Examining influences on observed counts from shoreline surveys of marine debris

A report for the NOAA Marine Debris Program

Version 1.0

Hillary K. Burgess, Timothy T. Jones, Jacqueline K. Lindsey and Julia K. Parrish

June 30, 2020

Contents

Executive Summary

Utilizing a combination of controlled field experiments and marine debris shoreline surveys, we explored factors that may influence observed counts of marine debris during shoreline monitoring.

This project was designed specifically to uncover causes of error in observed counts relative to actual counts of marine debris, and to make recommendations for a) minimizing such sources of error or b) informing results interpretation that can take such known errors into account. Key findings and recommendations of direct relevance to marine debris shoreline monitoring protocols include:

- Several uncontrollable factors influence the rate of debris detection: item size and color, distance from observer, and surrounding material or substrate
  - counts can be considered minimums especially for items that are smaller (between 2.5 and 5cm) and duller (less brightly colored) items
  - counts can be presented as a range that will be broader (more variable) for smaller, duller items and in situations where visual obstructions (e.g., variably sized and colored cobble compared to monotone sand substrate or a vegetated area where detection may depend on line of site) occur.
  - could consider increasing the minimum size of debris items searched for (currently 2.5cm) or switching to a smaller search area for smaller items
- For a single observer looking in one direction, the rate of detection of marine debris declines beyond a visual distance of greater than 2.5m from the observer, especially for items at the smaller end of the range tested.
- For a single observer searching in two directions at once (i.e. looking left and right), the rate of detection is lower than when looking only in one direction.
- Two observers had a higher survey-level rate of detection than individuals searching alone, and three observers had a higher rate of detection than two.
- Inter-observer variability (i.e. number of items detected by one observer compared to another) was considerable, but at the survey level, variability decreased with increased team size
  - survey effort (i.e., # observers) should be as consistent as possible across sites, with an ideal team size of three. However, we did not test group sizes larger than three.
  - personnel should be as consistent as possible within a site
  - if a survey must be conducted by an individual observer, they should scan no more than 2.5m at a time, and only searching in one direction at a time (replicating effort by two observers)
- We found no effect of removal of marine debris on observed counts during subsequent surveys

- observers removing debris during their surveys is unlikely to impact data, which should be considered a snapshot of debris presence in time, rather than a month over month accumulation rate
- consider designing special projects to investigate accumulation/turnover rates at shorter intervals between surveys incorporating either treatment (removal or not)

- Inclusion of the back barrier (vegetation) does increase overall debris loads on a transect (i.e., raises the average concentration), and anecdotally this effect appears to be linked to vegetation type, where denser "grabbier" vegetation contains more debris
  - consider incorporating the back barrier into surveys, but keep these counts separate from the rest of the transect. Because vegetation is variably present and accessible, it may be worthwhile to set a short distance that can be applied consistently across sites that have a vegetated area that can be accessed. We found that debris density decreases as distance into the vegetation increases, so variable distances would bias estimated densities.
  - Including a portion of the back barrier will document those items that have been deposited by very high tides or storm surges, were blown upward (weigh less) or were left behind by beachgoers

Introduction

The NOAA Marine Debris Program (NOAA MDP) published a shoreline survey field guide in 2012. The standing stock protocol (Opfer et al. 2012, Lippiatt et al. 2013) involves cataloging marine debris items that are ≥2.5cm in size within four randomly selected 5m wide transects sampled within a permanent 100m long plot and where debris is left in place. Initial analysis of data from this effort raised questions about the influence of individual observers, team sizes, effort (# observers participating),) actual debris presence and other aspects of volunteer surveys that influence debris detection (Hardesty et al. 2017).

The Coastal Observation and Seabird Survey Team (COASST) is a citizen science program based at the University of Washington which partners with local, state and federal agencies, tribes and environmental organizations, and involves over 1000 people in beach-based monitoring. COASST has designed, developed, field-tested and adaptively managed two modules: beached birds (est. 1999) and marine debris (est. 2014). The sampling design for the marine debris module was modelled after the NOAA MDMAP standing stock method with additional aspects designed to produce knowledge about the sources and impacts of shoreline debris.

This project had two main objectives:

1. To evaluate factors that may influence observed counts of marine debris during shoreline monitoring
2. To evaluate two areas of distinction between the two protocols (MDMAP vs COASST)
    a. Debris removal during the survey
    b. Inclusion of the vegetated back barrier in the search area

This effort is intended to inform protocol and training improvements, and results interpretation for volunteer-based shoreline marine debris monitoring, with the ultimate goal of facilitating a robust and coordinated network of marine debris monitoring efforts nationwide.

Using controlled field scenarios (referred to as “field trials”) and protocol iterations of the NOAA Marine Debris Monitoring and Assessment Project (MDMAP) standing stock method (referred to as “fixed plots”) we explored the following questions:

Question 1: What is the effect of an observer(s) search pattern on debris counts?

Question 2: What is the effect of number of observers per transect on debris counts?

Question 3: If an observer is short on time, should they complete a reduced number of full-width transects (5m) or decrease the width of the transect (e.g., 2.5m) while maintaining 4 replicates?

Question 4: Does removing debris during standing stock surveys affect future counts/load estimates?

Question 5: How does including a portion of the back barrier (i.e., dune line, vegetation) in the transect influence the number of debris items found?

Key Definitions

Field trial – a simulated survey event under which certain aspects of the survey are controlled and others are tested. Two types of field trials occurred. The first set was designed to examine Question 1 and occurred in the initial phase of the work. The second set was designed to address Questions 2 and 3 respectively, during a latter series of events, and were informed by results of Question 1. The distinctions between the field trials are defined by which factors were controlled for in the set-up of simulated survey transects, and which were modified.

Fixed Plot – a section of beach that is typically sub-sampled using transects. Plots may be 100m wide (parallel to the shoreline) as in standing stock surveys, or 25m wide in some field trials.

Transect – a sub-section of a plot that runs perpendicular to the shoreline, 5m wide in standard standing stock surveys. Width was a variable, modified (and examined) in some field trials.

Observer – a person who may or may not have scientific training, recruited from the general public and the University of Washington (where COASST is affiliated), who participated in the field trials to conduct simulated marine debris surveys.

Search – the act of looking for marine debris, different search patterns were prescribed during field trials (e.g. one directional vs left and right).

Survey – a complete search of a prescribed area (one or more transects). During field trials, surveys were conducted by one or more observers using prescribed search patterns, and formed the sampling unit of analysis. Fixed plots were also surveyed in uncontrolled (natural) conditions, following a modified MDMAP standing stock survey protocol as described below.

Gross detection rate – the proportion of items observed during a survey relative to the known number of possible items.

Detection rate – the conditional probability of detecting an item given its presence.

Study Sites

These studies were conducted across seven sites in the state of Washington. Sites were chosen for accessibility from Seattle (where COASST is based), availability of restrooms and parking for hosting field trials, ease of conducting repeated fixed-plot surveys (described below in Key Definitions), and representation of different substrates and features of Puget Sound and the Pacific Coast. Below we provide a description of the six sites.

Kalaloch is located in Olympic National Park on the Pacific coast just south of Kalaloch Lodge and faces west-southwest. A fixed-plot was established and regularly surveyed by COASST staff were conducted here, but no field trials were conducted due to prohibitive weather and limited volunteer recruitment to this remote area.

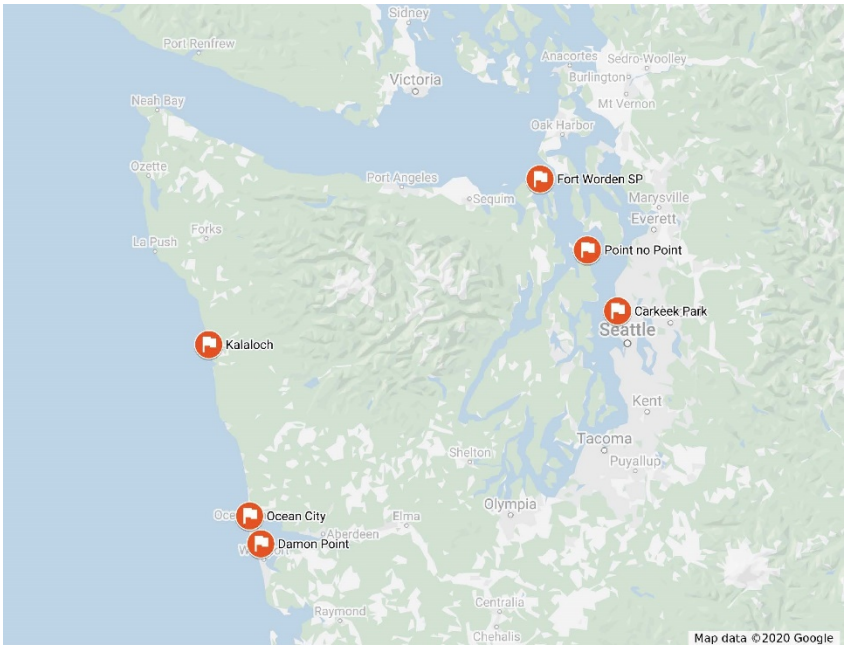
Two sites were used in Fort Worden Historical State Park, located at the end of the Quimper Peninsula**: Point Wilson**, which faces north into the Strait of Juan de Fuca was established as a fixed-plot survey site (to address questions 4 and 5), while **Port Townsend Marine Science Center (PTMSC)** which faces east-south-east into Port Townsend Bay was used as a field trial site (to address questions 1-3). The former was chosen because of its influence by the Strait, whereas the latter was chosen to provide access to restrooms and a parking lot for volunteers during field trials.

Point no Point is located at the end of the Kitsap Peninsula in Puget Sound. This beach faces east, running south from the Point No Point Lighthouse. Fixed-plots and field trials were conducted here.

Damon Point is a small isthmus at the southern tip of the Point Brown Peninsula at the entrance to Grays Harbor. The beach faces WSW and the Pacific Ocean. Fixed-plots and field trials were conducted here.

Ocean City is a state park located on the north end of the Point Brown Peninsula facing west to the Pacific Ocean. Field trials were conducted here.

Carkeek Park is a city park located in north Seattle facing north-northwest on Puget Sound. Fixed-plots and field trials were conducted here.



Map of study site located within Washington State.

Question 1. What is the effect of search pattern on debris counts?

Methods

We designed a set of controlled field trials to determine how detection rate of marine debris varies with distance from observer under a variety of conditions and search patterns. During each field trial, we set-up three, 5m wide transects that extended the length of water's edge to 5m into the vegetation (if present). Transect locations were chosen to achieve diversity in beach profile (attempting to encompass wrack, driftwood and vegetation) as well as substrate (sand, cobble), while also avoiding high-traffic areas of the beach to limit disrupting recreation and vice-versa.

Each transect was first thoroughly cleaned of all marine debris by several COASST staff members repeatedly passing over the entirety of the transect. Then, each transect was seeded with twenty known debris items (> 2.5cm). Cleaning the beach and seeding a known quantity, location, and type of debris allowed us to document which items were detected vs undetected during each survey.

To inform the number and types of debris needed to seed transects, COASST first analyzed our existing dataset to determine average debris densities in 5m wide transects as well as frequencies of item identity, size, color and material as defined in the COASST marine debris protocol (Parrish & Burgess 2017).

We found that across the COASST dataset, the mean number of items per transect is 4.43 with a SD of 16.6, so we set about determining a sample of 20 items per transect that would

be representative of typical debris found. This sample size was an intersection of realistic quantity allowing for a range of distances and characteristics, while being reasonable to manage and recover during a field trial. For composition of seeded debris items, we calculated the proportion of different debris types, materials, sizes and colors within the COASST dataset. We compared the proportion ranking of COASST data to MDMAP top 10 item types and found that they were largely aligned. We then treated the COASST proportions as probabilities to generate samples of 20 debris items representative of what is found during surveys. Debris kits for each transect were created based on these analyses, and each item in a kit was labeled with a serial number (directly writing on larger items, attaching aluminum tags for smaller/lighter items), and inventoried so that detection could later be related back to debris characteristics. See Appendix A for a table and images of kit contents.

During set up, debris items from the transect kit were seeded to approximate an *even* distribution throughout the transect to allow each observer to be exposed to the full range of potential visual distances that may be present in a transect. Debris items were spread apart to avoid aggregations close to the observer or farther away which would bias the visual distance levels and skew the analysis. Transects were delineated with lines of numbered colored flags (Figures 1A, 1B) to form a grid wherein the right (when facing the water) edge was lined with equally spaced orange flags, the center was lined with equally spaced pink flags, and the left side was lined with equally spaced blue flags. Flags were numbered from the vegetation toward the water, and debris was seeded to ensure that there was at least one item closest to each flag relative to the flags surrounding it using a coordinate system (e.g. 1-Orange in Figure 1A would be the orange flag within the vegetation, 6-Blue would be the blue flag closest to the water), locations of each item were marked on the datasheet to ensure that each flag had at least one associated item, and to aid in item retrieval at the end of the field trial. To ensure that items did not move throughout the field trial, smaller/lighter weight items were anchored to the substrate using the aluminum tag which was buried. Staff recorded the coordinate (e.g. 1-Orange), beach zone (surf, wrack, bare, driftwood, vegetation; Figure 1A) and substrate (cobble or sand) of each debris item placement. The length of each transect was also measured.

