect size?
3. How confident should we be in positive or negative results from these tests?

## 2. Methods

To accomplish the stated objectives, we:

1. Developed simulation methods, including introducing a desired amount of bias on observed trips at coverage rates from 5% to 85%;
2. Applied five published tests from three studies (Benoît and Allard, 2009; Faunce and Barbeaux, 2011; Rago et al., 2005) on those datasets;
3. Quantified the rates of true and false positives and negatives of bias detection; and
4. Replicated these steps to achieve a reliable distribution of results.

All analyses were done using R (R Core Team, 2022).

### 2.1. Data selection for conditioning simulations

Trips taken by commercial fishing vessels were selected from the Northeast Fisheries Science Center (NEFSC) Vessel Trip Report (VTR) database from the New England large mesh otter trawl fleet from April 2018 through March 2019 (aligning with the dates used in the Standardized Bycatch Reporting Methodology [SBRM], Wigley and Tholke, 2020). Groundfish landings in this fishery were used as an example metric to test because it has been previously identified as one that may be subject to an observer effect (Demarest, 2019). For each vessel $v$ with at least 20 trips over the year and at least one record of kept groundfish ($G_v$), the proportion of trips with groundfish landings $p_v$ was calculated, along with the mean $\bar{x}_v$ and standard deviation $s_v$ when groundfish landings were non-zero. These were converted into the parameters for a delta-lognormal distribution:

$$G_v \sim \text{Binomial}(p_v) * \text{Lognormal}(\mu_v, \sigma_v^2)$$

with the probability of encounter $p_v$ for the binomial component, and $\mu_v$ and $\sigma_v$ for the lognormal component calculated as:

$$\mu_v = \ln\left(\frac{\bar{x}_v}{\sqrt{1+\frac{s_v^2}{\bar{x}_v^2}}}\right)$$

$$\sigma_v = \sqrt{\ln\left(1+\frac{s_v^2}{\bar{x}_v^2}\right)}$$

To mimic these patterns for simulations, four distributions and their controlling hyperparameters were defined:

Number of trips: $n_v \sim \text{NegBinomial}(r_n, p_n)$
Proportion of non-zero trips: $p_v \sim \text{Beta}(\alpha_p, \beta_p)$
Log mean of positive values: $\mu_v \sim \text{N}(\mu_m, \sigma_m^2)$
Log standard deviation of positive values: $\sigma_v \sim \text{N}(\mu_s, \sigma_s^2)$

One set of parameters ($n_v$, $p_v$, $\mu_v$, and $\sigma_v$) was drawn for each vessel, and $n_v$ trips were drawn from that vessel's unique distribution set $G_v \sim \text{Binomial}(p_v) * \text{Lognormal}(\mu_v, \sigma_v^2)$ (the number of vessels was fixed at 100 for all exercises, roughly based on the 113 vessels in the year and fleet used for conditioning). From there, a percentage of trips were marked as "observed" ($G_O$) and the remainder as "unobserved" ($G_U$). Unless otherwise specified, this percentage was the same for all vessels within a simulation (i.e., no deployment effects or non-random selection of trips).

The hyperparameters ($r_n$, $p_n$, $\alpha_p$, $\beta_p$, $\mu_m$, $\sigma_m$, $\mu_s$, and $\sigma_s$) control the variability within and between vessels. For example, $\mu_s$ controls the mean of the within-vessel deviance (do all vessels get assigned a larger or smaller standard deviation term) while $\sigma_s$ controls some of the between-vessel variance (do different vessels get assigned more similar or more disparate standard deviation terms). By manipulating the hyperparameters, many different data sets were created. The hyperparameters for simulated groundfish landings (default) and sensitivity test ranges are listed in Table 1.

To replicate an observer effect, a bias term was added that would adjust the generated landings only on observed trips. For a given amount of bias $b$, expressed as a proportion of the original data value, all $G_O$ values were replaced by $G_O * (1 + b)$. Thus, a $b$ of −0.2 represents a 20% decrease on observed trips, while a $b$ of 0.05 represents an increase of 5%.[1] This adjustment was made after the original data were generated and selected as observed or unobserved. In reality, some vessels would exhibit more or less of an effect than others, and the amount of bias would also vary between trips on the same vessel. Such variability would likely make the detection of an observer effect more difficult, so the detection rates presented here may be considered upper bounds. If positive and negative effects occur concurrently in the same fishery, they may mask any per-vessel differences (Demarest, 2019).

To replicate a deployment effect (non-random observer coverage), the selection of "observed" trips was weighted by average landings ($\mu_v$), such that the vessels with the highest $\mu_v$ would be selected at twice the target coverage rate, and the vessels with the lowest $\mu_v$ would be selected at near 0%. This was termed the "high catch preference" scenario. In the "low catch preference" scenario, the weightings were reversed so that the vessels with the highest average landings received the lowest coverage. These situations can occur if differential deployment is based on some factor that scales with catch. For example, the "high catch preference" scenario could represent observers being preferentially assigned to larger boats that conduct multi-day, offshore fishing trips, and the "low catch preference" could represent preference given to smaller, inshore, day boats.

**Table 1**
Parameter values for simulation testing.

| Parameter | Description | Value for simulated groundfish landings | Sensitivity value range |
|---|---|---|---|
| $r_n$ | First parameter for negative binomial distribution of number of trips; controls mean number of trips taken per vessel | 8 | Constant |
| $p_n$ | Second parameter for negative binomial distribution of number of trips; controls variation in number of trips taken per vessel | 0.15 | Constant |
| $\alpha_p$ | First parameter for beta distribution of proportion of non-zero trips; controls spread of probability by vessel | 1 | Constant |
| $\beta_p$ | Second parameter for beta distribution of proportion of non-zero trips; controls spread of probability by vessel | 0.1 | Constant |
| $\mu_m$ | Mean of the lognormal for vessel mean; central mean around which each vessel and trip varies | 7 | 0.5 – 10.5 |
| $\sigma_m$ | Standard deviation of the lognormal for vessel mean; how much each vessel's mean deviates from the central mean | 2 | 0.5 – 4.0 |
| $\mu_s$ | Mean of the within-vessel lognormal variance; how much each trip deviates from that vessel's mean | 0.5 | 0.1 – 2.0 |
| $\sigma_s$ | Variance of the within-vessel lognormal variance; how much each vessel's variance deviates from the overall variance | 0.2 | 0.1 – 0.51 |

## 2.2. Observer effect detection

Three published studies were selected for simulation testing that apply alternative approaches to observer effect detection (based on Rago et al., 2005; Faunce and Barbeaux, 2011; Benoît and Allard, 2009). Some alterations to the published methodologies were necessary because of differences in data structure and to facilitate replication. We grouped analyses based on the study they were drawn from, so that each group may contain more than one statistical test or output.

### 2.2.1. t-test and F-test

Because only one stratum was used in this exercise, differences in observed and unobserved trips within each simulation were tested with a two-sided Welch's t-test for difference of means (no assumption of equal variances) and an F-test for difference of variances (ANOVA F-statistic) to implement a similar method to that described by Rago et al. (2005). This is nearly identical to the methods used by Liggins et al. (1997). For both tests, a p-value lower than 0.05 was used as the definition of a "positive" result.

### 2.2.2. GLMM

The GLMM metric $G_i$ for trip $i$ and vessel $v$ is given by:

$$G_{i,v}^{\lambda} = \beta_0 + X_i\beta_1 + u_v + \varepsilon_{i,v}$$

where $\lambda$ is the Box-Cox power transformation parameter (here always the log transform $\lambda$=0), $\beta_0$ is the model intercept, $X_i$ is a nominal variable that is either 0 (unobserved) or 1 (observed), $\beta_1$ is the fixed effect coefficient for the observer effect, $u_v$ is the random effect for vessel ($u_v \sim N(0, \sigma_u^2)$), and $\varepsilon_{i,v}$ is the residual deviation ($\varepsilon_{i,v} \sim N(0, \sigma^2)$) to implement the method developed by Faunce and Barbeaux (2011). Because a number of records had zero landings, those had to be adjusted by adding a small amount of landings (chosen from a uniform distribution between 0.00098 and 0.00102 pounds) to avoid errors in the

[1] Other studies present alternate numeric definitions of bias. For example, Kerr et al. (2020) define bias in terms of the "missing" catch, so a bias of 100% indicates that the true catch was twice as high as reported. Using our notation, that can be expressed as the "true" $G_O$ being replaced with $G_O/(1 + b_K)$. Thus the bias levels used in that paper (0, 50, 125, and 200%) correspond to $b$ values of 0, −0.33, −0.56, and −0.67, respectively.

log-transformation. Vessels with fewer than two observed and two unobserved trips were removed to minimize the potential influence of rare outliers. The model was first fit using all terms, and the AIC value was calculated. Then the model was refit with the observer effect term removed (i.e., an intercept-only model with random effect of vessel) and the new AIC calculated. If the difference between these was more than 2, it was taken as a "positive" result for an observer effect, following the threshold used by Faunce and Barbeaux (2011), by which an observer effect was rejected if the model with observer factor was worse (higher AIC) or within the "substantial support" range (within 2 AIC units) of the model without the observer factor.

#### 2.2.3. Triplet method

Sequences of trips within vessels were identified which consisted of either three sequential unobserved trips (U-U-U) or an observed trip between two unobserved trips (U-O-U). Triplet sequences from actual landings data were restricted to less than a 45-day span between the first and last trip in a sequence. Simulated data had no date or inherent order, so they were sequenced in the order in which they were drawn. From each sequence, the middle trip was compared with either the first or the last trip (chosen randomly) to create a pair of unobserved trips (U-U) or an observed trip paired with an unobserved trip (O-U). Pairs were removed if the earlier trip was the same as the latter trip in the previous pair when ordered by date or draw order. Vessels with fewer than 3 sequences were removed from the analysis.

For each trip pair $j$ on vessel $v$, the difference between the $G$ values was calculated and standardized by the average value on unobserved trips for that vessel ($\overline{G_{U_v}}$). The percent change in an observed-unobserved (O-U) pair ($\Delta_{O_{j,v}}$) is given by

$$\Delta_{O_{j,v}} = \frac{(G_{O_{j,v}} - G_{U_{j,v}})}{\overline{G_{U_v}}}$$

Similarly, the percent change in an unobserved-unobserved (U-U) pair ($\Delta_{U_{j,v}}$) is given by

$$\Delta_{U_{j,v}} = \frac{(G_{U1_{j,v}} - G_{U2_{j,v}})}{\overline{G_{U_v}}}$$

In the absence of an observer effect, the distribution of $\Delta_O$ should be similar to the distribution of $\Delta_U$. This was tested using a two-tailed Kolmogorov-Smirnov (K-S) statistic. Benoît and Allard (2009) also used a Kuiper statistic to test for differences at the extremities. That test was omitted here as it restricted the replication of analyses. Benoît and Allard (2009) did not find any metrics that were indicated as significant using one test but not the other. A p-value of less than 0.05 was taken as a "positive" result.

The median difference was calculated as

$$(M_{\Delta_U - \Delta_O}) = \text{median}(\Delta_U) - \text{median}(\Delta_O)$$

This metric is for all vessels within the fleet. Analysis can also be done at the individual vessel, to determine if some vessels display more of an observer effect than others. Benoît and Allard (2009) pooled $\Delta$ values over fleets as well as vessels, whereas Demarest (2019) calculated median values at the vessel level. Because of the relatively low sample size within some vessels, we pooled over vessels (hence the lack of subscript $v$ in the above equation).

To test if medians were significantly different, this statistic was sampled with replacement from the $\Delta_U$ and $\Delta_O$ values 1000 times. If the bootstrap 95% confidence interval did not include zero, that indicated a statistically clear difference in the median values. This was taken as a "positive" result for an observer effect.

### 2.3. Evaluating methodologies

To evaluate the effect of coverage rate on false positives, 500 replicates were run for coverage rates varying from 5% to 85%, in 10% increments. The false positive rate (Type I error rate) was calculated as the percentage of replicates for that parameter set that returned a positive result from simulations with no observer effect added. To evaluate the proportion of true positives (i.e., power of the test), 500 replicates were run for the same coverage rates but with observer effects varying from −50 to 0% in 5% increments. To determine the minimum detectable effect size, the true positive rates (for non-zero bias) were fit with a logistic regression against the introduced observer effect size. The point at which the fitted line of frequency of detecting an effect crossed 75% was taken as the minimum detectable effect size (i.e., the point where the power was 0.75). All simulations were repeated with simulated high and low preference deployment effects.

Our preliminary investigation showed that detection rates were similar for positive and negative effects (i.e., symmetrical around 0% bias). We applied negative bias because it is the more common concern based on incentives (Henry et al., 2019). Under the presumption of symmetry, the results in this study for $b = -0.5$ (50% decrease on observed trips) should be roughly similar to $b = 0.5$ (50% increase on observed trips).

If the test allowed for it, the magnitude of the effect ($\widehat{b}$) was estimated in a way that could be comparable to the bias parameter $b$. For the t-test, bias was derived from the difference in group means:

$$\widehat{b}_t = \frac{\overline{G}_O}{\overline{G}_U} - 1$$

The 95% confidence interval around $\widehat{b}_t$ was defined as:

$$\frac{\overline{G}_O \pm 1.96 s_t}{\overline{G}_U} - 1$$

where $s_t$ is the estimated standard error of the mean difference between $\overline{G}_O$ and $\overline{G}_U$.

For the mixed effects model, the transformed parameter was used:

$$\widehat{b}_g = \exp(\beta_1) - 1$$

The 95% confidence interval was calculated as:

$$\exp(\beta_1 \pm 1.96 s_g) - 1$$

where $s_g$ is the estimated standard error of the of mean of the $\beta_1$ (observer effect factor) coefficient.

In the median differences test, if the median $\Delta_U \approx 0$, then $(M_{\Delta_U - \Delta_O})$ simplifies to the median of $0 - \Delta_O$, which should be equivalent to median $-(G_O/G_U - 1)$, thus:

$$\widehat{b}_m = -(M_{\Delta_U - \Delta_O})$$

The 95% confidence interval was taken from the bootstrap as described earlier.

For evaluating estimated effect sizes, we categorized the results as "accurate" if the 95% confidence interval around the estimate included the true observer effect added, and "precise" if the width of the interval was less than 20% (i.e., a margin of error of +/- 10% in terms of observer effect magnitude).

To evaluate the sensitivity of the tests, the hyperparameters were adjusted individually (ranges listed in Table 1), with all other parameters held at their original value. Coverage rate was tested at 5, 25, 65, and 85%, and bias was tested at −50, −30, −10, and 0% (no observer effect). No deployment effects were added to sensitivity tests. Each combination was simulated 1000 times. Trends were identified using a logistic regression of test significance (i.e., 1 if the test was positive and 0 otherwise) against the value of the hyperparameter. This was evaluated independently for each coverage rate, test, and bias level. Results with p-values less than 0.0001 were considered significant trends. Note that here we use a smaller p-value due to the repeated tests to reduce Type I errors. This type of correction wasn't needed within the individual simulations, with the presumption that each iteration represents

a single test that a researcher would be conducting and thus using the p=0.05 standard.

Lastly, all tests were run on the true observed vs. unobserved groundfish landings for New England large mesh otter trawl in SBRM year 2018, to compare with the simulated distributions. A null distribution was created by randomly assigning observed or unobserved to each trip at the realized coverage rate (14.2%) and calculating the test statistics. This was replicated 500 times to create a dataset with no observer effect and no deployment effect. This was compared to the real results to determine what proportion of the null distribution was at least as large as the observed test statistic using real data.

## 3. Results

The term "true positive" refers to test results that are statistically significant when there is an observer effect, and "false positive" to significant test results when there is no observer effect. When no bias was simulated, any differences between observed and unobserved values are due to the stochasticity of the data generating process, just as there would be variability between trips in real life. In any random subset of trips, there may be differences between observed and unobserved trips that are not the result of a systematic observer effect, and it is these cases that we have tried to distinguish as "false positives".

### 3.1. Simulated groundfish landings

#### 3.1.1. Detection rates

With no observer effect and no deployment effect, the false positive results for simulated groundfish landings were low for most conditions (Fig. 1). The t-test tended to perform worse at lower coverage (8–13% false positive rate at 5% coverage) but otherwise stayed less than 10% false positives (2–9%, Fig. 1a). The median difference test was less reliable at higher coverage (11–18% false positive at 85% coverage), but otherwise had less than 5% false positives (0–4%, Fig. 1e). The other tests did not vary much with coverage rate, with the K-S test having a slightly lower false positive rate (1–7%, Fig. 1d) than the GLMM test (2–8%, Fig. 1c) and F-test (2–7%, Fig. 1b).

When deployment effects were added, the t-test and F-test returned at least 75% false positives for all coverage rates and both preference types (Figs. 1f, 1g, 1k, and 1l). The GLMM results were slightly higher than in the no preference scenario (2–11%, Figs. 1h and 1m), as was the K-S test (2–8% (Figs. 1i and 1n). The median difference test did not show the high increase in false positives at high coverage as in the no deployment scenario; otherwise, it did not differ much by coverage or between scenarios (0–4%, Figs. 1j and 1o).

When no deployment effects were present, the true positive rates (power) were generally higher with increasing magnitude of observer effect added for all tests (Fig. 2). The t-test and F-test had slightly lower power at the smallest and largest coverage rates for a given amount of bias (Figs. 2a and 2b). Power was above 96% for both tests at coverage rates between 25% and 65% and the highest bias (-50%). At moderate bias (-30%), the power dropped to between 73% and 87% at medium coverage, and at low bias (-10%) the power was only 11–25%. The GLMM test results were similar (100% power for high bias, 86–91% for moderate bias, 11–17% for low bias; all lower at the highest and lowest coverage rates; Fig. 2c). The K-S test and median difference tests, on the other hand, were more strongly affected by coverage rate (Figs. 2d and 2e). At high and moderate bias, the power was above 96% when coverage rates were 65% or lower. As coverage increased, the power sharply decreased to less than 40% at the highest coverage (85%). At low bias, the highest power was 44–68% at 25% coverage, above and below which the power decreased.

When both observer and deployment effects were present, the GLMM test was the most consistent across all scenarios, with true positive rates being only slightly lower when deployment effects were introduced compared to the no preference scenario (particularly for the high catch preference scenario) and a more pronounced decrease at higher coverage (Figs. 2h and 2m). In the low catch preference scenario, the t-test and F-test returned positive results for over 94% of simulations regardless of bias or coverage (Figs. 2k and 2l). In the high preference scenario, at low coverage rates the tests became less powerful as bias increased; at 15% coverage the power of the t-test was 97–99% for low bias, 77–84% for moderate bias, and only 5–9% for high bias (Figs. 2f and 2g). The K-S test and median difference test showed similar patterns to the no preference scenario (high power at high and moderate bias, highest power for low bias occurred around 25% coverage) with the exception of the sharp decrease in power at the highest coverage rates (Figs. 2i, 2j, 2n, and 2o).