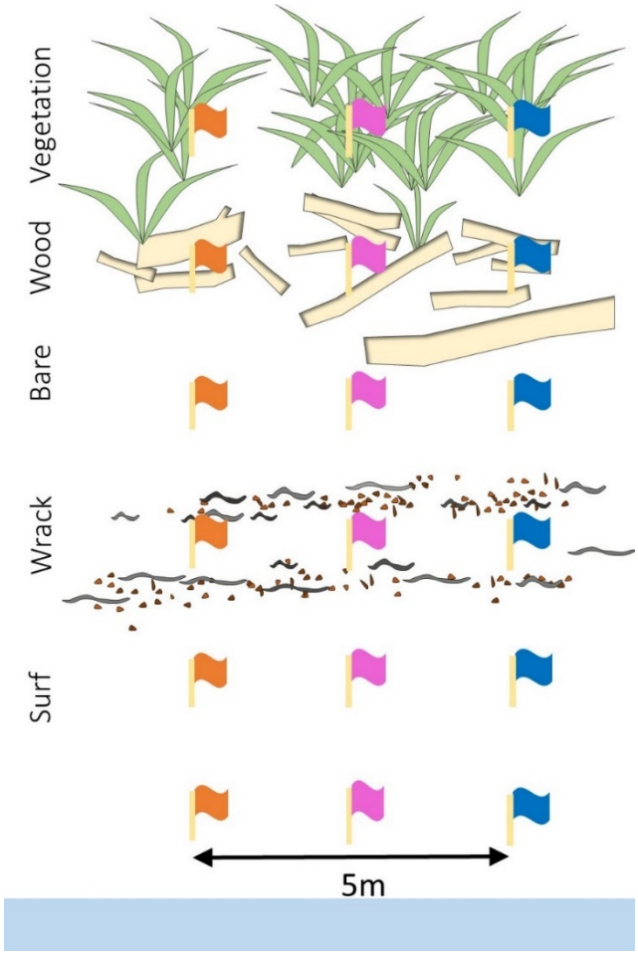


Figure 1A. Items were seeded in 5m wide transects delineated by flag lines and beach zones. Transects ran the length of the beach from the edge of the water up to 5m into the vegetation (if present and accessible). Flags were numbered from the vegetation toward the water and used to map seeded debris and approximate even distribution. For example, at least one item was placed closest to 1-Orange, one closet to 2-Orange, 1-Pink, 1-Blue…6-Blue.

Figure 1B. Layout of a field trial at PTMSC. Lines of different colored flags were used to delineate the edges and midline of 5m wide transects.

Volunteer observers for these and other field trials described later were recruited for each event from existing COASST participants, as well as advertisement through regional partner organizations.

Upon arrival, each volunteer observer was provided instructions and assigned to one of three search patterns (Figure 1C) to survey every transect:

- Edge-left: required searching from the left edge of the transect (when facing the water) looking in toward the midline
- Midline: required walking down the middle of the transect searching in both directions.
- Edge-right: required searching from the right edge of the transect looking in toward the midline

Our goal was to have each observer survey each transect once, conducting a different search pattern each time, and to have each transect surveyed by multiple different observers conducting different search patterns.

Instructions included to:

- begin the survey from the back of the beach and walk toward the water
- look for manmade items roughly the size of a bottle cap (2.5cm) or larger

- walk toward the water scanning for items without backtracking or looking backwards
- stop and call out to a COASST data recorder when an item is sighted
- not stray from the assigned flag line

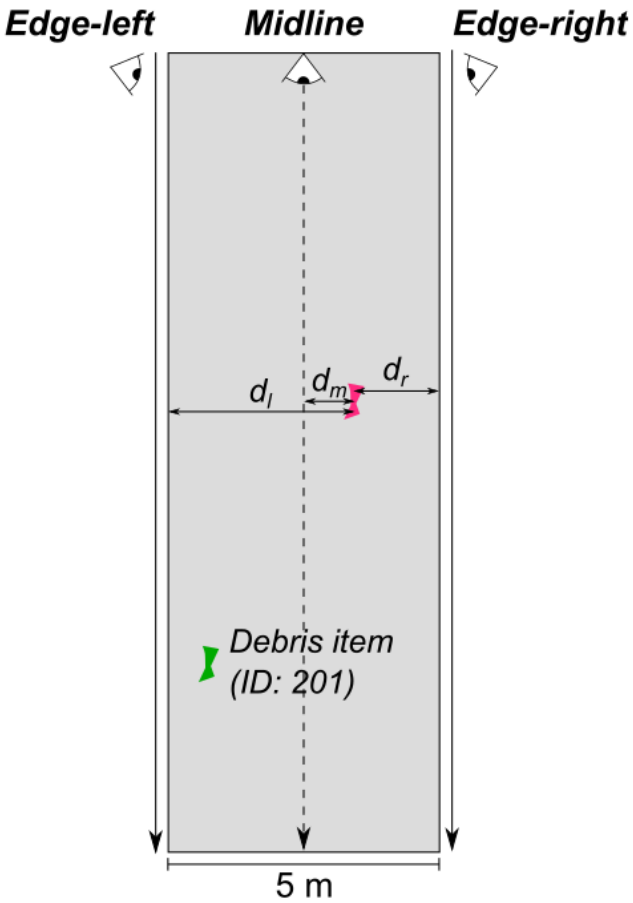


Figure 1C. Search patterns employed in the first set of field trials. Transects were seeded with debris, and dependent on the search pattern (Edge-left, Midline or Edge-right), measurements of distance from observer (dl, dm, and dr, respectively) were made from the corresponding search line (delineated by a line of flags), to the piece of debris found. Each debris item was labelled with a unique ID code to enable item characteristics (size and color) to be associated with detection/non-detection.

KIT 1		ZONES		Veg:		Wood:		Bare:		Wrack:	
HOST		DATE				Vol ID	Vol ID	Vol ID	Vol ID	Vol ID	Vol ID
PACER		LOCATION									
					Circle Search:	EL Mid ER	EL Mid ER	EL Mid ER	EL Mid ER	EL Mid ER	EL Mid ER
LOCATION (color-#)	ITEM #.	ID	COLOR	SUBSTRATE	ZONE	Distance (cm)	Distance (cm)	Distance (cm)	Distance (cm)	Distance (cm)	Distance (cm)
	101	cap	clear								
	102	plastic chunk	white								
	103	plastic chunk	white								
	104	plastic chunk	white								
	105	rope	yellow								
	106	cup part	multi/clear								
	107	plastic chunk	white								
	108	plastic chunk	white								
	109	cup part	multi/clear								
	110	car part	blue								
	111	styrofoam square	white								
	112	cup part	multi/clear								
	113	styrofoam circle	white								
	114	styrofoam	white								
	115	cup part	multi/clear								
	116	shotgun wad	white								
	117	rubber	black								
	118	wrapper	brown								
	119	lumber	brown								
	120	lumber	grey								

Figure 1D. Example datasheet. The location of each item was recorded using proximity to numbered colored flags to aid in even distribution and item recovery after the field trial.

During the survey, observers were accompanied by a COASST staff member who followed behind (as to not inadvertently influence gaze or interfere). When an observer sighted an item, the staff member recorded the serial number and measured the distance between the observer and the item. To measure distance, the observer held the 0 mark of the measuring tape at their waist and the staff member stretched the tape maintaining the height perpendicular to the ground over the item, and looked straight down to record where item overlapped with tape (Figure 1D).

Following each field trial, all seeded debris items were removed and inventoried for each transect to ensure that no debris was introduced into the environment. Although COASST staff scrutinized transects during and between surveys to reduce chances of accidental burial or movement of items, there were three instances when items were not recoverable at the end of a field trial due to interference from playful public children. In these cases, the item was replaced in its respective kit with another of like type and characteristics.

Field trials of this type were conducted on three dates at a different site each date: PTMSC, Damon Point, and Carkeek Park.

Table 1A. Field trial sampling details.

Location	Date	Number of observers	Number of transects	Number of replicates		
				Edge-left	Midline	Edge-left
PTMSC	9/16/2018	19	3	16	16	13
Damon Point	11/3/2018	19	2	3	12	4
Carkeek Park	4/28/2019	18	3	16	18	17

Analytic approach

For each survey we calculated gross detection rates as the proportion of items observed out of the twenty seeded (e.g. if 10 items were sited the detection rate was 50%). Subsequently we calculated mean and 95% confidence intervals of mean gross detection rates via bootstrap resampling (1000 permutations) for each of the search patterns. To control for inter-transect variability in detection rate we also calculated a bootstrap mean (±95% CI) of paired differences in detection rate between midline and edge search patterns. The bootstrap procedure consisted of selecting a midline search pattern survey (midline survey) at random from the pool of all such surveys, and then selecting an edge search pattern survey (edge survey) at random from the pool of such surveys carried out at the same site and transect as the selected midline survey. The difference in gross detection rate between these paired surveys was then calculated. This was repeated N times, where N was the total midline surveys available (there were fewer midline surveys than edge surveys, and so this N represents the effective sample size for paired differences) to generate a set of paired differences. The mean paired difference was calculated for that set. This process was repeated 1000 times to generate a bootstrap distribution of the mean paired difference, which was then converted into a bootstrap mean and 95% confidence interval based on the distribution of mean paired differences across permutations.

In order to identify how search pattern, visual distance, item size, color and observer affect detection rate, we tested a series of mixed effects models in R (glmer in package lme4). Our dependent variable was a binomial variable for each debris item within each transect, taking a value of 1 if the item was detected, and 0 otherwise. Modelling detection at the item level allowed us to model detection curves as a function of debris visual distance from observer. Because we seeded transects with known quantities of debris, modelled detection rate was absolute as total debris quantity was known for each survey. We considered eight independent variables consisting of five fixed effects (visual distance, item size, site, search pattern, item color), and three random effects that account for observer, transect, and survey level variability, respectively. The distance effect only entered into the model as an interaction with search pattern because search patterns covered different distance ranges (edge = 0-5m, midline = 0-2.5m). We tested all possible combinations of these factors (Table 1B). Because item color had many levels, we conducted an exploratory analysis to determine if some colors could be grouped (Appendix B). We ultimately grouped colors into bright (blue, yellow, red), dull (black, brown, grey), clear (including transparent with some colorful lettering as in a “compostable” water cup), and white albeit without random effects, to reduce the number of parameters fitted for color. Although the initial analysis showed that clear and white could be grouped together, we chose to keep them separate as white is the most common color of marine debris in the COASST dataset. To represent item size, we included maximal item surface area calculated as the product of the item’s two longest dimensions, which we subsequently square root transformed to reduce the spread of points and leverage effects of larger items. Data from Damon Point was not included in

these analyses, because with limited staff and observers who were primarily students on a class field trip, not all factors were documented (substrate, beach zone) and we felt it was not representative in terms of quality of observer effort; detection rates were overall lower and more variable

Because response variables were binomial, we used binomial Generalized Linear Mixed Models (GLMM; logistic link function), such that all continuous factors (distance and maximum surface area) are modelled as a logistic function on the response scale. Models representative of all possible predictor combinations were fitted to the data, and subsequently ranked based on small sample size corrected Akaike's Information Criterion (AICc). Small sample size corrected Akaike's Information Criterion was used for model selection. AICc reflects model fit, penalized by model complexity (number of model parameters), which scales with sample size (greater penalization for smaller sample sizes), to reduce the selection of overfitted models when analyzing smaller datasets. Among a set of models, those that minimize AICc are considered the most parsimonious.

Table 1B. Predictors used in detection rate models.

Name	Type	N_{levels}	Levels/range	Fixed/random
visual distance	continuous		0 - 5	fixed
item size[a]	continuous		1.5-18.5	fixed
site	factor	2	Carkeek, PTMSC	fixed
search pattern	factor	2	edge, midline	fixed
item color	factor	4	bright, dull, clear, white	fixed
observer id	factor	37		random
survey id	factor	96		random
transect id	factor	6		random

[a]*Item size was included in the model as the square root of the item's maximal visual surface area (product of each item's longest two dimensions)*

Within the initial GLMMs we tested size and distance as linear model components, which result in logistic curves due to the logistic link function. However, detectability may not vary exactly according to a logistic function. To account for this, we modified our modelling approach by including quadratic terms for item size and distance and examined whether they improved model fit. We constructed all possible models including site, item size (linear + quadratic), color, search pattern and distance (linear + quadratic) as well as random effects of transect, survey and observer and ranked them based on AICc, and identified whether the resultant best model (minimum AICc) was better than the model selected using only linear terms for item size and visual distance.

Item location information was collected in the form of beach zone (wrack, bare, wood and vegetation) and substrate (cobble, sand) at all transects at the Carkeek and PTMSC field trials except one transect from Carkeek. To examine the effects of item location further we

repeated the modelling procedure on the subset of data that had item location and substrate information, but limited the predictors to those found to be important when analyzing the entire dataset.

Results

Summary of surveys

A total of 115 surveys of transects were performed over three field trials (Table 1C). Edge surveys (no matter the side) had a slightly lower gross detection rate (proportion of items found that were present) than midline surveys (Edge = 70.5% with bootstrap 95%CI of 66.8-73.8%; Midline = 72.4% with 95% CI of 68.0% - 76.4%). Paired differences between midline and edge surveys (i.e. controlling for inter-transect differences by pairing comparisons within transects) corroborated this, with an average difference of +3.4% for midline surveys over edge surveys, although the 95% CI of this mean overlapped with zero (-2.7% - 9%) suggesting that, when site/transect variation was controlled, differences among search patterns were not statistically significant. To ensure that Damon Point was not driving the overall pattern, we calculated summary statistics excluding Damon Point, and found they largely reflected overall statistics, albeit with marginally higher detection rates (Edge = 72.5% CI=69.2 - 76%, Midline = 76% CI=72.2 - 79.9%). Paired differences between edge and midline surveys were unaffected by the exclusion of Damon Point, due to the relatively low sample size at Damon Point (Delta = 3.3% CI=-2.7 - 9.2%).

Table 1C. Summary statistics of field trials representing totals show the median across site×transect combinations, except for the bootstrap, which is the bootstrap 95% CI across surveys for that search pattern.