The minimum detectable effect size differed strongly by test, coverage rate, and presence of a deployment effect (Fig. 3). Overall the K-S test and median difference test were the most sensitive, reliably detecting observer effects as low as 12%. They both performed best when coverage rates were between 15% and 55%. Under the no deployment effect scenario, both tests were unable to reliably detect even large observer effects at high coverage rates (>75%, Fig. 3a), which did not occur when deployment effects were added (Figs. 3b and 3c; see Trip Filtering, below). The GLMM test was the next most reliable, detecting observer effects as small as 22% and performing best at coverage rates between 25% and 65%. At the lowest coverage rates, the GLMM test was unable to reliably detect observer effects smaller than 45%. It performed slightly worse with the high catch preference deployment effect scenario (Fig. 3b).

The t-test and F-test performed worst overall, only able to reliably detect effect sizes as low as 25% in the best scenario (25–75% coverage rate with no deployment effect, Fig. 3a). When deployment effects were added, a minimum detectable effect size could not be calculated (Figs. 3b and 3c). Under the high catch preference scenario, the rate of positive responses decreased with additional bias for a given coverage rate (i.e., opposite to the other tests), and under the low catch preference scenario nearly all results were positive, regardless of coverage or bias added (including no bias).

#### 3.1.2. Estimated effect sizes

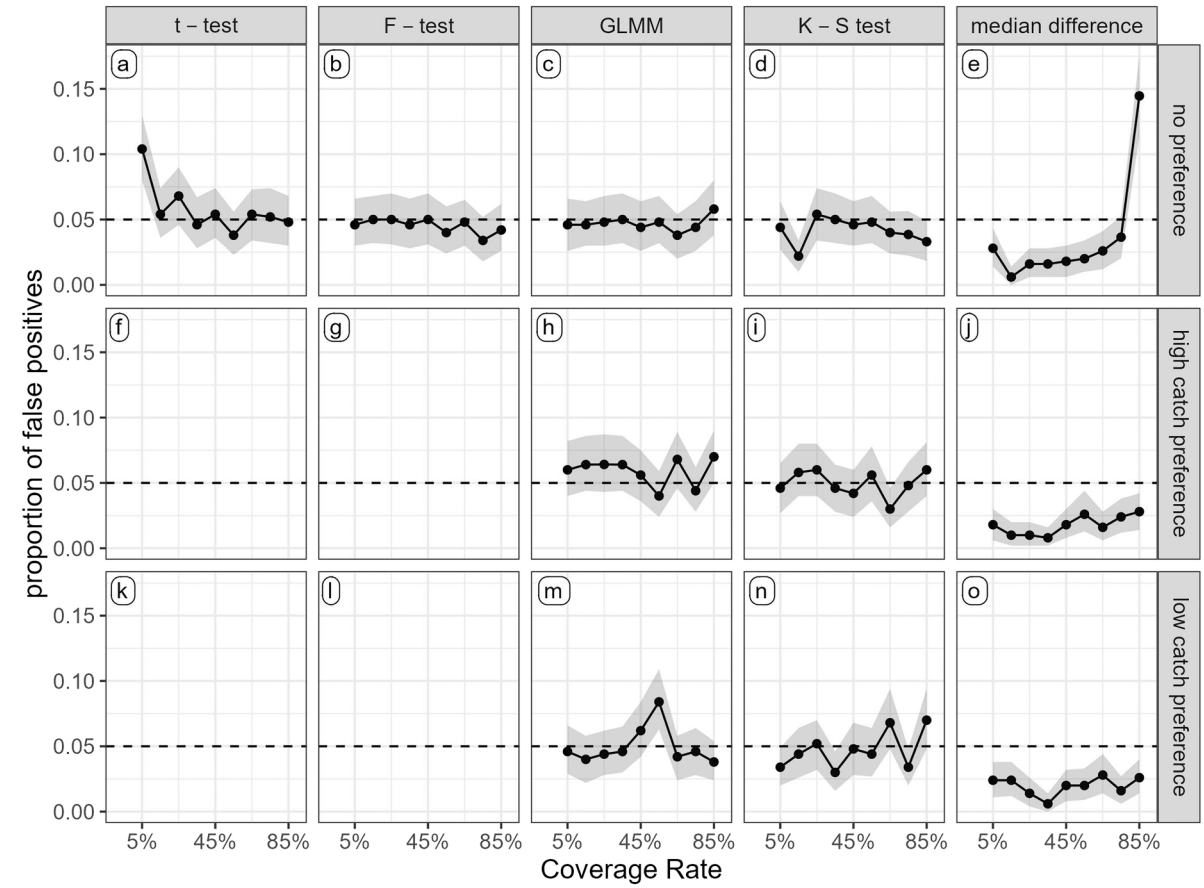
With no deployment effect, the $\widehat{b}_t$ estimates from the t-test were over 90% accurate (Fig. 4a) but less than 13% precise (Fig. 4b). Confidence intervals were generally centered on and symmetric around the true bias added (Figs. 5a and 5d). With high-preference deployment effects the estimated effect size was always more positive than the true bias (Fig. 5 g and 5j), and with low-preference deployment effects, the estimated effect size was always more negative than the true bias (Fig. 5 m and 5p), including when no observer effect was added.

The estimated effect size $\widehat{b}_g$ associated with the GLMM observed factor tended to be slightly more positive than the observer effect introduced (Figs. 5b, 5e, 5 h, 5k, 5 n, and 5q). This trend was similar for all coverage rates, with a much larger variability between iterations at low coverage. Accuracy was between 84% and 97% for all scenarios (Figs. 4c, 4i, and 4o), but precision was less than 10% for all but the high bias scenarios (Figs. 4d, 4j, and 4p).

The estimated effect size $\widehat{b}_m$ from the median difference generally had much narrower confidence intervals than $\widehat{b}_t$ or $\widehat{b}_g$, but consistently underestimated the magnitude of the actual observer effect (Fig. 5 f, 5 l, and 5r). Accuracy decreased noticeably with increasing observer effect magnitude, particularly at moderate coverage levels where precision was also highest (Figs. 4e, 4f, 4k, 4l, 4q, and 4r). Accuracy was highest at the largest and smallest coverage rates where precision was lowest due to the very wide confidence intervals.

#### 3.1.3. Trip filtering

The inconsistent pattern of performance on the K-S and median differences tests can be explained by the triplet selection process (Fig. 6). With no deployment effects, at 5% coverage a large proportion of the



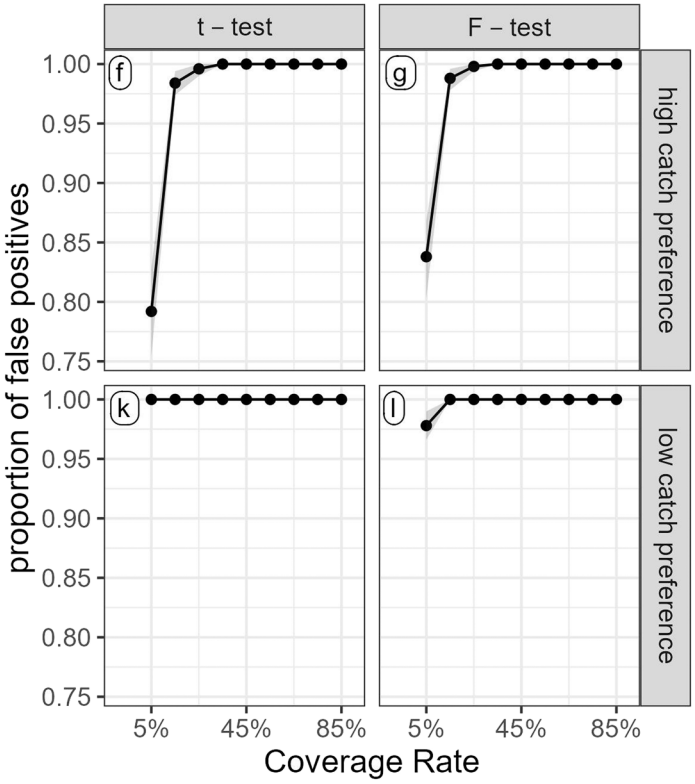




**Fig. 1.** Rate of false positives on simulated groundfish landings with no observer effect added. Panels f, g, k, and l shown on a different axis scale because they were far outside of the range of the other panels. Dashed line represents $p = 0.05$. Shaded areas represent central 95% intervals from 500 resamples.

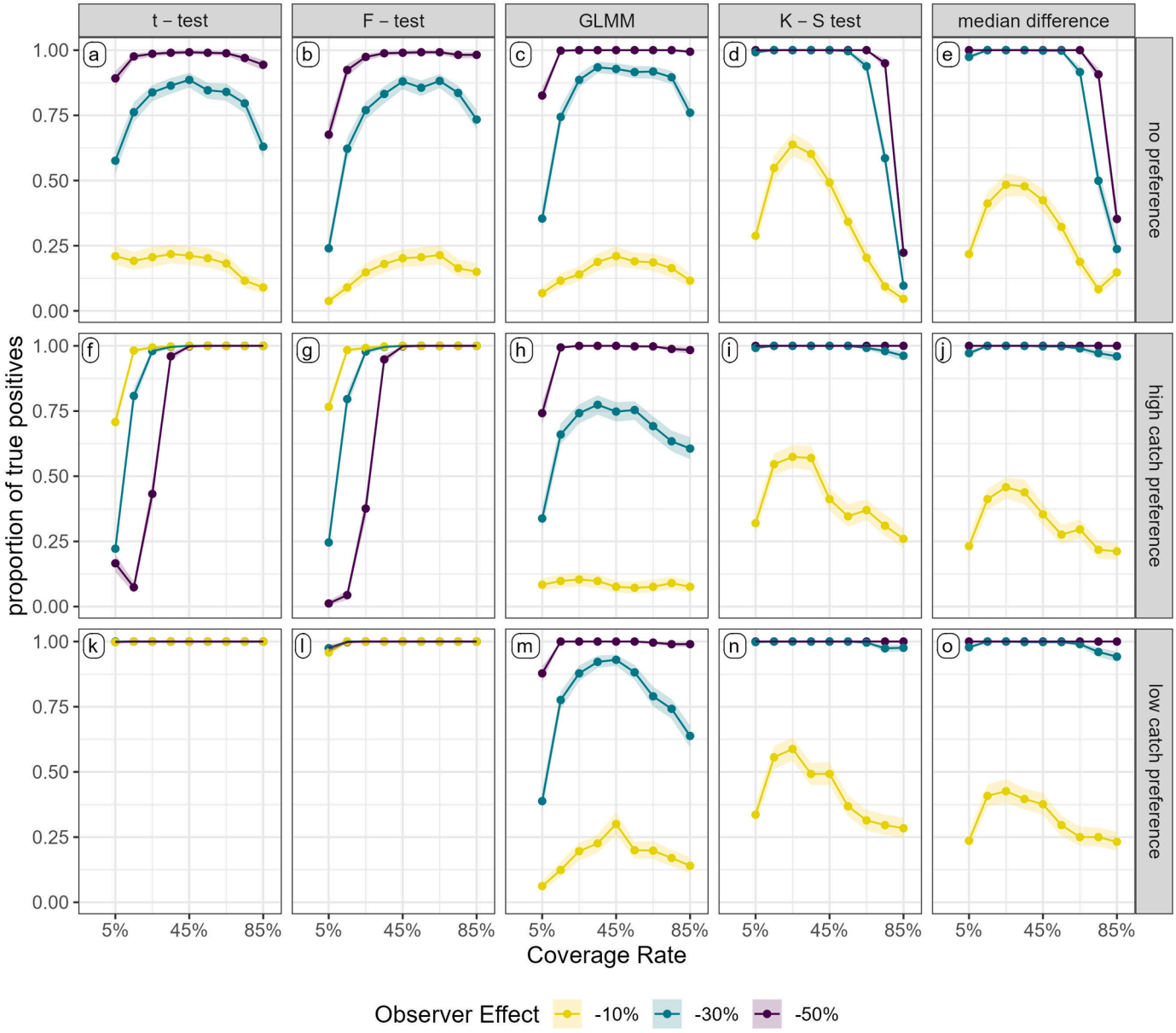



**Fig. 2.** Rate of true positives (power) on simulated groundfish landings with observer and deployment effects added. Shaded areas represent central 95% intervals from 500 resamples.

data were used in the final analysis (67% of observed trips, 54% of unobserved trips, and 99% of vessels, on average). As coverage rates increased, the probability of finding two consecutive observed trips increased, so the frequency of U-O-U triplets decreased, leading to a lower proportion of the observed trips being used in the final analysis (to less than 1% at 85% coverage), and higher coverage meant that the likelihood of finding three consecutive unobserved trips to create the U-U-U triplets also decreased, so that fewer than 8% of unobserved trips were used, and those were mostly in the O-U pairs (Fig. 6a). The number of vessels remained high so long as coverage was below 45%, but then dropped such that on average only 9% of the original vessels remained at 85% coverage (Fig. 6b).

More trips occurred on vessels with either high or low average landings when there was a deployment effect. At moderate coverage, those vessels were more likely to have 3 sequences, and so were more likely to be retained, while non-preferred vessels tended to drop out. At higher coverage rates the preferred vessels were more likely to have back-to-back observed trips, which would never be included in the triplet filtering algorithm. At 25% coverage, 99% of vessels were retained, the same as in the no preference scenario, then decreased as coverage exceeded 35%, but remained higher at the highest coverage (42% at 85% coverage, Fig. 6d and 6f). The percent of observed trips retained still dropped from 65% at the lowest coverage to 5% at the highest, while the percent of retained unobserved trips stayed near 50% at all coverage rates above 45% (Fig. 6c and 6e). This explains why the median difference and K-S tests showed low reliability at high coverage rates for no deployment effect but were seemingly "fixed" by including a deployment effect – there were still enough trips in the selected pool to estimate the simulated observer effect (which in this simulation was the same for all trips), but only for the portion of vessels that remained in the analysis.

### 3.2. *Sensitivity to parameter values*

False positive rates did not vary consistently with changes in parameter values (Fig. 7). The major exception was the increase in false

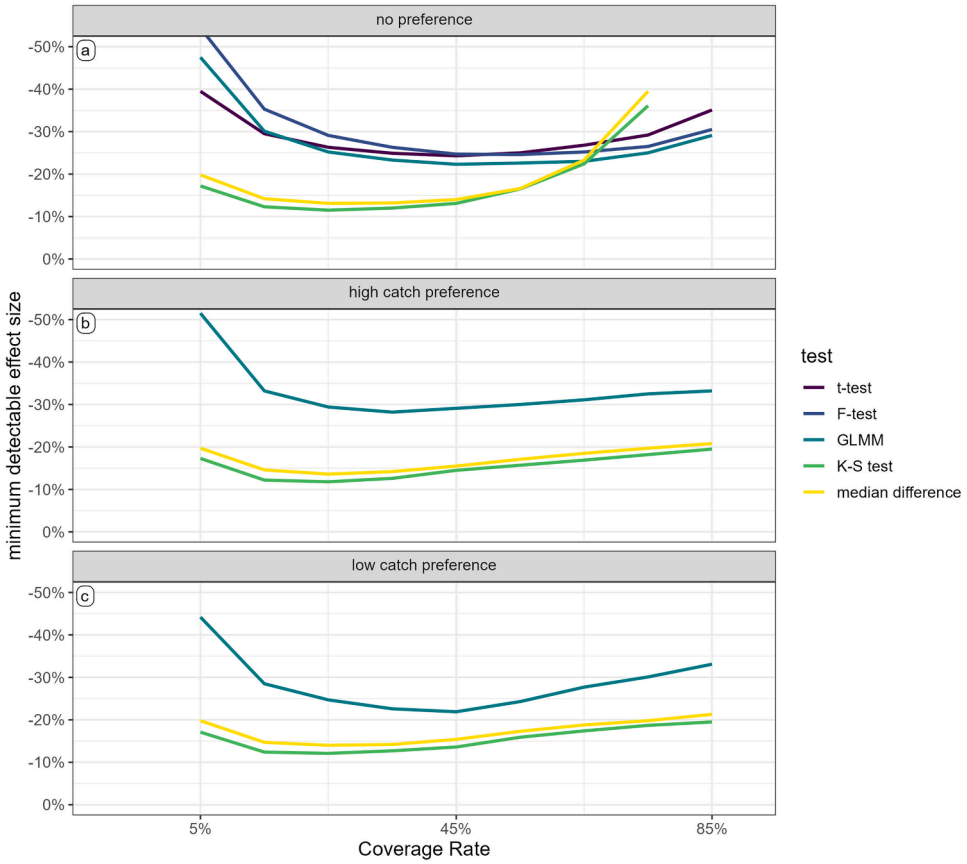



**Fig. 3.** Minimum detectable effect sizes for five tests on simulated groundfish landings by coverage rate. Minimum sizes could only be calculated for t-test and F-test under the no preference scenario.

positives from the t-test as $\mu_s$ (how much each trip deviates from that vessel's mean) increased at the highest and lowest coverage rates (Fig. 7c). There was also an increase in t-test false positives with increasing $\sigma_m$(how much each vessel's mean deviates from the central mean) at the lowest coverage rate (Fig. 7b), and a slight decrease in the false positive rate of the median difference test with larger $\sigma_s$ (how much each vessel's variance deviates from the overall variance, Fig. 7t).

The minimum detectable effect sizes responded to changes in the hyperparameters more clearly (Fig. 8). The t-test and F-test were less able to detect small effects as $\sigma_m$ and $\mu_s$ increased (Figs. 8b, 8c, 8f, and 8g) and to a lesser extent with increasing $\sigma_s$ (Figs. 8d and 8h). The GLMM test was relatively unaffected by all changes except for a decrease in power with increasing $\mu_m$ (the central mean around which each vessel and trip varies, Fig. 8i) and a slight decrease with increasing $\mu_s$ (Fig. 8k). The K-S test and median difference test were less likely to detect smaller effects with increasing $\mu_s$ (Figs. 8o and 8s) and somewhat more likely to detect smaller effects with increasing $\sigma_s$ (Figs. 8p and 8t). Minimum detectable effect sizes could not be determined for the K-S test and median difference test at the 85% coverage rate level because all levels of bias had power less than 0.75 due to trip filtering.

### 3.3. Application to New England groundfish landings data

When applied to the actual New England groundfish landings data, 4 of the 5 tests did not detect an observer effect, with the GLMM test suggesting a significant decrease of 38% on observed trips (AIC difference = 13.67, Table 2). The t-test indicated an increase of 15% landed groundfish weight on observed trips but was not significant (t-test p = 0.14, F-test p = 0.13). The K-S test was also not significant (p = 0.11) and the median difference was 0%. Because of the random selection of the first or last trip in a triplet, the K-S and median difference test results varied even though the observed trip assignments were unchanged: 42% of 500 iterations returned a significant result on the K-S test and 0.4% returned a significant result on the median difference test; the $\widehat{b}_m$ values also changed slightly by iteration. The other tests did not change when re-running analyses on the same data.

Compared to the null distribution, 13.0% of iterations had a t-statistic at least as far from 0 as the observed value (2-tailed). For the F-test, 13.2% of replicates had an F-statistic at least as large as the observed value. For the K-S test, 6.4% of replicates had a D-statistic at least as large as the observed value. These differed slightly from the p-values calculated for each test but are similarly non-significant. Only 1.0% of iterations under the null hypothesis had a median difference not equal to the observed value of 0. For the GLMM test, the observed statistic was larger than all simulated values, suggesting strongly significant results.