				DETECTION RATE					
Site	Tran.	Search	N	Mean	Median	Min	Max	sd	95% CI
Carkeek	1	Edge-L	3	58	60	55	60	3	[55-60]
Carkeek	1	Midline	9	76	80	60	90	10	[69-82]
Carkeek	1	Edge-R	4	78	80	65	85	9	[69-84]
Carkeek	2	Edge-L	6	73	73	70	75	3	[71-74]
Carkeek	2	Midline	6	72	70	60	85	8	[66-78]
Carkeek	2	Edge-R	5	64	60	40	85	17	[52-78]
Carkeek	3	Edge-L	7	71	75	55	85	11	[64-79]
Carkeek	3	Midline	3	75	70	70	85	9	[70-85]
Carkeek	3	Edge-R	8	59	60	40	70	10	[52-66]
Damon	2	Midline	8	63	70	20	85	20	[47-74]
Damon	2	Edge-R	1	65	65	65	65		
Damon	3	Edge-L	3	62	65	50	70	10	[50-70]
Damon	3	Midline	4	60	65	35	75	18	[44-73]
Damon	3	Edge-R	3	40	45	30	45	9	[30-45]
PTMSC	1	Edge-L	4	90	90	80	100	9	[82-98]
PTMSC	1	Midline	6	84	80	80	95	7	[80-89]
PTMSC	1	Edge-R	4	60	65	40	70	14	[46-69]
PTMSC	2	Edge-L	6	90	93	80	100	8	[84-96]
PTMSC	2	Midline	4	79	85	50	95	20	[60-93]
PTMSC	2	Edge-R	5	87	90	75	95	9	[79-94]
PTMSC	3	Edge-L	6	73	73	65	80	7	[68-78]
PTMSC	3	Midline	6	72	70	50	85	13	[63-80]
PTMSC	3	Edge-R	4	69	70	60	75	6	[63-74]
		Edge-L	**35**	**73**	**77**	**68**	**83**	**7**	
Total (median)		**Midline**	**46**	**75**	**76**	**62**	**89**	**11**	
		Edge-R	**34**	**66**	**71**	**53**	**80**	**11**	

Statistical analysis of detection rate

Modelling detection rate at the individual item level, the model with the lowest AICc value included the fixed effects of item size and color, visual distance from observer, site and random effects of observer and transect (Table 1D). Four models had AICc within 2 of the best overall model ($\Delta_{AICc} < 2$), but differed only based on the inclusion of survey or transect random effects, and the inclusion of search pattern (Table 1D). Of the random effects, observer id had the highest summed Akaike weight, and survey id had the lowest weight and was omitted from the highest ranked model (Table 1D). Among the fixed effects, the interaction between search pattern and visual distance was included in all the highest ranked models, as were item size and color, whereas site and search pattern (affecting the intercept) was omitted from several of the highest ranked models (Table 1D). Although search pattern was omitted from the highest ranked models, and received the lowest

summed Akaike weight, differences among survey patterns were modelled via the interaction with distance, suggesting that differences in detectability among search patterns were negligible at short distances, but began to diverge at greater distances from the observer (Table 1D).

Table 1D. Model selection table. All models with $\Delta_{\mathrm{AICc}} < 4$ are shown. Model specific Akaike weights, $\omega_{\mathrm{AICc}} = e^{-\Delta_{\mathrm{AICc}}/2}$, are given, as are the summed Akaike weights for model predictors, Σ_{AICc}, calculated as the sum of ω_{AICc} for models containing the corresponding predictor. Predictors noted as (1|volunteer) refer to random effects of the stated grouping variable.

#	Model equation	AICc	Δ_{AICc}	ω_{AICc}
1	site + item_color + item_size + pattern:distance + (1|volunteer) + (1|transect)	1816.1	0.00	0.22
2	site + item_color + item_size + pattern:distance + (1|volunteer) + (1|transect) + (1|survey)	1817.0	0.85	0.15
3	site + item_color + item_size + pattern:distance + (1|volunteer) + (1|survey)	1817.6	1.48	0.11
4	site + item_color + item_size + pattern + pattern:distance + (1|volunteer) + (1|transect)	1818.0	1.90	0.09
5	item_color + item_size + pattern:distance + (1|volunteer) + (1|transect)	1818.8	2.64	0.06
6	site + item_color + item_size + pattern:distance + (1|volunteer)	1818.8	2.66	0.06
7	site + item_color + item_size + pattern + pattern:distance + (1|volunteer) + (1|transect) + (1|survey)	1818.9	2.75	0.06
8	site + item_color + item_size + pattern:distance + (1|survey)	1819.3	3.16	0.05
9	site + item_color + item_size + pattern + pattern:distance + (1|volunteer) + (1|survey)	1819.5	3.40	0.04
10	item_color + item_size + pattern:distance + (1|volunteer) + (1|transect) + (1|survey)	1819.7	3.56	0.04

		Item		Search		Random effects		
	site	color	size	pattern	pattern: distance	volunteer	survey	transect
Σ_{AICc}	0.85	1	1	0.28	1.	0.89	0.52	0.7

Because replicate search patterns were unbalanced among transects within a field trial (Table 1B), the broader effect of site or transect effects included in the model need to be addressed. Unbalanced survey effort was most prominent at Carkeek transects 1 and 3, with transect 1 being dominated by midline surveys, and transect 3 dominated by edge surveys. At the site level (included as a random controlling factor in the model) this is unlikely to contribute to any differences relative to a more uniform spread of survey effort, as overall the Carkeek field trial (i.e. across transects) received uniform effort across the search patterns. However, the transect effect (also included as a random controlling factor) may be influenced by unbalanced survey effort as the majority of data from Carkeek transect 1 was from midline surveys, and Carkeek transect 3 edge surveys. The potential outcome of this is that variability due to differences between edge and midline surveys may be being accounted for by the random transect effect, rather than the fixed effect of search pattern, or vice versa (i.e. fixed effect of search pattern absorbing some of the among transect

differences). However, because the design was not completely unbalanced (i.e. edge surveys represented a minimum of 44% and a maximum of 83%, relative to a completely balanced 66%), the effects are likely to be minimal on model conclusions.

The inclusion of a quadratic term for item size resulted in a better fit model than including size as a linear term only, whereas quadratic terms for distance were not supported (Table 1E). The highest ranked model among these was the same as in the previous analysis, but with the addition of a quadratic term for item size (Table 1E).

Table 1E. Model selection table including quadratic terms for distance and item size. All models with Δ_AICc < 3 are shown. Model specific Akaike weights, $\omega_{\mathrm{AICc}} = \mathrm{e}^{-\Delta_{\mathrm{AICc}}/2}$, are given, as are the summed Akaike weights for model predictors, Σ_{AICc}, calculated as the sum of ω_{AICc} for models containing the corresponding predictor. Predictors noted as (1|volunteer) refer to random effects of the stated grouping variable.

#	Model equation	AICc	Δ_{AICc}	ω_{AICc}
1	site + item_color + item_size + item_size^2 + pattern:distance + (1|volunteer) + (1|transect)	1808.6	0.0	0.13
2	site + item_color + item_size + item_size^2 + pattern:distance + (1|volunteer) + (1|survey) + (1|transect)	1809.5	0.86	0.09
3	site + item_color + item_size + item_size^2 pattern:distance + (1|volunteer) + (1|survey)	1810.2	1.54	0.06
4	site + item_color + item_size + item_size^2 + pattern + pattern:distance + (1|volunteer) + (1|transect)	1810.6	1.95	0.05
5	site + item_color + item_size + item_size^2 + pattern:distance^2 + (1|volunteer) + (1|transect)	1810.8	2.16	0.05
6	item_color + item_size + item_size^2 + pattern:distance + (1|volunteer) + (1|transect)	1811.3	2.65	0.04
7	site + item_color + item_size + item_size^2 + pattern:distance + (1|volunteer)	1811.4	2.72	0.03
8	site + item_color + item_size + item_size^2 + pattern + pattern:distance + (1|volunteer) + (1|survey) + (1|transect)	1811.5	2.82	0.03
9	site + item_color + item_size + item_size^2 + pattern:distance^2 + (1|volunteer) + (1|survey) + (1|transect)	1811.6	2.92	0.03

	Item				Search			Random effects		
	site	color	size	size^2	pattern	distance	distance^2	vol	surv	transect
Σ_{AICc}	0.86	1	1	0.98	0.31	0.77	0.41	0.90	0.52	0.68

Our final fitted model included fixed effects for item color (factor), linear and quadratic terms for item size, site, and an interaction effect of search pattern and distance from observer, as well as random controlling effects of observer and transect.

The highest ranked model estimated strong positive effects of item size, and that detectability of items with different colors scaled as bright >> white > clear >> dull (Table 1F). Distance effects were more strongly negative for the midline search pattern than for the edge search pattern, suggesting that detection efficacy is largely the same for nearby

items (distance = short) between the two search patterns, but that midline searches ==tend to miss more items as distance increases when compared to side search patterns (Table 1F). Detection rate was estimated to be higher at PTMSC than at Carkeek, equivalent to an absolute difference of 12% (61% to 73%, respectively) for a small item (roughly equivalent to a bottle cap) situated at 1 m from the observer. However, the estimated difference was uncertain, with lower 95% confidence interval bound almost overlapping with zero (Table 1F). The model estimates for random effects of observer id was higher than for transect id among transects, suggesting that inter-observer variability was higher than the effect of location and item arrangement given the same set of items (Table 1F).

Table 1F. Parameters for the highest ranked model based on AICc. The estimated parameter for site (Carkeek), and color (bright) are represented in the model intercept, and alternate levels for these factors are represented as differences from these reference levels.

	Estimated parameters (lin. model)			Estimated parameters (quad model)		
	Mean	SE	95% CI	Mean	SE	95% CI
Fixed Effects						
Intercept	1.46	0.27	0.92, 2.00	0.86	0.33	0.20, 1.52
sqrt_size	0.30	0.03	0.23, 0.37	0.52	0.08	0.36, 0.68
$sqrt\_size^2$				-0.014	0.004	-0.022, -0.006
site = PTMSC	0.60	0.25	0.10, 1.11	0.61	0.25	0.11, 1.12
item color = clear	-1.63	0.19	-2.02, -1.24	-1.71	0.19	-2.09, -1.33
item color = dull	-2.36	0.25	-2.85, -1.87	-2.50	0.26	-3.02, -1.98
item color = white	-1.18	0.19	-1.55, -0.81	-1.29	0.20	-1.69, -0.90
distance [midline]	-0.73	0.11	-0.95, -0.51	-0.75	0.11	-0.97, -0.53
distance [edge]	-0.45	0.05	-0.55, -0.35	-0.46	0.05	-0.56, -0.36
Random Effects						
	SD			SD		
vol id	0.43			0.44		
transect id	0.20			0.20		
	Resid. dev. = 1796 (1894 resid. df.)			Resid. dev. = 1787 (1893 resid. df.)		

The resulting detection curves illustrate that at short distances, detectability is largely the same for edge and midline search patterns. These curves use two examples: a small white and a large white item. For small items (roughly equivalent to a bottle cap) situated 1 meter from the observer this difference amounts to 5% (edge = 73.5%, CI = 66 – 80%, midline = 68%, CI = 59 – 75%), increasing to 13% at 2 meters (edge = 64%, CI = 54 – 73%, midline = 51%, CI = 42 – 60%) (Figure 1E). The difference for larger items (~ size of a beverage can) is minimal at 1 m, but amounts to a 4% difference at 2 meters (edge = 93%, CI = 91 – 95%, midline = 89%, CI = 85 – 92%) (Figure 1E). Variation in detection was higher for a smaller item than for a larger one, and larger items were more reliably detected overall (Table 1G).

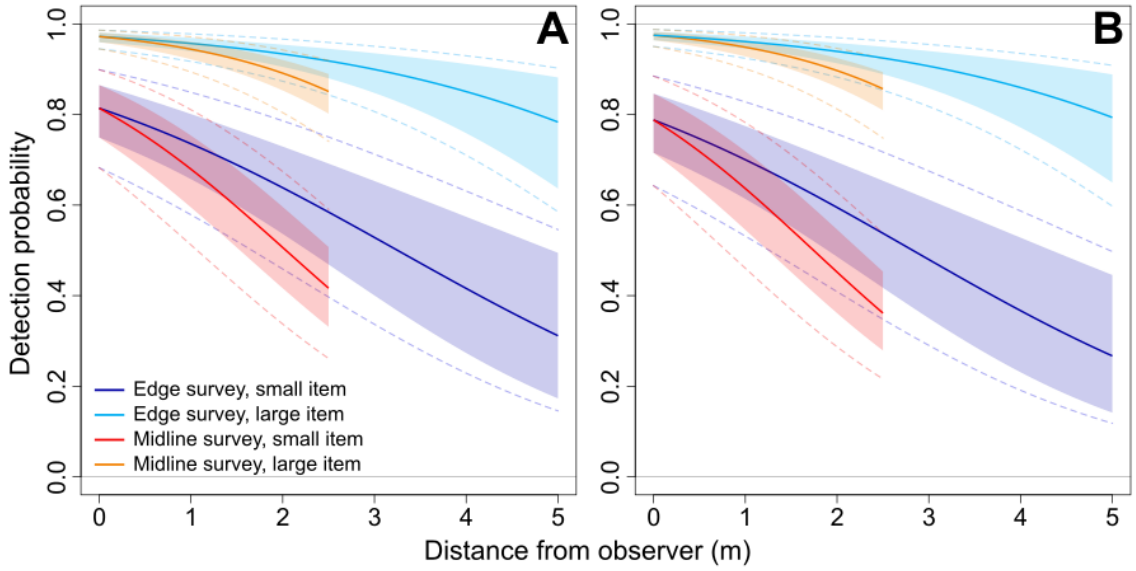


Figure 1E. Detection (Burnham et al. 1980; Buckland et al. 2005) as a function of distance from observer for edge and midline search patterns resulting from linear (A) and quadratic (B) model fits. Items modelled are white with maximal surface areas of 9cm^2 (small) and 100cm^2 (large). Shaded areas indicate 95% confidence interval of the mean, whereas the dashed lines show a 95% prediction interval that includes inter-observer variability.

Table 1G. Fitted probabilities and 95% CI for example items of different sizes, colors, distance from observer and search patterns.

Search pattern	Distance from observer (m)	√Item Area	Bright	White	Clear	Dull
Edge	1	3	90 [87, 93]	74 [66, 80]	64 [57, 70]	46 [39, 53]
Edge	1	10	99 [98, 99]	96 [94, 97]	93 [91, 95]	87 [83, 90]
Midline	1	3	87 [83, 91]	68 [59, 75]	57 [50, 64]	39 [32, 47]
Midline	1	10	98 [97, 99]	94 [92, 96]	91 [89, 94]	84 [79, 88]
Edge	2	3	85 [80, 89]	64 [54, 73]	53 [44, 61]	35 [28, 43]
Edge	2	10	98 [97, 99]	93 [91, 95]	90 [86, 93]	81 [75, 86]
Midline	2	3	77 [71, 82]	51 [42, 60]	39 [33, 46]	24 [19, 30]
Midline	2	10	96 [95, 98]	89 [85, 92]	84 [79, 88]	72 [65, 78]

Detection rate models including quadratic terms for item size modelled lower detection rates for smaller items and a more rapid increase in detection as debris size increases compared to models containing only a linear component for item size. This model reaches an asymptote at 100 cm^2 (10×10cm) suggesting items above this size are almost always detected. Below 25cm^2 (5×5cm) there is a sharp decrease in detection rate suggesting items smaller than this are most prone to not being detected (Figures 1F).