Estimated effect sizes differed by test. Under the null hypothesis, 63.2% of $\widehat{b}_t$ ranges contained the observed $\widehat{b}_t$ value; 2.2% of $\widehat{b}_g$ ranges contained the observed $\widehat{b}_g$ value; and 100% of $\widehat{b}_m$ ranges contained the observed $\widehat{b}_m$ value. This is consistent with the tests for significance.

## 4. Discussion

Through simulation, we demonstrated that the existing methods for detecting an observer effect can be reliable under certain conditions. An ideal test would have a low false positive rate, have a high true positive rate (power), provide a precise and unbiased estimate of the effect size, and be reliable across various underlying distributions. No single test reviewed here fulfilled all these criteria.

All statistical tests are predicated on some underlying assumptions about the data distribution and sampling mode. The underlying data structure used here should have invalidated several of the test

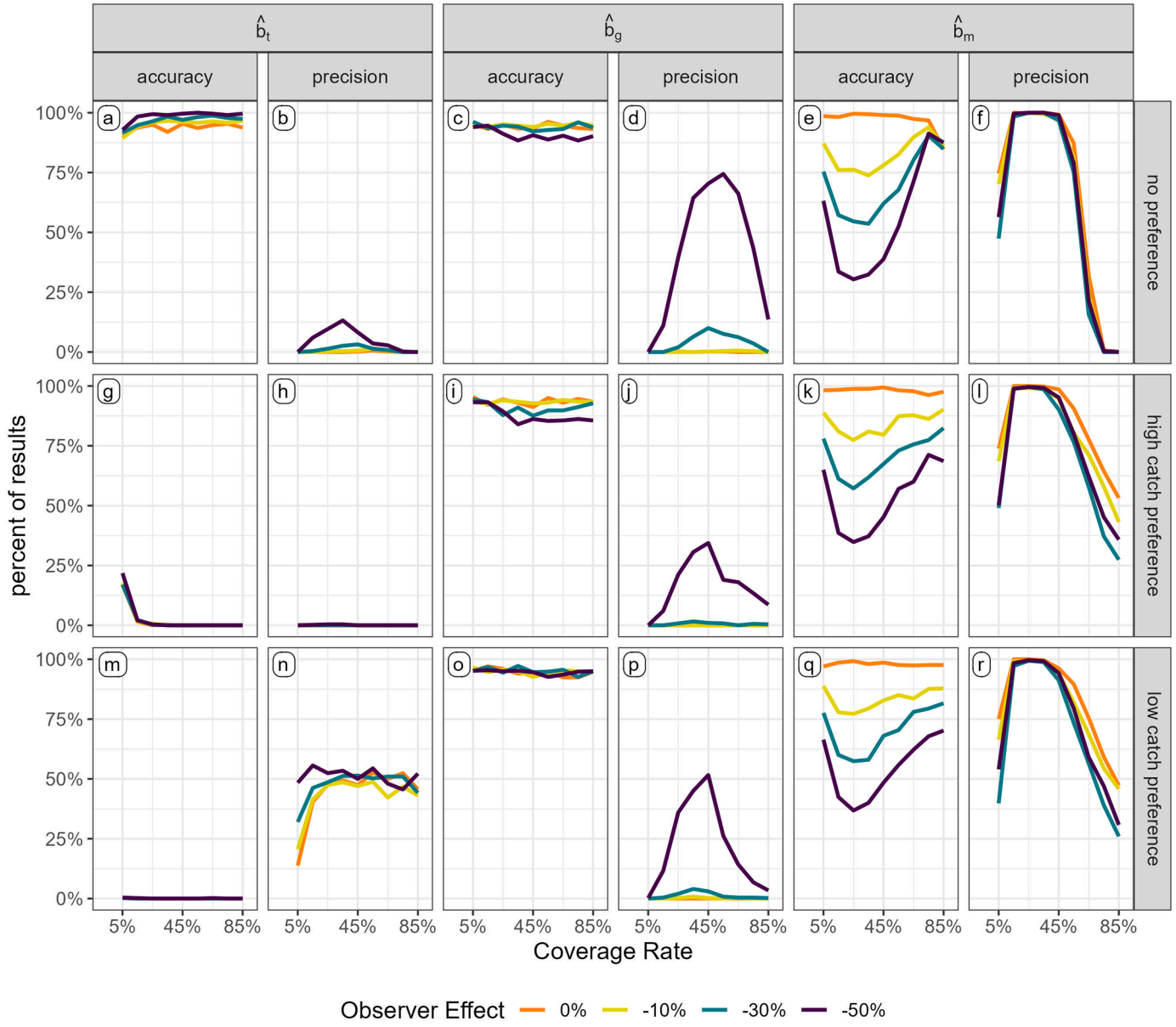



**Fig. 4.** Accuracy and precision of the estimated observer effect magnitude from the t-test ($\hat{b}_t$), GLMM ($\hat{b}_g$), and median difference ($\hat{b}_m$) in simulated groundfish landings. Accuracy is the percent of iterations for which the 95% confidence interval contained the true level of observer effect added in the simulation, and precision is the percent of iterations for which the interval width is less than 20%.

assumptions, such as the assumption of normality for t-tests. The large sample sizes may have improved the outcome because of the Central Limit Theorem, but this might not hold for a smaller dataset, such as a fishery with few overall trips and/or low coverage rates. Some fleets in the Northeast region require less than 1% of trips to be covered per the SBRM recommendations (Wigley and Tholke, 2020). For these fisheries, it would be difficult to interpret the results of observer effect detection, as both positive and negative results may be unreliable.

Simulations always require some simplification of real-world dynamics. In this study, observer effects were intentionally consistent across all observed trips (i.e., always a fixed percent reduction from the original value drawn) so that results could been compared to a single true value. In application to fishery monitoring data, we would not expect such consistency. Some vessels may alter their behavior every time, some may change depending on various factors (familiarity with the individual observer, amount of quota remaining, prior intentions for the trip, etc.), and some vessels may never change their behavior. Modeling human behavior is difficult, but we may be able to identify expected reactions to certain circumstances. For example, Demarest (2019) suggests that the shift in management measures from an input-control (fishing effort) to an output-control (removals) regime in the groundfish fisheries created higher economic incentives for observer effects because discards now represent a direct cost to the industry. Extension of this work could include additional (but plausible) individual heterogeneity to understand how this would affect the performance of tests.

### 4.1. *t-test and F-test*

The t-test and F-test were unable to distinguish deployment and observer effects. These methods could be used for answering overall questions such as "are observed trips representative of unobserved trips?" but should not be used to evaluate observer effects specifically (as opposed to any other type of systematic bias) unless there is

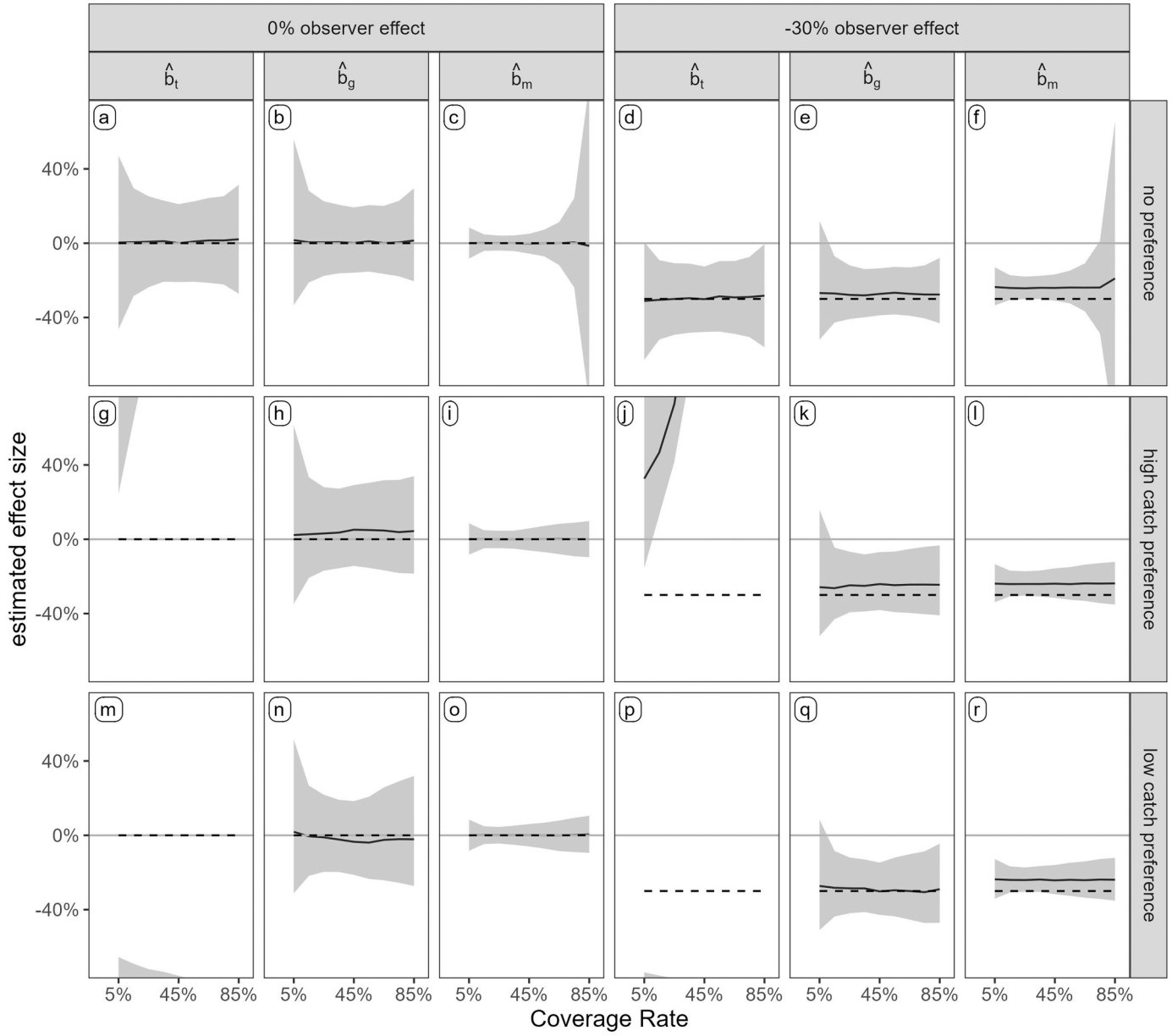



**Fig. 5.** 95% confidence intervals for the estimated effect size of the bias factor from the t-test ($\widehat{b}_t$), GLMM ($\widehat{b}_g$), and median difference ($\widehat{b}_m$) for groundfish landings from simulated data. The solid horizontal grey line represents 0 (no observer effect), the thin dotted line is the level of observer effect added to the simulations, the thin black line is the mean estimated effect size from 500 simulations, and the grey ribbon is the mean of the 500 95% confidence intervals around the estimated effect sizes.

sufficient evidence that other sources of non-randomness are not occurring. In the absence of deployment effects, both tests tended to perform best at moderate to high coverage rates (15–75%). There was an increase in false positive rates at lower coverage and an increase in false negatives at higher coverage, likely due to the low number of either observed or unobserved trips leading to higher variance in that category. There is also the possibility of higher false positive and false negative rates with larger between-vessel variance. Thus, these tests may be preferred for metrics with smaller ranges, such as days absent, the number of areas fished, the number of market categories landed, etc. Because these tests use an assumption of normality, transforming or scaling the metric (e.g., centering at zero) before testing may make them more appropriate for use with larger metrics. Values could also be standardized to the mean for each vessel, or paired differences by vessels could be used. The data could also be scaled on other variables such as trip duration or vessel size, neither of which was simulated directly in this study.

The estimated effect size from the t-test tended to be accurate but with wide margins of error that reduce its utility in drawing clear conclusions. With no deployment effects, both tests were at best able to detect observer effect sizes as small as 25%. Therefore, the significant differences reported by Liggins et al. (1997) which used this method should be interpreted as a potential mix of deployment and observer effects. Due to low coverage rates, the non-significant results reported by Liggins et al. (1997) and Rago et al. (2005) should be interpreted either as an observer effect with low magnitude, or as a deployment effect masking an observer effect in the opposite direction (e.g., a preference towards vessels with higher landings coinciding with a reduction in landings on observed trips).

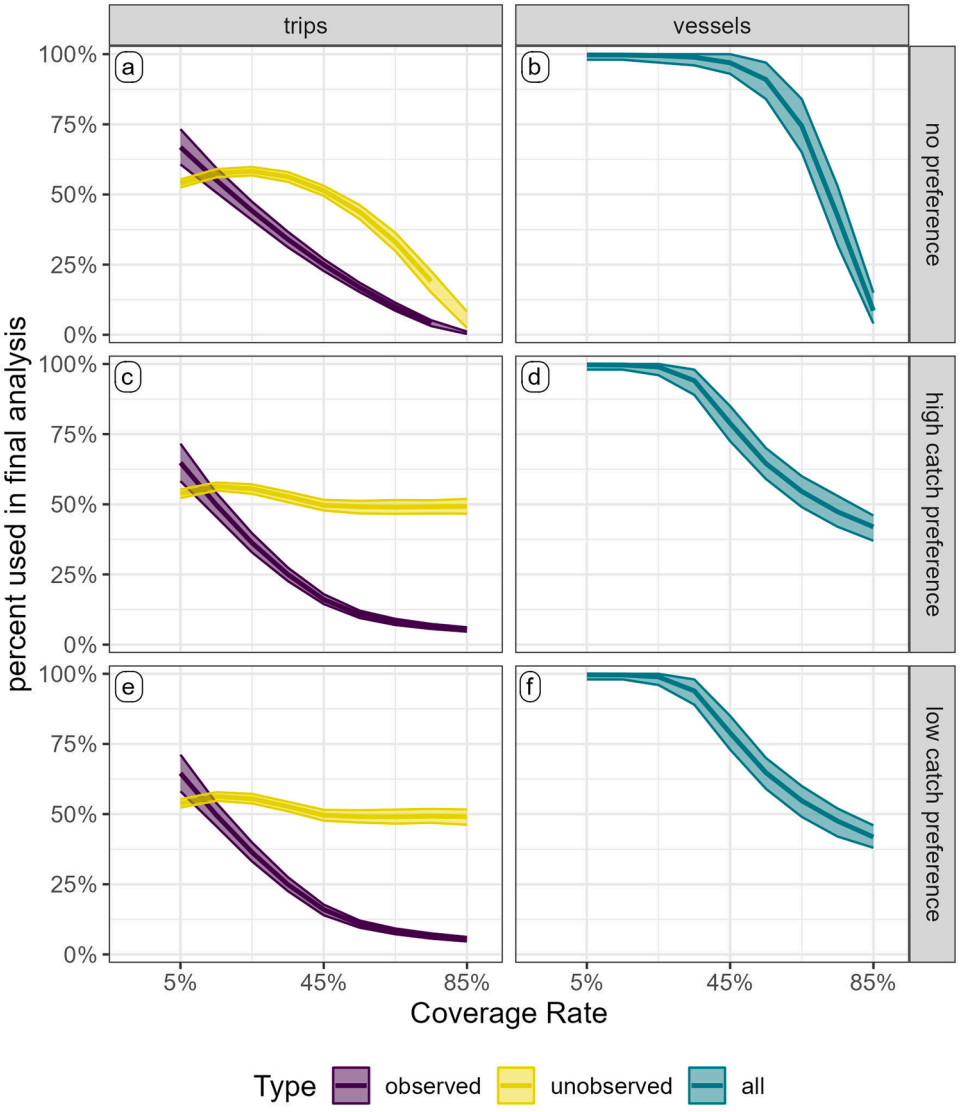



**Fig. 6.** Proportion of original trips and vessels that remained after triplet analysis filtering from simulated data, by coverage rate and deployment effect.

### 4.2. GLMM

The GLMM test performed most consistently over the range of our simulations, particularly in the moderate coverage range of 25–75%. False positive rates were above 5% in several scenarios but rarely above 10%. The GLMM test was generally insensitive to changes in hyper-parameters; there was some decrease in performance with the addition of deployment effects but the changes were less substantial than those for the other two methods. However, the GLMM was not able to detect smaller effect sizes, particularly at low coverage. Applying these conclusions to the Faunce and Barbeaux (2011) results, the two fisheries with significant GLMM results likely had observer effects of at least moderate magnitude, the three fisheries with non-significant results could have had observer effects of smaller magnitude, and the deployment effects identified in the same study likely did not affect these conclusions.

Estimated effect sizes with the GLMM test tended to be accurate but also suffered from wide margins of error, though not as large as the t-test estimate. Minimum detectable effect sizes were relatively consistent across all parameter changes, except for a slight increase with higher mean values. This method possibly suffered because it could not be optimized for any single dataset. If an investigator was performing this method on a single fishery, they perhaps could improve the estimates somewhat by manipulating details (such as the Box-Cox parameter) or including additional covariates that could not be done for thousands of replicates. One could also use a different value for the AIC difference threshold. Here we used 2 to follow the original study, which is based on Burnham and Anderson's (2004) rule of thumb; in essence, an observer effect is only declared when the model without the observer factor is outside the range of "substantial support" relative to the model with the observer factor. Using a threshold of 4 (not shown) resulted in the reduction of the false positive rate (to the lowest rate among these tests) but also a decrease in the ability to detect smaller effect sizes.

### 4.3. Triplet method

The K-S and median difference tests had high reliability at moderate coverage (10–40%) without deployment effects, but performance decreased as coverage increased. At high coverage (>60%) these methods should be avoided because of the high rates of both false positives and false negatives. Even in the ideal range, the median difference tended to underestimate the magnitude of the observer effect, with a misleadingly narrow confidence interval that failed to capture the true value more often than would be expected. These methods were able to detect smaller effect sizes than the other methods, and so could be useful if there are concerns that a small observer effect is having a large impact.