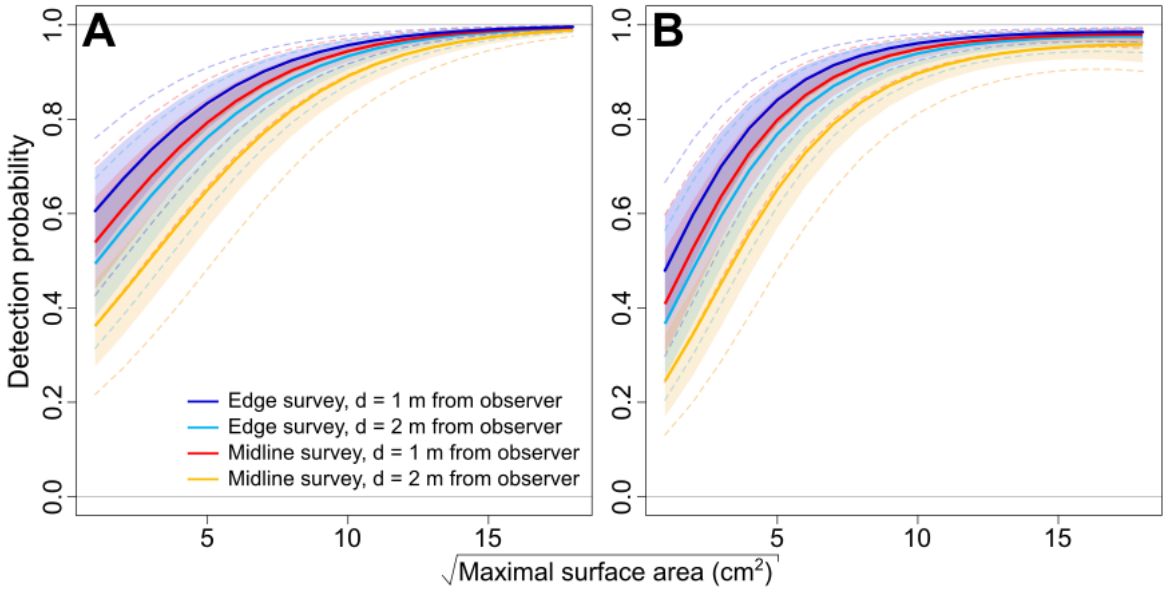


Figure 1F. Detection curves as a function of object size for edge and midline surveys resulting rom linear (A) and quadratic (B) model fits. Items modelled are white with maximal surface areas ranging from 1cm² to 20cm² situated at 1 & 2 m from the observer. Shaded areas indicate 95% confidence interval of the mean, whereas the dashed lines show a 95% prediction interval that includes inter-observer variability.

There is an interaction between size and visual distance from observer. The detection probability for larger items (i.e. ≥10 cm long) is close to 1 regardless of distance, whereas items smaller than 10 cm are somewhat less likely to be detected at a distance greater than 2.5 meters (Figure 1G, A). For even smaller items, 5 cm or less, visual distance from observer accounts for a detection probability range of .95-.80 for items near the observer, to .70-.30 for items as far as 5 meters away. to .30 at a distance of 5 meters away (Figure 1G, A).

This interaction is also affected by search pattern. When searching in one direction up to 2.5m, the smallest items (2.5 cm) have a detection probability of .70 compared to .40 when searching in both directions during midline surveys (Figure 1G). The effect of search pattern on detection rate within a visual distance of 2.5m of the observer was further examined directly using a paired bootstrap routine. In each bootstrap permutation, 34 (minimum sample size) edge surveys were selected at random and each paired with a midline survey from the same transect. The difference in detection rate, Δ = Edge-Midline, was calculated for each of the 34 pairs, and the mean, $\underline{\Delta}$, calculated across paired surveys. Repeating this procedure ($N_{permutations}$ = 1000) enabled us to build a distribution for $\underline{\Delta}$, which we used to infer the bootstrap mean and 95% confidence interval of the difference in detection rate, while controlling for differences in detectability among transect. This analysis revealed that on average, detection rates were 5% higher (95% CI = -1 to 11%) for edge surveys (edge: detection rate = 81%, 95% CI = 76-86%) than for midline surveys (midline: detection rate = 76%, 95% CI = 72-80%) with the caveat that midline searches covered twice the area.

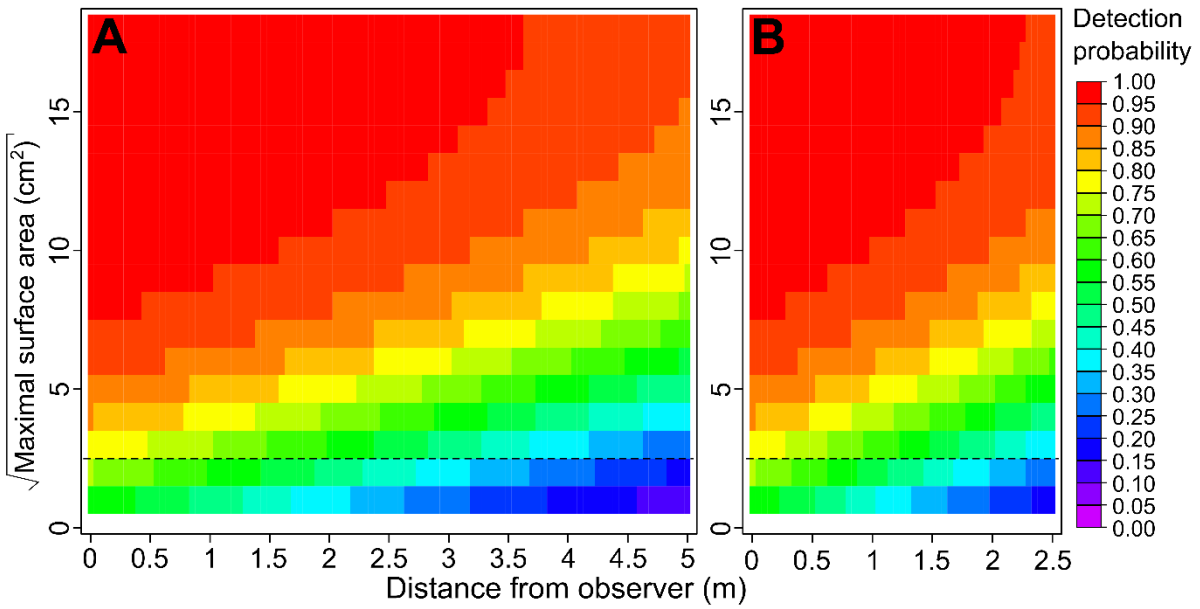


Figure 1G. Size by distance heatmap for edge (A) and midline (B) surveys. Larger items are more often detected than smaller items, especially as visual distance from item increases. The impact of size is greater when attention is split looking side to side during a midline search (B) versus an edge only search (A). The black dotted line delineates 2.5cm items, the smallest that were tested in these field trials. Values below this line are extrapolated from the model.

Further modifying our models by considering beach zone and substrate (albeit on a reduced dataset; beach zone and substrate was only collected on a subset of transects, see Methods), both of these predictors included in the highest ranked models, and had high Akaike weight, indicating they were important in determining detection rate (Table 1H). The inclusion of these variables resulted in the omission of site and transect id (random) factors from the model when compared to the best model identified using all data as they likely explained the same variability (i.e. substrate types among sites, and item locations within transect). However, the highest ranked model retained item color and size terms, distance from observer and observer random effects (Table 1H).

Table 1H. Model selection table including item beach zone and substrate information. All models with $\Delta_{\mathrm{AICc}} < 4$ are shown. Model-specific Akaike weights, $\omega_{\mathrm{AICc}} = \mathrm{e}^{-\Delta_{\mathrm{AICc}}/2}$, are given, as are the summed Akaike weights for model parameters, Σ_{AICc}, calculated as the sum of ω_{AICc} for models containing the corresponding parameter. Predictors noted as (1|volunteer) refer to random effects of the stated grouping variable.

#	Model Equation	AICc	Δ_{AICc}	ω_{AICc}
1	Item_color + Item_size + Item_size^2 + Distance + Zone + Substrate + (1|Volunteer)	1442.2	0	0.36
2	Site + Item_color + Item_size + Item_size^2 + Distance + Zone + Substrate + (1|Volunteer)	1442.4	0.19	0.32
3	Item_color + Item_size + Item_size^2 + Distance + Zone + Substrate + (1|volunteer) + (1|transect)	1443.5	1.35	0.18
4	Site + Item_color + Item_size + Item_size^2 + Distance + Zone + Substrate + (1|volunteer) + (1|transect)	1444.3	2.10	0.13

	Item				Search	Location		Random	
	Site	Color	Size	Size^2	Distance	Beach zone	Substrate	Observer	Transect
Σ_{AICc}	0.45	1	1	0.99	1	1	0.99	0.99	0.31

The highest ranked model resulted in similar parameters to the previous model for item size, color and distance effects (Table 1I). The fitted model suggests that detection rate was higher on sand than on gravel, and that detection rate for beach zones followed bare > wrack > wood >> vegetation (Table 1l). Fitted probabilities indicate that detection is similar across bare, wrack and wood zones, but is much lower in the vegetation zone, likely due to objects being obscured by vegetation, and that cobble can obscure the detection of items by up to 20% for small (surface area = 3 cm^2; roughly equivalent to a bottle cap) items (Table 1J).

Table 1I. Parameters for the highest ranked model based on AICc evaluated on all data, and for a subset of data including terms of item zone and substrate. For models containing the corresponding effects of site, item color, beach substrate and zone, the estimated parameter for site (Carkeek), item color (bright), substrate (cobble) and zone (bare) are represented in the model intercept, and alternate levels for these factors are represented as differences from these reference levels.

	Models without substrate or zone			Models including substrate and zone		
	Mean	SE	95% CI	Mean	SE	95% CI
Fixed Effects						
Intercept	0.86	0.33	0.21, 1.51	0.67	0.41	-0.15, 1.49
sqrt_size	0.52	0.08	0.36, 0.68	0.52	0.08	0.36, 0.68
sqrt_size^2	-0.014	0.004	-0.02,-0.006	-0.017	0.005	-0.027, -0.007
site = PTMSC	0.61	0.25	0.11, 1.11			
item color = clear	-1.71	0.19	-2.08, -1.34	-1.47	0.23	-1.93, -1.01
item color = dull	-2.50	0.26	-3.02, -1.98	-2.02	0.28	-2.58, -1.46
item color = white	-1.29	0.20	-1.69, -0.89	-0.71	0.24	-1.19, -0.23
distance [midline]	-0.75	0.11	-0.97, -0.53	-0.82	0.13	-1.08, -0.56
distance [edge]	-0.46	0.05	-0.58, -0.34	-0.48	0.05	-0.58, -0.38
substrate = sand				1.01	0.22	0.57, 1.45
zone = veg				-1.89	0.30	-2.49, -1.29
zone = wood				-0.35	0.22	-0.79, 0.09
zone = wrack				-0.21	0.24	-0.69, 0.27
Random Effects						
	SD			SD		
vol id	0.44			0.43		
transect id	0.20					
	Resid. dev. = 1787 (1893 resid. df.)			Resid. dev. = 1416 (1551 resid. df.)		

Table 1J. Fitted probabilities and 95% CI for items of different item locations, sizes (item = white), and distance from observer for edge surveys only.

Distance from Observer (m)	√Item Area	Substrate	Bare	Wrack	Wood	Veg
1	3	cobble	77 [68 ,83]	73 [63 ,81]	70 [62 ,76]	33 [24 ,43]
1	3	sand	90 [85 ,94]	88 [81 ,93]	86 [80 ,91]	57 [46 ,68]
1	10	cobble	96 [95 ,98]	95 [93 ,97]	95 [93 ,96]	80 [72 ,86]
1	10	sand	99 [98 ,99]	98 [97 ,99]	98 [97 ,99]	92 [88 ,94]
2	3	cobble	67 [56 ,77]	62 [50 ,73]	59 [48 ,68]	23 [16 ,33]
2	3	sand	85 [76 ,91]	82 [71 ,89]	80 [69 ,87]	46 [33 ,58]
2	10	cobble	94 [91 ,96]	93 [90 ,95]	92 [89 ,94]	71 [61 ,80]
2	10	sand	98 [96 ,99]	97 [96 ,98]	97 [95 ,98]	87 [81 ,91]

Post-hoc simulations to theoretically examine differential search effort

We used the best fit model (fitted to all data; i.e. excluding zone and substrate) to examine the theoretical differences in detection rate among either edge (left or right), midline, left + right (edges combined), and left + right + midline (all search patterns combined). We performed 1,000 permutations of detection rate for the following scenario. One hundred marine debris items were selected at random from the pool of items used in field trials to generate a set of debris characteristics. Each item of debris selected was then assigned a visual distance from the observer based on a uniform distribution between 0 and 5 meters. We then used the model to make predictions for whether each item was detected according to a left-edge survey (observer walks along the 0m line), a midline survey (observer walks along the 2.5m line) and a right-edge survey (observer walks along the 5m line) with distances assigned accordingly (see Figure 1C). For each of the simulated surveys we then totaled the number of detections for each of the individual surveys, and also aggregated counts to simulate all edge searches combined and all searches combined. This was calculated on an item-by-item basis, whereby if any item was detected in any one of the two (left+right) or three (left+right+midline) surveys of that simulated survey arrangement it was tallied into the total count. Across 1000 permutations we were then able to calculate the mean and range of detection rate for each of the 5 simulated search patterns.

Based on these simulations, midline surveys detected an average of 4% more items than left or right, which matches the 3% we calculated based on the raw data. Combining left and right-edge surveys within a transect resulted in an increase in detection rate from x% to 92%, representing an 18% increase compared to midline surveys alone. Simulation of adding a third observer (all search patterns combined) only resulted in a 5% increase in detection rate, but approached 100% for some arrangements (Table 1K).

Table 1K. Simulated detection rates for different individual and combined search patterns and levels of effort.

	Detection rate			
Simulated search pattern	Mean	Median	Min	Max
Left-edge	73.8	74	65	82
Right-edge	73.8	74	66	82
Midline	77.5	78	70	85
Left+Right	91.8	92	86	97
Left+Right+Midline	97	97	94	100

Question 2. What is the effect of number of observers per transect on marine debris counts?

Methods

We conducted a second series of controlled field trials to more directly examine the effect of survey team size. The field trial set up was designed to demonstrate how alternate effort and search pattern affect detection rates under procedures that are closer to true surveys. As in the previous field trials, we established three to four (depending on staffing) 5m wide transects seeded with the same kits of 20 representative marine debris items described above. Instead of an even distribution, as required in the earlier field trials, seeded items approximated random distribution through haphazard placement of items across the transect to approximate true survey conditions. And, unlike the prior field trials, volunteers were permitted to move within the transect to view and record the serial numbers of marine debris as they searched. Entry of volunteers into the transect represents real world survey protocols, whereas the previous field trials included a COASST staff member to control for proximity to debris.