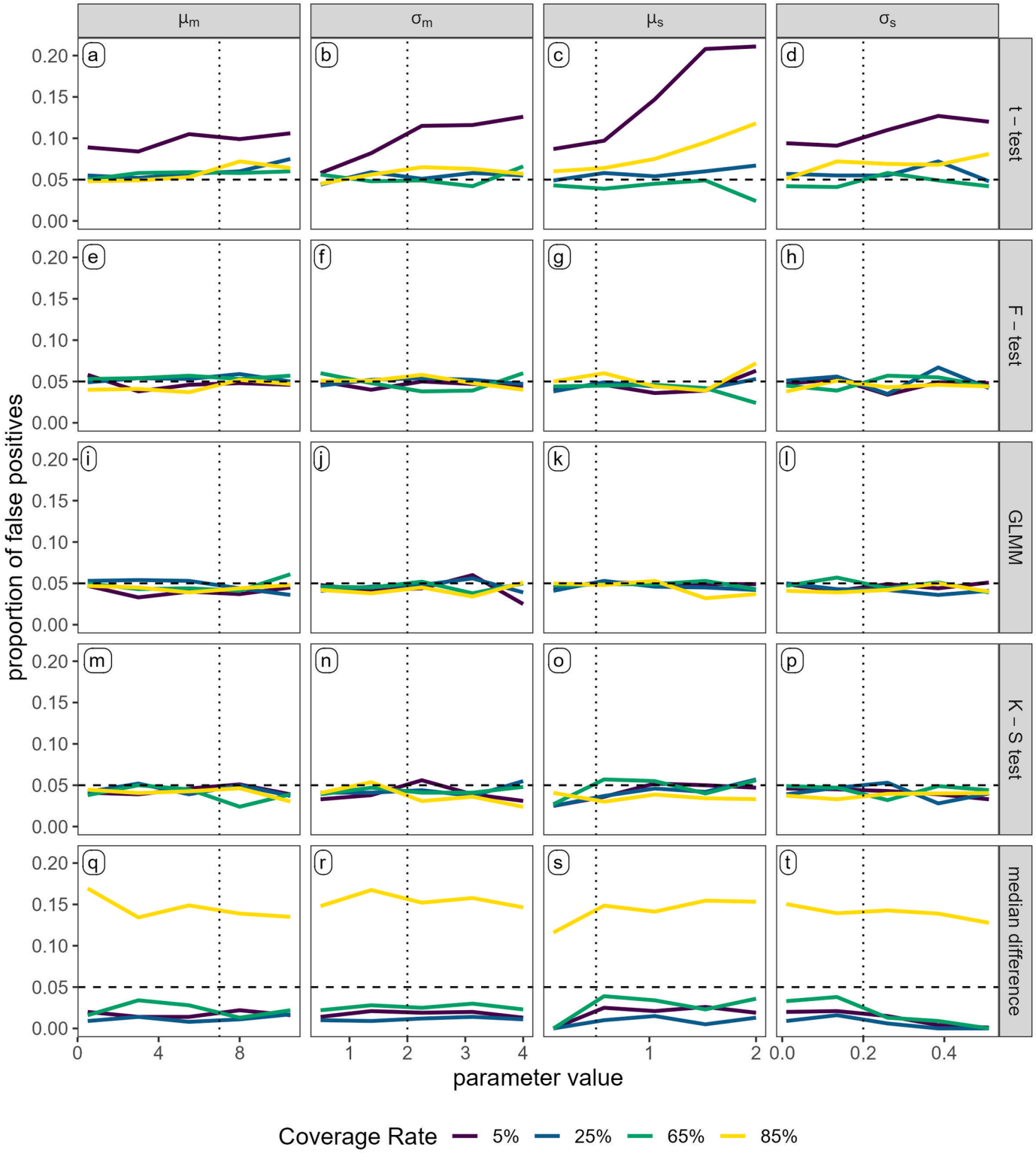



**Fig. 7.** Rate of false positives from observer effect detection tests on unbiased simulated data by changes in hyperparameter values. The horizontal dashed line indicates 0.05 (Type I error rate benchmark) and the vertical dotted lines indicate hyperparameter values used in the groundfish landings simulations.

These methods would be best used in fisheries with moderate observer coverage and for metrics that do not have a wide variation from trip-to-trip within a particular vessel. Both tests were able to distinguish between observer and deployment effects, making them important tools for determining the source of bias. The Anderson-Darling two-sample test could have been used in place of the Kolmogorov-Smirnov test, being more sensitive to deviations in the tails of the distributions, although a small scale simulation (not shown) suggested that overall strengths and weaknesses were the same. Another possible modification is the use of bias-corrected and accelerated bootstrap intervals (DiCiccio and Efron, 1996) which in a small simulation (not shown) improved detection rates for small effect sizes, but also increased false positive rates.

Benoît and Allard (2009) found statistically significant deployment

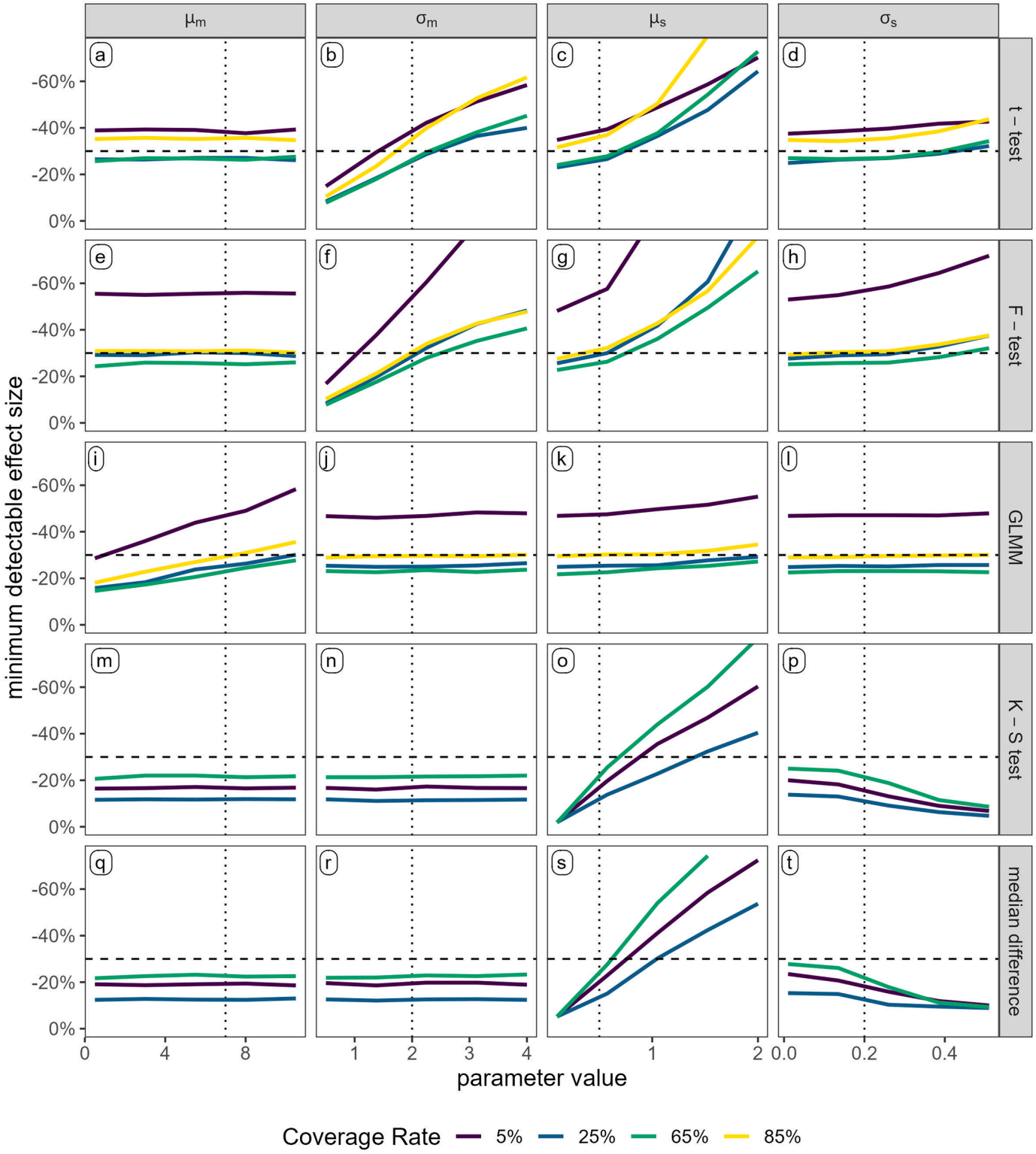



**Fig. 8.** Minimum detectable effect sizes for five tests on simulated data by changes in hyperparameter values. The horizontal dashed line indicates observer effect size of −30% and the vertical dotted lines indicate hyperparameter values used in the groundfish landings simulations.

effects prioritizing higher catches and all fisheries reported had significant observer effects using the triplet analysis. With coverage rates around 6%, these fisheries likely had observer effects of at least moderate magnitude, and the true difference was likely larger than the reported median difference. Demarest (2019) reported a mix of significant and non-significant results across various coverage rates. It is possible that a consistent observer effect could be detectable in years with higher coverage but not detectable in years with lower coverage. For example, our simulations suggest a fixed 10% decrease in groundfish landings on observed trawl trips could result in the median differences reported by Demarest in all four time periods.

The major caveat with the triplet selection method is that it does not use the entire dataset. First, trips that do not fall into a U-U-U or U-O-U sequence are removed, and then vessels which do not have at least three

**Table 2**
Results of tests for observer effects applied to actual groundfish landings data from New England large mesh bottom trawl, April 2018 through March 2019.

| Test | Statistic value | p-value | Significant? | $\widehat{b}$ estimate | $\widehat{b}$ range |
|---|---|---|---|---|---|
| t-test | -1.47 | 0.14 | No | 0.15 | -0.05 – 0.34 |
| F-test | 2.29 | 0.13 | No | | |
| GLMM | 13.67 | | Yes | -0.38 | -0.51 – −0.21 |
| K-S test | 0.08 | 0.11 | No | | |
| Median difference | 0.00 | | No | 0.00 | -0.04 – 0.00 |

sequences are dropped. Further, vessels must have at least one unobserved trip to calculate the mean unobserved value (the denominator for the calculation of the Δ values). As coverage rates rise, the likelihood of back-to-back observed trips increases, reducing the number of U-O-U sequences. Every observed trip bounded by another observed trip would be removed. At the highest coverage rates, the only vessels remaining in the analysis are those with the fewest observed trips. With deployment effects, the vessels that are preferentially covered will have a lower proportion of their trips used in the analysis. This could provide another pathway for confounding observer and deployment effects, if for example vessels with higher or lower coverage rates were more likely to change behavior on observed trips. In other words, the triplet method can only draw conclusions about a subset of trips and vessels, not the entire fleet. To a lesser extent these concerns also apply to the mixed effects model, which removed vessels that did not have at least two observed and two unobserved trips. This may explain the poor performance at the highest and lowest coverage rates, as many vessels would not contribute their data to the analysis. But the GLMM test filtering only applied to vessels, whereas the triplet method filtered at both the trip and vessel level. Randomization within the triplet selection step (i.e., choosing the first or last trip within a sequence to form a pair) also led to inconsistent results when the test was run multiple times on the same data. In general, therefore, a method that uses more of the data and provides consistent results should be preferred.