These field trials were carried out at three locations on four dates (Table 2A), and on each occasion, participants were asked to survey according to one of four search patterns (Figure 2A);

- **1O**: **One observer** searches alone walking down the midline of the transect and records all information
- **2O**: **Two observers** search together walking down the edges (opposite one another) of the transect while one records all information
- **2O+1R**: **Two observers and one data recorder**: two observers search walking down the edges of the transect as in B, and a third person stands outside the transect, not contributing to the search but recording information called out by the others
- **3O**: **Three observers** where two search walking down the edges of the transect and a third searches down midline, one of which records information.

Table 2A. Summary of observer and survey data for variable survey effort study

Location	Date	Number of observers	# Surveys	**1O**	**2O**	**2O+1R**	**3O**
PTMSC	23 Jun 2019	22	23	8	6	5	4
Ocean City	2 Nov 2019	35	48	14	14	9	11
Carkeek Park	20 Jul 2019	21	27	8	8	5	6
Carkeek Park	12 Oct 2019	14	16	6	5	2	3

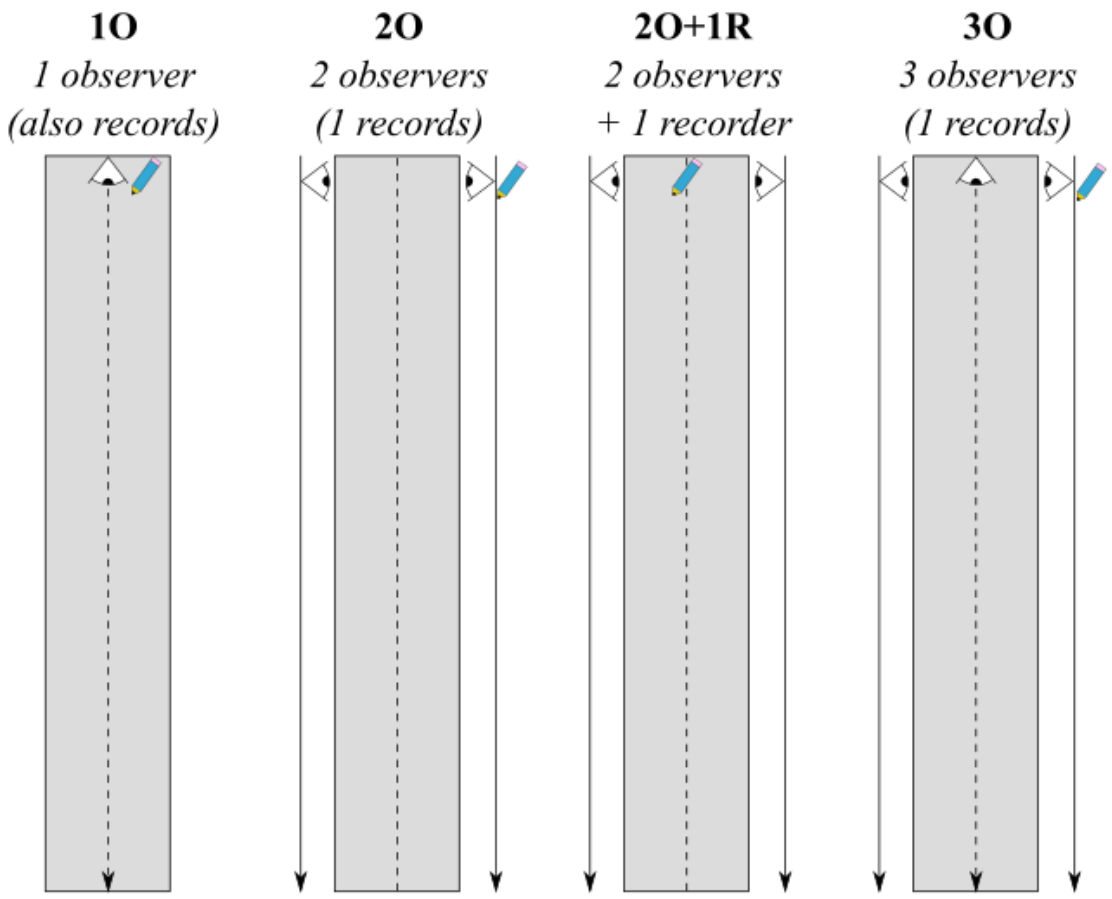


Figure 2A. Field trial transect effort types rom left to right.

Our goal was to have each transect surveyed at least once for each effort type, and for individuals to participate in different effort type surveys at each of the transects they visited (Table 2B).

Table 2B. Summary of surveys completed during each field trial event. Each transect was seeded with 20 debris items and surveyed at least once using each effort type.

			Surveys by effort type and transect			
Site	Date	Tran.	1O	2O	2O+1R	3O
Carkeek	Jul-2019	1	2	2	1	2
Carkeek	Jul-2019	2	2	2	2	1
Carkeek	Jul-2019	3	2	2	1	1
Carkeek	Jul-2019	4	2	2	1	2
Carkeek	Oct-2019	1	1	1	1	2
Carkeek	Oct-2019	2	2	2	0	1
Carkeek	Oct-2019	3	3	2	1	0
Ocean City	Nov-2019	1	3	3	3	3
Ocean City	Nov-2019	2	3	3	3	2
Ocean City	Nov-2019	3	5	4	1	3
Ocean City	Nov-2019	4	3	4	2	3
PTMSC	June-2019	1	2	1	1	1
PTMSC	June-2019	2	3	1	2	1
PTMSC	June-2019	3	1	2	1	1
PTMSC	June-2019	4	2	2	1	1
	TOTAL	15	36	33	21	24

Analytic Approach

We previously fitted detection curves established during the initial field trials to make predictions about detection rates within a transect for different numbers of observers and to identify how much overlap (i.e. items detected by multiple observers) there is. However, actual field methods, where participants move within the transect to collect items, may influence the overall detection rate. Therefore, this second approach replicates three different levels of effort (1, 2 or 3 observers) and examines how detection rates increase with number of participants, and in turn, how reported marine debris density may vary with search effort.

As a summary of surveys we calculated the gross-average detection rates for each of the e effort types (1O, 2O, 2O+1R, 3O; Figure 2A) and its bootstrapped 95% confidence interval, and also calculated inter- and intra- effort-type differences. We calculated all possible pairwise differences for effort search type performed at that transect (Table 2C), such that pairwise differences represented differences among effort types and/or surveyors given the same arrangement and type of items and did not represent differences among transects. For alternate effort types (i.e. the difference between 1O and 2O) we calculated the difference as 2O-1O, whereas for the same effort type we calculated the absolute difference among surveys at the same transect. Therefore, among effort-type differences represent how much the detection rate differs among effort types, whereas within effort-type provides a measure variability given the specific type of survey. The bootstrap mean and

95% CI of inter- and intra- pairwise differences were calculated by generating 1000 sets of n observations, sampled randomly with replacement, where n is the number of observed pairwise differences. For each bootstrap sample we calculated the mean pairwise difference, and calculated the overall mean and 95% CI of the mean as the mean and 95% range of the resultant distribution of sample means.

Results

Gross averages of detection rate for each effort-type increased with increasing participant number, with a single observer (1O) recording an average detection rate of 80% compared to 85% for two observers (2O), and 88-89% for three observers (2O+1R, 3O) (Table 2C). Pairwise differences in detection rate (i.e. controlling for transect effects) for 3O type surveys were 3-4% higher than 2O/2O+1R types, and 9% higher than 1O type surveys (Table 2C). Differences between 2O and 2O+1R type surveys were small, and the 95% CI of the mean difference overlapped with zero, suggesting that there is no effect on detection rate of adding an independent data recorder (Table 2C). Within effort-types, the largest differences were evident for 1O survey types, where the inter-observer variability was ±12% on average (Table 2C). Increasing the number of observers to two reduced variability to ±7%, with no significant difference between 2O and 2O+1R type surveys (Table 2C). Adding a third observer reduced this variability to ±3.5% on average (Table 2C), suggesting that increasing observer effort increases the overall detection rate, and reduces the level of inter-observer (or in the case of 2 or more observers, inter-team) variability. Results were broadly similar when looking at detection of small items only (< 25 cm^2), but detection rates were lower by ~ 10% for single and double-observer surveys, but by 6% for three-observer surveys (Table 2C). For small items pairwise differences were also larger comparing three-observer to solo surveys, but smaller and less consistent between pairs and singles (Table 2C), suggesting that while the addition of another observer increases detection rate of small items it is not guaranteed. The addition of a third observer had a marked effect on the detection of small items over singles and pairs, resulting in consistently higher detection rates (Table 2C). Inter-team variability was also higher for smaller items, but showed the same general pattern of increasing consistency (decreasing differences among teams) with increasing personnel. The exception was inter-team differences for 1O+1R type surveys, which was particularly high for smaller items (Table 2C). The sample size for 1O+1R type surveys was considerably lower, and inter-team differences were largely driven by comparisons at Ocean City, a field trial distinct in that (similar to Damon Point in the previous set of field trials), observers were comprised of undergraduate students on a field trip (Table 2C: inter-team differences could only be calculated on transects with two or more teams).

Table 2C. Gross averages, and pairwise differences among and within survey types. All values were calculated via bootstrap resampling. Sample sizes (N) represent the number of surveys (gross averages), or the number of pairwise comparisons (pairwise differences) for that comparison.

Survey effort type	N	All items	Small items (< 25 cm2)
Gross average detection rate (%)			
1O	36	80.2 [77, 83.75]	70.5 [64.6, 75.6]
2O	33	85.1 [81.4, 88.2]	73.7 [68.3, 78.8]
2O+1R	20	88.5 [85.5, 91.5]	78.3 [72.0, 84.7]
3O	24	88.8 [86.0, 91.3]	82.6 [77.5, 87.4]
Average pairwise differences in detection rate (% difference)			
2O - 1O	88	5.1 [1.3, 8.6]	2.3 [-2.4, 8.8]
2O+1R - 1O	54	6.3 [2.4, 10.6]	5.5 [-2.0, 13.4]
3O - 1O	63	8.9 [5.2, 13]	13.5 [7.6, 19.8]
2O+1R - 2O	50	-0.1 [-3.2, 3.2]	1.3 [-4.9, 7.5]
3O - 2O	61	4.0 [0.8, 7.3]	10.2 [5.7, 15.2]
3O - 2O+1R	38	2.5 [-0.8, 6.1]	7.1 [1.1, 13.2]
Average among team differences in detection rate (% absolute difference)			
1O	32	12.1 [9.2, 15.5]	16.1 [11.3, 20.9]
2O	26	6.9 [3.5, 11.2]	9.7 [5.5, 15.1]
2O+1R	9	7.2 [3.3, 11.1]	18.2 [14.9, 22.2]
3O	13	3.5 [1.2, 6.5]	4.3 [1.5, 7.5]

Question 3. If a surveyor is short on time, should they complete a reduced number of full-width transects or decrease the width of the transect while maintaining 4 replicates?

Methods

During the field trials used to examine Question 2, an additional area was set up in a location with debris confirmed to be present (not seeded). We chose not to seed debris because knowing the rate of detection was unnecessary and we wished to approximate real-world survey conditions and density estimates. We set up a 25m wide plot divided into ten 2.5m wide transects (equivalent to 5 × 5m wide transects searched from each edge) delineated by lines of flags. Each observer visiting this plot searched all 10 transects by walking along the right edge of each transect, searching up to 2.5m away toward the next transect (line of flags), tallying the number of items found in each transect (but not removing or disturbing them). We chose this search pattern based on the results of Question 2 (i.e. an edge survey rather than a midline survey searching in both directions), which indicated that a single person has a higher detection probability at a visual distance of 2.5m when searching only in one direction compared to both directions, especially for items at the smaller end of the range tested.

Compiled data from the 25m wide plot was used to compare debris densities observed under different sampling effort scenarios, specifically determining the best number and widths of transects to survey to ensure that the data are representative of the entire sampling area.

Table 3A. Summary statistics of the first set of field trials where item detection is only for items within 2.5m of the observer. Rows representing totals show the median across site×transect combinations, except for the bootstrap, which is the bootstrap 95% CI across surveys for that search pattern.

				Detection					
Site	Tran.	Search	N_{survey}	Mean	Median	Min	Max	sd	95% CI
Carkeek	1	Edge-L	3	96	100	88	100	7	[88-100]
Carkeek	1	Midline	9	81	85	65	95	10	[74-87]
Carkeek	1	Edge-R	4	82	79	79	93	7	[79-89]
Carkeek	2	Edge-L	6	85	83	78	100	9	[80-91]
Carkeek	2	Midline	6	72	70	60	85	8	[66-78]
Carkeek	2	Edge-R	5	71	73	45	91	17	[58-84]
Carkeek	3	Edge-L	7	86	91	64	100	14	[77-95]
Carkeek	3	Midline	3	75	70	70	85	9	[70-85]
Carkeek	3	Edge-R	8	54	56	33	67	13	[46-61]
PTMSC	1	Edge-L	4	89	95	64	100	17	[73-100]
PTMSC	1	Midline	6	84	80	80	95	7	[80-89]
PTMSC	1	Edge-R	4	89	91	73	100	14	[77-100]
PTMSC	2	Edge-L	6	85	86	64	100	12	[76-94]
PTMSC	2	Midline	4	79	85	50	95	20	[60-93]
PTMSC	2	Edge-R	5	96	100	78	100	10	[87-100]
PTMSC	3	Edge-L	6	86	91	73	100	11	[79-94]
PTMSC	3	Midline	6	72	70	50	85	13	[63-80]
PTMSC	3	Edge-R	4	78	78	78	78	0	[78-78]
		Edge-L	32	86	91	68	100	12	
Total (median)		Midline	34	77	78	65	92	12	
		Edge-R	30	80	90	79	96	8	

Analytic Approach

We compared three scenarios related to standard versus reduced effort surveys.

a. Standard practice: Four, 5m-wide transects accomplished by one participant walking each edge.
b. Reduced effort 1: Fewer transects, maintaining transect width of 5m
c. Reduced effort 2: Narrower (2.5m) transects, maintaining transect number (four).

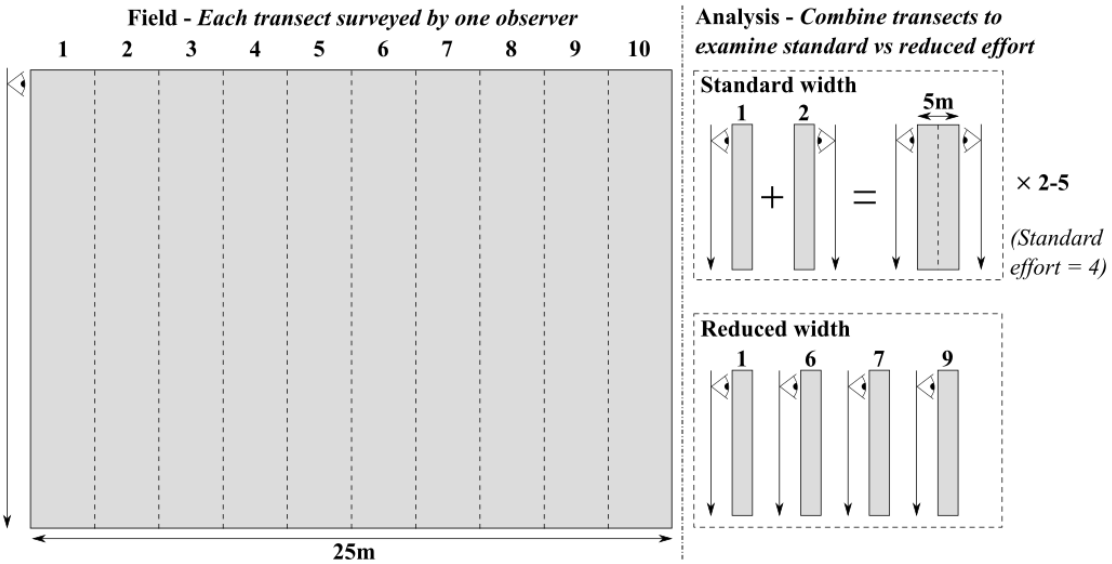


Figure 3A. Sampling setup and analysis schematic.

To examine how comparable reduced effort schemes are to complete sampling we performed surveys of 25m plots, delineated into 10, 2.5m-wide transects. Within each transect participants were asked to record all debris encountered. As transects were adjacent to one another, this allowed us the opportunity to pool data across adjacent transects to replicate a 5m wide transect surveyed from either side.

This data was analyzed by calculating the proportional difference between marine debris density estimates resulting from reduced effort for alternate transect sizes (1-9 2.5m transects, or 1-4 5m transects). For each participant at each transect we calculated their reference density, $D(i)$ equal to the summed count across all of their transects divided by the total transect area

$$D(i) = \frac{\sum_{j=1}^{10} \boldsymbol{d_j(i)}}{A}$$

where *i* denotes observer, $d_{\cdot}(i)$ is the vector of transect counts made by observer *i*, and A is the total area A = 10×2.5×*w*, where *w* is beach width in meters. Estimates of density resulting from reduced effort were calculated by randomly sampling transects within all possible volunteer by transect combinations and calculating the resultant mean density. This ensures that the difference in density is representative of reduced effort alone, rather than including differences among sites and/or volunteers if we had selected transects from among the total pool of data. For a reduced effort scheme defined by number of transects,

t, we selected *t* values from a specific observer's set of transects, $d_{\cdot}(i)$, and calculated an estimated density, $\widehat{D(i)}$ as

$$\widehat{D(i)} = \frac{\sum_{j=1}^{10} I_j \boldsymbol{d_j(i)}}{2.5tw}$$

where I_j is a binary vector equal to one if that transect was chosen and zero otherwise. We then calculated the relative difference in density, Δ, as the absolute difference between the 'true' density for that observer, $D(i)$, and the density resulting from reduced effort, $\widehat{D(i)}$, as a proportion of the 'true' density

$$\Delta = \frac{\left|\widehat{D(i)} - D(i)\right|}{D(i)}$$

To generate a distribution of values we performed 1000 permutations for each level of sampling effort, with each permutation selecting a transect and observer at random, as well as choosing a random selection of transects for that level of sampling effort. This was equivalently performed for narrow (*t* = 1-9) and wide (*t* = 1-4) transects. We then calculated the mean, median and 95% range of Δ across permutations.

Results

For narrow transects (i.e., 2.5m), performing a single transect resulted in density estimates that differed from the 'true' density by ± 50% on average but could be as high as ± 160%. Increasing sample size to 5 narrow transects resulted in an average relative difference of less than ± 20%, but could be as high as ± 50%. Eight or more narrow transects would be required to achieve relative differences of < 10% on average, although even at this sample size, differences in density could be as high as ± 30%. For wide (i.e., 5m) transects, performing 2 transects resulted in density differences of ±20% on average, but could be as high as ±60%. Four or more wide transects would be required to be within ± 10% (Table 3B, Figure 3B).

Table 3B. Relative differences in reported density for alternate sample sizes and transect dimensions. Differences in density were calculated as the absolute difference between reported density for reduced effort, and the "true" density, calculated as the observed density across all 10 transects, controlling for (i.e. selecting within) observer and transect. Presented means, medians and 95% range values were calculated via bootstrap resampling across observers, transects, and alternate samples of transects.

# of Transects	Transect Width	Relative difference in density (%)		
		Mean	median	95% range
1	narrow (2.5m)	54	43	[3, 159]
2		36	31	[0, 108]
3		27	23	[1, 76]
4		22	19	[0, 62]
5		18	15	[0, 50]
6		15	13	[0, 41]
7		12	11	[1, 34]
8		9	8	[0, 27]
9		6	5	[0, 19]
1	wide (5m)	38	32	[0, 108]
2		23	20	[0, 62]
3		16	13	[1, 41]
4		9	8	[0, 27]

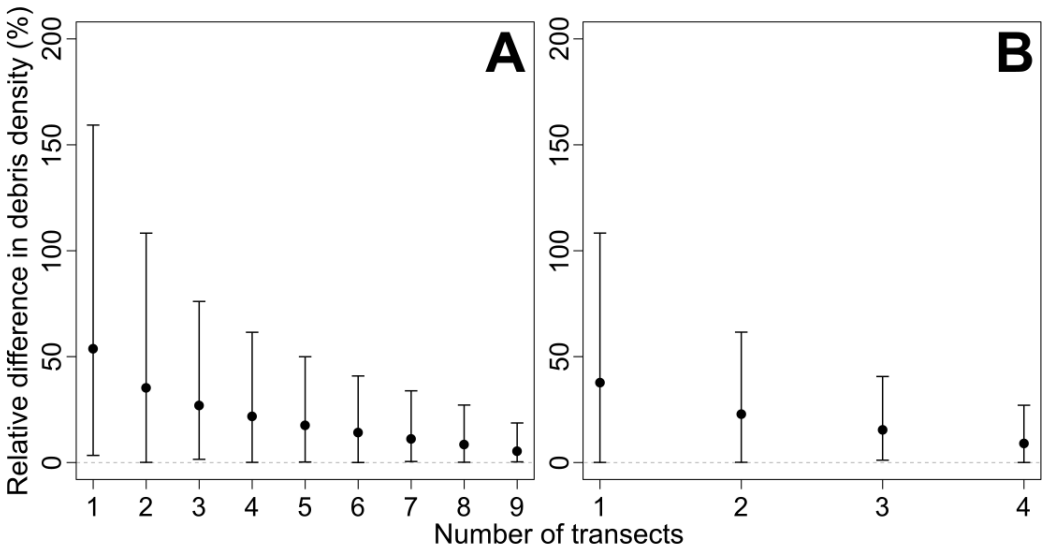


Figure 3B. Relative difference in debris density as a function of number of transects for narrow (2.5m; A) and wide (5m; B) transect widths. Values represent the mean and 95% range of relative differences.

Inter-observer variability

As we had multiple observers survey the same transect, this affords us the opportunity to examine between observer differences in density under *in-situ* debris densities and arrangements. For each transect, we calculated the density recorded by each observer and plotted histograms of the distribution of densities, calculating the mean and standard deviation of observer-density estimates, and also comparing these estimates to the theoretical maximum, calculated as the sum of the max count per transect for that transect.

This enables us to calculate a pseudo detection rate (i.e. assuming the theoretical maximum included all items) and how much this detection varies among individuals.

Recorded density varied considerably among individuals ranging by ±40-50% compared to the mean (Table 3C). Coefficients of variation were approximately equal across transects at ~ 0.3, suggesting that among observer variability was approximately constant (Table 3C). Framed relative to the theoretical maximum for that transect, detection rate on average varied between 45-55% for each transect, with the exception of Ocean City where detection rate was ~ 10% lower (Table 3C). The range of rates among individuals varied from ~25% up to 70-75%, with the exception of one observer at Carkeek in July who detected 81% (Table 3C).

Table 3C. Summary of marine debris concentration observed at the four experimental field trials. Concentration values are calculated across observer-specific concentrations, whereas the transect observed maximum was calculated as the mean density calculated across the max count observed for each flag line within that transect. Pseudo-detection rates are also given as the median and range in concentration relative to the observed maximum.

			Concentration (per 100m2)							Median
Transect	N	Length (m)	mean	med.	min	max	sd	CV	Obs. max	detection rate [min, max]
PTMSC	11	28.1	2.6	2.6	1.4	3.7	0.7	0.26	5.41	47 [26, 68]
Carkeek-Jul	11	21.8	7.8	8.1	2.8	11.7	2.7	0.34	14.50	56 [19, 81]
Carkeek-Oct	11	14.5	6.0	6.3	3.6	9.1	1.9	0.32	12.41	51 [29, 73]
Ocean City	15	44.3	2.6	2.3	1.6	4.2	0.9	0.32	6.05	37 [27, 69]

Question 4. Does removing debris during standing stock surveys affect future counts/load estimates?

Methods

COASST established a single 100m plot at each of 5 beaches (see study sites above) chosen to encompass natural variation in substrate and exposure to visitation, tides and currents. Each plot was assigned four sets of two paired 5m wide transects. Within each pair, one transect was assigned as either “debris removed” or “debris left in place” (Figure 4A). Locations and the assignment of each transect within a pair (debris removed or not) were marked with wooden stakes and remained constant throughout the study.

Surveys were conducted monthly from June 2018 to November 2019 by staff with the assistance of undergraduate interns who remained constant throughout the project. During each survey, we counted the number of items greater than 2.5cm found in each transect and beach zone, removing or leaving items according the fixed assignment of the transect.

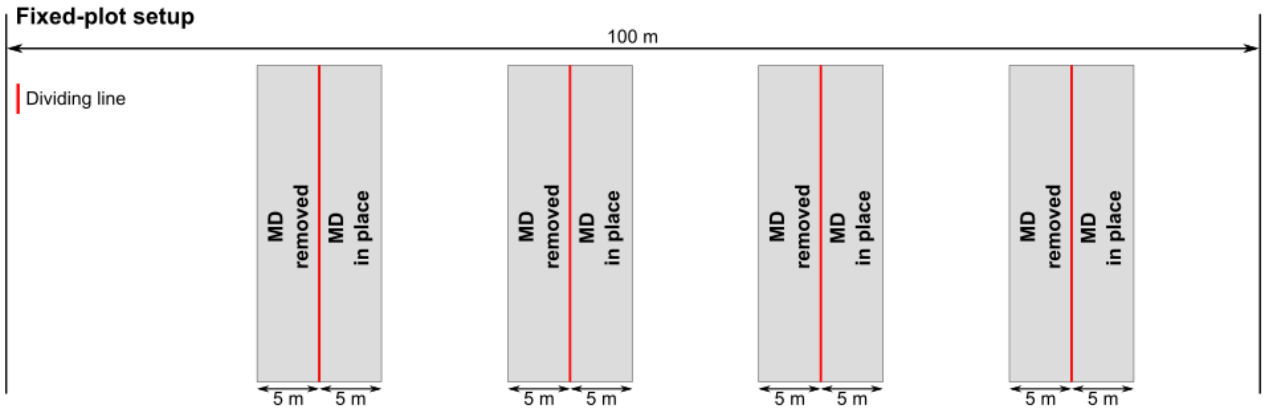


Figure 4A. Four replicates of paired 5m transects were established and surveyed ~monthly on five beaches.

Analytic Approach

Monthly surveys conducted over eighteen months at each plot provided two contrasting time-series of marine debris density; (1) debris counted and then removed on a monthly basis and (2) debris counted and left in place. If debris density is sufficiently dynamic over a month period (i.e. the amount of debris entering and/or exiting the transects through other forces is greater than the treatment effect of removal) then there would be no discernible difference in reported density between treatments. Conversely, we would expect that the act of removing debris should result in lower debris density for the transect with that treatment compared to the one where debris was left in place to “accumulate”. By testing the difference in mean debris density between treatments we were able to identify how the act of debris removal affects marine debris density during subsequent surveys. To identify whether debris removal had a significant effect on debris load, we ran a series of GLM models including treatment (two levels: debris removed vs left in place) as a predictor, and examined whether treatment effect was significant. Our models considered factors of

treatment (left in place, removed), site, the interaction between site and treatment, and the interaction between season (summer: May – Oct, winter: Nov – Apr) and site, collectively modelling spatiotemporal variation in counts. We also included random effects of survey occasion and transect pair to model among survey event variability, synchronous across transects within site and occasion, and small scale variability (i.e. adjacent transects) that was consistent across survey occasions, respectively (Table 4A). As the response was a count of debris items found, we used generalized linear mixed effects models assuming data were distributed according to a negative binomial distribution (glmer.nb in lme4 package R). The assumption of negative binomial data accounts for possible overdispersion of data. Because transects differed in length, we also included an offset term of log(area), where area = transect width*length. With the exception of the offset term, all possible combinations of remaining model effects were trialed and the resulting models were compared and ranked based on AICc. After identifying the best model, we tested whether the inclusion of treatment type, or the interaction between treatment type and beach resulted in a significant improvement of the model based on a likelihood ratio test.

Table 4A. Factorial predictors considered in model fitting of removal treatments.

Name	Type	N_{levels}	Levels/range	Fixed/random
beach	factor	5	Carkeek, Point Wilson, Damon Point, Point no Point, Kalaloch	fixed
treatment	factor	2	left in place, removed	fixed
season	factor	2	summer, winter	fixed
survey	factor	64		random
transect pair	factor	20		random

Results

Average debris densities were variable among survey dates, but appeared to vary similarly among transects with debris left in place and transects with debris removed (Figure 4B). Carkeek, an urban park within Puget Sound, showed seasonal variation, with summer surveys (June – September) showing higher debris density than other times. Point no Point and Point Wilson sub-urban sites in Puget Sound and the Strait of Juan de Fuca, displayed little to no variation among treatments, and with the exception of the September survey at Point Wilson, showed little variation among survey time points (Figure 4B).