### *4.4. Application to New England groundfish landings data*

The goal of including a real-life example was not to determine conclusively whether an observer effect was occurring, but to demonstrate how conflicting results on these tests may be interpreted. Starting with the only significant result, the GLMM suggests an effect size of −38%, possibly as large as −51% or as small as −21%. At this coverage rate, the GLMM test can reliably detect effects of this size at least 75% of the time, with false positive rates less than 10%, so results can be considered reliable. The contradiction in direction of effect from the t-test estimating a positive effect while GLMM estimated a negative effect could indicate the presence of a deployment effect with a preference towards vessels with higher landings. With such a deployment effect and an observer effect around −40%, the t-test and F-test would be expected to return a false negative about 50% of the time. The K-S and median difference results do not fit neatly into this narrative, however. In this scenario, we would expect both of these tests to return a significant result nearly every time, regardless of randomization. The estimated effect size should also be in the range of about −40 to −25%, not −4 to –0%. It is possible that the observed trips that remained after triplet filtering were more similar to the unobserved trips, and the observed trips that were dropped reflected different behavior.

In the simulated data, deployment effects and observer effects were independent, and the magnitude of the observer effect was the same on all unobserved trips. In the real world, we would not expect such uniformity, and it is not unreasonable to suspect that vessels that experience disproportionately higher coverage may behave differently from vessels that are only rarely observed. For example, if vessels that were often covered on back-to-back trips were more likely to change behavior than those that were only covered sporadically, the median difference of the remaining pairs would likely be small. Thus, one explanation could be that non-random allocation of observers caused higher coverage on vessels that tended to have high landings; at the same time, those vessels tended to land less catch on trips with observers than they would otherwise, whereas vessels with lower coverage tended to have similar landings on all their trips. The combined impact of the observer and deployment effects would be that the overall observed landings were slightly, but not significantly higher than on unobserved trips, but when standardized by vessel the observed trips had roughly 40% lower landings than observed trips. Note that this is one hypothetical scenario that explains the results; others may exist.

Fisheries dynamics are complex and require thoughtful interpretation. It would be inaccurate to label any difference arising from random processes as a deliberate shift in behavior. Likewise, it is incorrect to assume that mitigation measures intended to reduce observer effects would also reduce other effects and uncertainties (such as deployment effects, observer error, etc.). When observer data are suspected of being not representative of total fishing effort, the reason for that discrepancy should be evaluated before potential mitigation actions are taken. It is also important to consider how those actions will impact the ability to determine whether the issue has been resolved. For example, suppose a management agency used the median difference test to conclude that an observer effect of magnitude −30% was occurring in a year when coverage rates were near 30%. They decide to increase observer coverage to 70% in the following year. When they analyze the second year's data, they may get a non-significant result and conclude that the issue has been resolved, when in fact the vessels were still changing behavior on the 30% of trips that were not covered.

## 5. Conclusions

The ability to accurately identify observer effects and estimate their magnitude will be important for effective fisheries management decision-making, as modifications to monitoring programs can be costly and complicated. Unfortunately, the tests investigated here did not always provide simple, accurate results. They were often contradictory with each other for the same dataset, sensitive to changes in coverage or other parameters, confounded by deployment effects, or imprecise in their estimates of effect sizes. This was under simplified conditions with constant observer and deployment effects across the fleet; real conditions would offer even more confusion and contradictions.

The selection of which test to use, if only one is to be performed, should be based on the specifics of the fishery and metric of interest. If one is only interested in the overall differences between observed and unobserved trips and not in the source of differences or vessel-specific impacts, then the t-test and F-test should be sufficient. If it is important to identify observer effects specifically and deployment effects may or may not be present, then the GLMM method would be preferred but is likely to miss smaller effect sizes. If smaller observer effects are important to identify and the coverage rates are low or moderate, then the triplet analysis may be the preferred test, with the caveat that the magnitude of the estimated effect size from the median difference is likely underestimated. Using all five tests in combination may require interpretation of conflicting results, but likely provides the most complete understanding of the fishery under investigation.

## CRediT authorship contribution statement

**Steven X. Cadrin:** Writing – review & editing, Supervision, Conceptualization. **Debra Duarte:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data Availability

The data that has been used is confidential.

## Acknowledgements

This study was conducted as part of Debra Duarte's doctoral dissertation, which was supported by the National Marine Fisheries Service Advanced Studies Program. We are extremely grateful to her committee members Anna Mercer, Gavin Fay, Geret DePiper, and Pingguo He, all of whom provided thoughtful guidance and feedback throughout this project. We also acknowledge the hard work of hundreds of fisheries observers and observer program staff in collecting and processing this invaluable data.

## References

Babcock, E.A., Pikitch, E.K., & Hudson, C.G., 2003. How much observer coverage is enough to adequately estimate bycatch? Pew Institute of Ocean Science. ⟨https://oceana.org/reports/how-much-observer-coverage-enough-adequately-estimate-bycatch⟩.

Benoît, H.P., Allard, J., 2009. Can the data from at-sea observer surveys be used to make general inferences about catch composition and discards? Can. J. Fish. Aquat. Sci. 66 (12), 2025–2039. https://doi.org/10.1139/F09-116.

Burnham, K.P., Anderson, D.R., 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociol. Methods Res. 33 (2), 261–304. https://doi.org/10.1177/0049124104268644.

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

Davies, S.L., & Reynolds, J.E., Eds., 2002. Guidelines for developing an at-sea fishery observer programme. FAO Fisheries Technical Paper No. 414. 116p. ⟨https://www.fao.org/fishery/en/publications/77350⟩.

Demarest, C., 2019. Evaluating the observer effect for the Northeast U.S. groundfish fishery. https://d23h0vhsm26o6d.cloudfront.net/210427_Amendment-23_Volume-II-Appendices.pdf, 248-274.

DiCiccio, T.J., Efron, B., 1996. Bootstrap Confidence Intervals. Stat. Sci. 11 (3), 189–212. https://doi.org/10.1214/ss/1032280214.

Faunce, C.H., Barbeaux, S.J., 2011. The frequency and quantity of Alaskan groundfish catcher-vessel landings made with and without an observer. ICES J. Mar. Sci. 68 (8), 1757–1763. https://doi.org/10.1093/icesjms/fsr090.

GARFO (Greater Atlantic Regional Fisheries Office), 2021. Summary of analyses conducted to determine at-sea monitoring requirements for multispecies sectors FY2019. ⟨https://www.greateratlantic.fisheries.noaa.gov/ro/fso/reports/Sectors/ASM/FY2021_Multispecies_Sector_ASM_Requirements_Summary.pdf⟩.

Hall, M.A., 1999. Estimating the ecological impacts of fisheries: What data are needed to estimate bycatches? Proceedings of the International Conference on Integrated Fisheries Monitoring, 175–184. ⟨https://www.fao.org/3/x3900e/x3900e04.htm⟩.

Henry, A., Demarest, C., & Errend, M., 2019. Modelling Discard Incentives for Northeast Multispecies (Groundfish) Stocks. Groundfish Plan Development Team Document. ⟨https://d23h0vhsm26o6d.cloudfront.net/210427_Amendment-23_Volume-II-Appendices.pdf⟩, 484-520.

Kerr, L.A., Weston, A.E., Mazur, M.D., & Cadrin, S.X. 2020. Evaluating the impact of inaccurate catch information on New England groundfish management. ⟨https://d23h0vhsm26o6d.cloudfront.net/210427_Amendment-23_Volume-II-Appendices.pdf⟩, 707-782.

Liggins, G., Bradley, M., Kennelly, S., 1997. Detection of bias in observer-based estimates of retained and discarded catches from a multi species trawl fishery. Fish. Res. 32 (2), 133–147. https://doi.org/10.1016/S0165-7836(97)00053-2.

NEFMC (New England Fishery Management Council), 2021. Northeast Multispecies Fishery Management Plan, Amendment 23. ⟨https://s3.amazonaws.com/nefmc.org/210809_Groundfish_A23_FEIS_final_submission.pdf⟩.

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨https://www.R-project.org/⟩ (URL).

Rago, P.J., Wigley, S.E., Fogarty, M.J., 2005. NEFSC bycatch estimation methodology: Allocation, precision, and accuracy. NEFSC Ref. Doc. 05–09. ⟨https://repository.library.noaa.gov/view/noaa/3455⟩.

Rudd, M.B., Branch, T.A., 2017. Does unreported catch lead to overfishing? Fish Fish. 18 (2), 313–323. https://doi.org/10.1111/faf.12181.

Suuronen, P., Gilman, E., 2020. Monitoring and managing fisheries discards: New technologies and approaches. Mar. Policy 116, 103554. https://doi.org/10.1016/j.marpol.2019.103554.

Wigley, S.E., & Tholke, C., 2020. 2020 discard estimation, precision, and sample size analyses for 14 federally managed species groups in the waters off the Northeastern United States. NOAA Technical Memorandum, NMFS-NE-261. ⟨https://repository.library.noaa.gov/view/noaa/25521⟩.