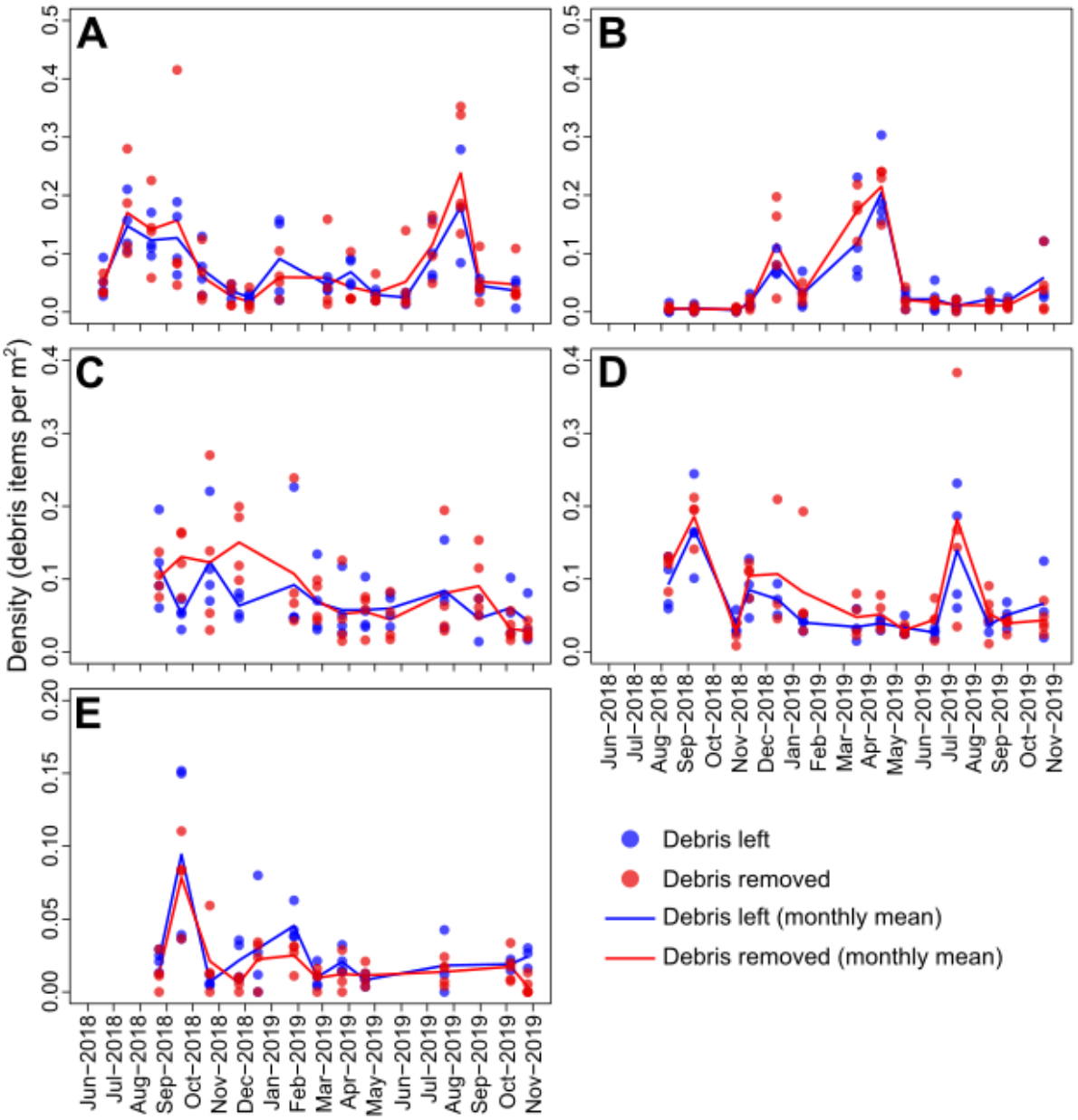


Figure4B. Time series of debris density at Carkeek Park (A), Kalaloch (B), Point No Point (C), Damon Point (D) and Point Wilson (E) for debris removed and left in place treatments. Points represent average density per transect for surf into vegetation (surf to wood was very similar) and lines represent the average across transects on each survey occasion.

At outer coast sites, debris density was highly variable among surveys at Kalaloch but more consistent at Damon Point (Figure 4B). At Kalaloch, from winter to late spring (December – May) surveys had higher density than other months, whereas Damon Point peaked in September 2018 and again in July 2019, but outside of these dates was largely the same (Figure 4B).

Among GLMs fitted to debris count data, the best model based on AICc included season and beach as an interaction effect, as well as random effects of survey and transect pair. The addition of treatment and beach interacting with treatment resulted in increased AICc

values (Table 4B), and were not significant based on likelihood ratio tests (Table 4B). The most informative predictor was the interaction between season and beach, suggesting that different beaches have different seasonality and different overall marine debris densities

Table 4B. Model statistics, and likelihood ratio test results when adding treatment effects to the best-fitting models, and models consisting of random effects only. Likelihood ratio tests provide a means of examining whether the addition of model predictors to a given null model are supported by the data, based on the reduction in model deviance.

Mod #	Model	AICc	Likelihood ratio tests			
			Null model	$\Delta_{deviance}$	df	p
0	count ~ rand(survey) + rand(transect)	3510.1				
1	count ~ beach:season + rand(survey) + rand(transect)	3488.7	Mod 0	40.1	9	7.0×10^{-6}
2	count ~ treatment + rand(survey) + rand(transect)	3511.8	Mod 0	0.18	1	0.67
3	count ~ treatment + beach:season + rand(survey) + rand(transect)	3489.8	Mod 1	0.16	1	0.69
4	count ~ treatment:beach + beach:season + rand(survey) + rand(transect)	3492.2	Mod 1	5.7	5	0.34

We then repeated this analysis considering counts for the lower beach (surf, wrack and bare zones), wood and vegetation zones to see if removal had a differential effect in different parts of the beach (i.e. would retention/turnover differ in places where debris could get "stuck"?). The results were the same for the lower beach and wood zones, with only beach-specific seasonal effects being important, and the addition of treatment effects were not supported based on likelihood ratio tests. For the vegetation zone the only supported predictor was beach, suggesting that seasonal effects are not as important in the high beach compared to the lower portions of the beach. However, similar to all other zones, the addition of removal treatment effects was not supported based on likelihood ratio tests (Table 4C).

Table 4C. Model statistics, and likelihood ratio test results when adding treatment effects to the best-fitting models, and models consisting of random effects only, when fitted to vegetation zone counts only. Likelihood ratio tests provide a means of examining whether the addition of model predictors to a given null model are supported by the data, based on the reduction in model deviance.

Mod #	Model	AICc	Likelihood ratio tests			
			Null model	$\Delta_{deviance}$	df	p
0	count ~ rand(survey) + rand(transect)	1998.6				
1	count ~ beach + rand(survey) + rand(transect)	1966.4	Mod 0	38.1	3	2.0×10^{-8}
2	count ~ treatment + rand(survey) + rand(transect)	1999.9	Mod 0	0.68	1	0.4
3	count ~ treatment + beach + rand(survey) + rand(transect)	1967.8	Mod 1	0.6	1	0.44
4	count ~ treatment: beach + rand(survey) + rand(transect)	1973.3	Mod 1	1.07	4	0.9

Question 5. Does including a portion of the back barrier in the transect influence the number of debris items found?

Methods

During monthly surveys of 100m plots established for assessing the effect of debris removal (Question 4), special attention was given to the back barrier vegetation area to assess:

1. The density of debris and overall contribution of this zone to transect loads
2. The effect of survey distance into back barrier vegetation (i.e. is there a distance at which marine debris density precipitously drops?)

While surveying the vegetation zone, we measured the distance of items from the edge of the vegetation, continuing into the vegetation up to 50 meters, or until no items were found in a 5m section, or a natural or engineered barrier was encountered (whichever came first). The location of each debris item was recorded from the beach facing edge of the vegetation zone to enable assessments of debris density as a function of distance into vegetation zone (Figure 5A)

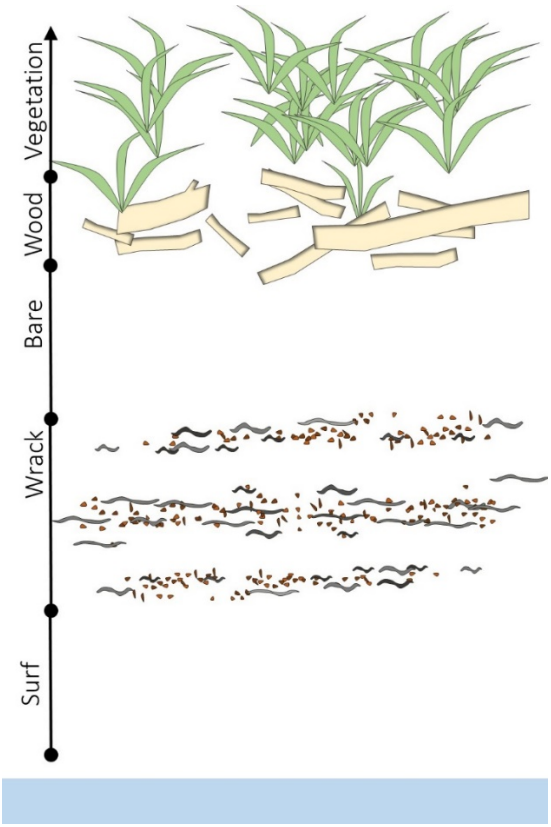


Figure 5A. Schematic of typical zones of a beach. Surf is the expanse of beach that shrinks and grows with the tide cycle, and runs from the water's edge to the first tide-line where floating debris (wrack) has been deposited. The wrack zone encompasses all visible deposits of wrack from fresh to older/higher lines. If recent high tides have not reached into the back beach, there may also be a "bare" area until the upper, sometimes present wood zone is found – where large driftwood is deposited by extreme high tides and storm surges. Finally, many beaches are backed by a vegetated area.

Analytic Approach

In order to evaluate how the inclusion of the vegetated back barrier affects density estimates, we used our data to simulate a set of scenarios. For each survey we calculated debris density assuming only the lower beach (everything but the vegetation) was surveyed, and then secondarily the density if transects extended different distances into the back barrier. We then examined how the proportional difference between the lower beach density and the lower beach + vegetation density changed as a function of distance into the vegetation for that transect. We then calculated the mean and 95% CI of the mean by bootstrap resampling the calculated differences for each distance into the vegetation. This was performed separately for each site, as well as across sites, except for Kalaloch which lacks a vegetated back barrier and backs onto a cliff.

Results

Across different beach zones, debris density was highest in the wood and vegetation zones, with the exception of Point Wilson where debris density was highest in the wrack zone (Table 5A). Point Wilson had overall lower debris densities than at other locations. At the broadest level, inclusion of the vegetation zone increased debris densities by 10-30%.

Table 5A. Average zone widths and absolute densities within beach zones at each of the four transects.

Zone	Site				
	Point Wilson	Point No Point	Carkeek Beach	Damon Point	Kalaloch
Width (m)					
Surf	4	15.4	9.2	13.3	38.3
Wrack	17.3	29.4	13.7	19.7	32.5
Bare	7.6	6.7	3.2	3.9	2.5
Wood	3.9	12.5	19.5	11.4	12.74
Veg	7.18	6.3	4.7	33.7	0
Surf to Wood	32.8	64	45.6	48.2	86
Surf to Veg	40	70.2	50.3	81.98	86.04
Density (per 100m^2)					
Surf	0.36	0.14	1.22	0.08	0.32
Wrack	5.54	3.62	4.06	4.62	4.14
Bare	2.34	0.44	6.18	5.02	1.14
Wood	3.02	22.26	10.78	16.26	15.1
Veg	2.26	24.16	26.56	8.38	
Surf to Wood	2.4	5.8	6.1	6.4	4.6
Surf to Veg	2.3	7.7	7.9	7.2	

Incorporating portions of the back barrier at Carkeek and Point no Point consistently elevated the recorded density above that recorded for the low beach (Figure 5B). This reached an asymptote in both cases, corresponding to the depth of the vegetation zone at

each site. This shows that vegetation at these two locations had a consistently higher density of objects in the vegetation than on the low beach. At Point Wilson there was no significant change in relative density, but changes in absolute density increased when incorporating 2-3m of the back barrier, but then decreased, suggesting that at Point Wilson shoreward facing vegetation had higher debris density than the low beach, but deeper into the vegetation had lower debris loads (Figure 5B). At Damon Point, relative changes in debris density were positive, suggesting higher debris loads in the vegetation than the low beach, but comparatively these changes were smaller than at Carkeek and Point no Point (Figure 5B).

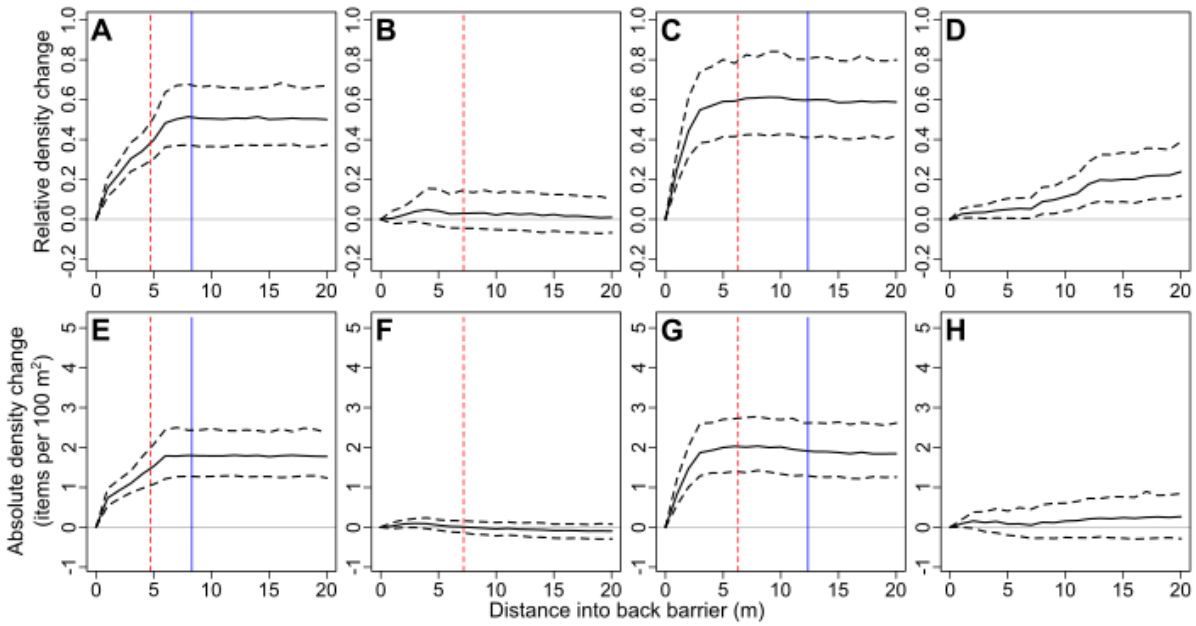


Figure 5B. Change in vegetation density as a function of distance. The top row (A-D) displays relative density (proportional change) whereas the bottom row (E-H) displays absolute change in density as observed at Carkeek Park (A & E), Point Wilson (B & F), Point no Point (C & G) and Damon Point (D & H). The red line represents the mean vegetation width at that site and the blue line represents the maximum recorded vegetation width (distance of an item into the vegetation) at that site. Lines for Damon Point are not represented as they extend beyond twenty meters.

Differences between overall density of marine debris in the vegetation zone at each of the four sites are dramatic, with Carkeek and Point no Point having higher loads than the other two sites (Figure 5C). We noted that vegetation types backing the plots at Carkeek and Point no Point both consist of dense shrubs including blackberry thickets, whereas Point Wilson and Damon Point are largely comprised of dune grass and smaller plants.

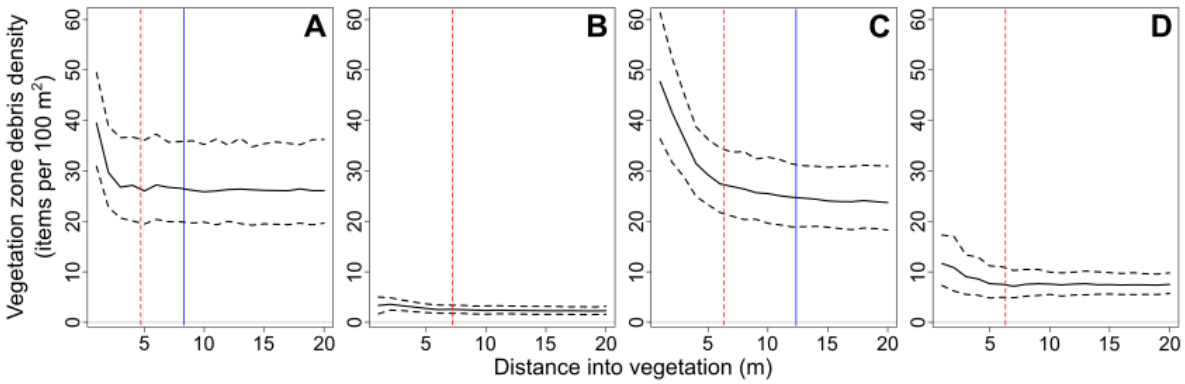


Figure 5C. Marine debris density in the vegetation zone as a function of distance searched into vegetation zone at Carkeek Park (A), Point Wilson (B), Point no Point (C) and Damon Point (D). The red line represents the mean vegetation width at that site and the blue line represents the maximum recorded vegetation width (distance of an item into the vegetation) at that site.

Discussion

Data quality is a topic that has been widely explored in the literature on citizen science and is a known barrier to achieving science-oriented outcomes of these projects. However, with intentional design, evaluation, and adaptive management, many potential pitfalls can be avoided (Burgess et al. 2017, Parrish et al. 2018). This study elucidates opportunities for design, protocol and training improvements, as well as considerations for data analysis and interpretation in shoreline marine debris monitoring by citizen scientists.

In general, best practices for citizen science project management hinge on intended data use, as well as the particularities of data collected by a known corps of volunteers. If shoreline marine debris data are to be used to compare patterns over time or between sites where actual densities as well as influences of detection rate (the proportion of items seen by volunteers relative to known presence) will vary, it is important to limit or control influences on detection rate as best as possible. Touch points worthy of consideration include sampling design, survey protocols, training, calibration, built-in quality checks and controls, team size and tasks, and achievable sample sizes (Parrish et al. 2018).

Of specific importance to marine debris monitoring, we found that detection rate varied among volunteers, with distance from volunteers, with characteristics of the debris items, and with the context of the search (i.e. presence of visual obstacles and contrast). Angelini et al. (2019) conducted laboratory experiments to explore detection rates of marine debris under controlled conditions. Although they did not examine the function of distance focused on smaller debris items (less than 2.5cm), and a narrower set of colors, they found that white and clear plastics were under counted (50% and 55% detection, respectively) relative to blue plastic (95%), and their results point to the importance of debris item contrast to substrate and presence of “visual noise” as factors that influence detection.

We found that variability (among observers) increases and detection decreases for a single observer searching beyond a distance of 2.5 meters, and this is especially true for smaller (between 2.5 and 5cm), less bright items. So if we want to maintain the 2.5cm minimum debris size for surveys, we should limit our individual search distance to within 2.5m. Because detection is imperfect and varies between people, there will be benefit to reducing variability between surveys and sites by recommending equal effort (e.g. some people might be inclined to enter and search a transect on their hands and knees back and forth and would find more items than someone sticking to the edge and walking a single pass), so recommending a single search type and potentially even pace (Bjorn et al. 2019, Banning et al 2011) is important.

While we know of no other studies examining real-world detection rates for marine debris, there is extensive literature on search theory and optimization with applications ranging from search and rescue missions to mineral exploration to archaeology (Frost 1999, Drew 1967, Hey 2001). The latter field is likely of most direct relevance to our context and where similar experimental studies have been undertaken to document detection rates during surveys of plowed fields seeded with artifacts. In such studies, detectability varied with distance, artifact characteristics, walking speed and whether a search was conducted toward or away from the sun (Banning et al. 2016, 2011), where the latter was only statistically important for highly reflective items like glass and blue-transfer ware.

Using both simulated (Table 1K) and actual surveys (during the second set of field trials) we found that two observers is better than one and that three is better than two. The result that two is better than one is generally consistent with literature on quality assurance in other tasks and contexts, from TSA baggage checks to military operations (Wiener 1964, Garcia et al 2011). Keeping team size even across sites, and personnel consistent within a site, will benefit data comparability and the likelihood of detecting true differences in debris densities across time and space. We did not gather demographic information or test any impact that observer characteristics may have on debris detection but recognize these could be important predictors for inter-observer variability (Angelini et al 2019).

Intentional biases notwithstanding (that is, we assume that volunteers do not choose to ignore debris or gather and count debris from outside the sampling area), debris counts can be considered minimums. We demonstrated that estimates of debris density are conservative, and that the extent of under estimation varies with size and color of items, as well as the visual “background” of the search. For example, larger bright items are detected close to 100% of the time, so density estimates of those items are more precise than smaller dull items that may be missed half the time. And detection rates in cobble are lower than on sand. These findings are consistent with other research on visual detection including the study of evolved camouflage in prey species - where size, pattern, color and contrast are of

import (Karpestam et al 2017). Where such factors effecting detection are known and quantified, they should be factored into analyses (e.g. Johnson et al 2019).

When interpreting and analyzing marine debris data from shoreline surveys, we should consider what these data are measuring and represent within a system of sources, sinks, and transport of marine debris items. There are several potential immediate inputs of manmade objects found on site: deposition from tides and storm surges, items left behind (intentionally or not) by beachgoers on site and items brought in or revealed by wind. At the same time, many of these drivers also remove debris from beaches: tides and storm surges can wipe the beach clean, beachgoers perform clean-ups, wind moves items off the beach or out to sea or covers them with sand. Given these factors, it is not surprising that removal of debris within a 5m wide transect does not impact densities in that location measured one month later.

It is conceivable that both transects in a pair where debris was removed from one and not the other during a survey, are repeatedly cleared of debris from a combination of forces (manual removal by beachgoers, movement and burial by wind, storm surges and high tides) such that there is limited to no accumulation on average, month to month. Also, movement of debris across the length of the beach due to these same forces, may mean that the influence of the untreated surrounding beach could overwhelm the effect of removal on a narrow transect. It is possible that over a larger area removal would impact future debris loads, especially toward the middle of such an area. This study focused on items 2.5cm or larger, but did not examine the potential relationship between item size and persistence. Larger, less mobile items may persist longer and contribute to accumulation. It is also possible that over a shorter interval, the impact would be noticeable, but on the scale of a month the dynamics of a beach are too great to measure accumulation.

One result that we found surprising was that the treatment effect (debris removed vs not) was not present anywhere on the beach, including the back barrier (vegetation) or driftwood, where debris might be expected to get stuck barring manual removal or extreme tidal energy and which have been posited as debris sinks (Olivelli et al. 2020). Generally, this part of the beach contained debris loads that contributed to (increased) the overall density of debris within a transect, but there was no indication of accumulation over time. Notably, vegetation was associated with the lowest relative detection rate across the profile of the beach, where plants likely present visual obstacles that obscure marine debris or visually compete for attention. In turn, we recommend a consistent approach to sampling this area, and to sample it separately from the “active beach” so that low detection rates can be taken into account during analysis. And, for data to be comparable across sites and times, the back barrier should be characterized and surveyed at a distance that is consistently achievable. Including some portion of the back barrier will capture those items that have blown upward

(lighter things) or pushed upward by extreme high tides as well as litter that could form a source of debris that will be transported to the ocean.

Compliance

The protocol for this study was reviewed by the University of Washington Human Subjects Division IRB, receiving activity approval under FW #00006878. Required documentation of consent was waived.

Staff participating in this project completed NOAA Contractor Sexual Assault and Sexual Harassment Prevention and Response Training on 10/2/2018.

Research permits were obtained and maintained from the National Park Service (OLYM-2018-SCI0058) and Washington State Parks (180503) for activities at Kalaloch, Ocean City, Damon Point and Fort Worden. The city of Seattle (Carkeek Park) and Kitsap County (Point no Point) were both contacted to seek permission.

Acknowledgements

This work was conducted in partnership with the NOAA Marine Debris Program, and we thank them for their input and feedback. Several undergraduate students and staff at the University of Washington were instrumental in data collection, including Abby Bratt, Ellie Davis, Allison DeKerlegand, Yurong He, Colin Piwtorak and Charlie Wright. Nir Barnea enthusiastically attended several field trials and took photos. The Olympic Coast National Marine Sanctuary, Port Townsend Marine Science Center, Olympic National Park and Washington State Parks graciously helped with participant recruitment and fixed plot establishment. We would especially like to thank the many members of the public who volunteered their time to help with this project.

References

Angelini Z., Kinner N., Thibault J., Ramsey P., Fuld K. (2019). "Marine debris visual identification assessment." *Marine Pollution Bulletin*. 142: 69-75. doi: 10.1016/j.marpolbul.2019.02.044.

Bates D., Mächler M., Bolker B., Walker S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

Banning E.B., Hawkins A.L., S.T. (2011). "Sweep widths and the detection of artifacts in archaeological survey.*" Journal of Archaeological Science*. 38 (12): 3447-3458. doi: 10.1016/j.jas.2011.08.007.

Banning, E., Hawkins, A., & Stewart, S. (2006). Detection Functions for Archaeological Survey. *American Antiquity, 71*(4), 723-742. doi:10.2307/40035886

Burgess H.K., DeBey L.B., Froehlich H.E., Schmidt N., Theobald E.J., Ettinger A.K., HilleRisLambers J., Tewksbury J., Parrish J.K. (2017). "The science of citizen science: Exploring barriers to use as a primary research tool." *Biological Conservation*. 208: 113-120. doi: 10.1016/j.biocon.2016.05.014.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L. (2005). *Distance sampling*. John Wiley & Sons, Ltd.

Burnham K.P., Anderson D.R., Laake J.L. (1980). "Estimation of density from line transect sampling of biological populations." *Wildlife monographs*. 1 (72): 3-202.

Drew, L.J., 1979. Pattern drilling exploration: optimum pattern types and hole spacings when searching for elliptical shaped targets. Math. Geol. 11, 223e254.

Frost, J.R., 1999a. Principles of Search Theory, Part I: Detection. Soza and Company, Ltd., Fairfax, VA.

Garcia, A., Baldwin, C., Funke, M., Funke, G., Knott, B., Finomore, V., & Warm, J. (2011). "Team Vigilance: The Effects of Co-Action on Workload in Vigilance." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. *55* (1): 1185–1189. doi: 10.1177/1071181311551247

Hardesty, B.D., Wilcox, C., Schuyler, Q., Lawson, T.J., and Opie, K. (2017b) Developing a baseline estimate of amounts, types, sources and distribution of coastal litter - an analysis of US marine debris data (CSIRO: EP167399), Final Report to the NOAA Marine Debris Program in fulfillment of NOAA Award Number NA15NOS4630201. https://marinedebris.noaa.gov/sites/default/files/publications-files/An_analysis_of_marine_debris_in_the_US_FINAL_REP.pdf

Hey, G., Lacey, M., 2001. Evaluation of Archaeological Decision-making Processes and Sampling Strategies. Kent County Council and Oxford Archaeological Unit, Oxford.

Isaac N.J.B., Michael J. O. Pocock M.J.O. 2015. "Bias and information in biological records." *Biological Journal of the Linnean Society*. 115 (3): 522–531, https://doi.org/10.1111/bij.12532

Johnston A., Hochachka W.M., Strimas-Mackey M.E., Ruiz Gutierrez V., Robinson O.J., Miller E.T., Auer t., Kelling S.T., Fink D. "Best practices for making reliable inferences from citizen science data: case study using eBird to estimate species distributions" bioRxiv 574392; doi: 10.1101/574392

Johnston, A., Fink D., Hochachka W. M., Kelling S. (2018). Accounting for observer expertise improves ecological inference from citizen science data. *Methods in Ecology and Evolution* **9**: 88– 97.

Karpestam E., Merilaita S., Forsman A. (2018). "Size variability effects on visual detection are influenced by colour pattern and perceived size." *Animal Behaviour*. 143: 131-138. doi: 10.1016/j.anbehav.2018.07.013.

Lardner B., Yackel Adams A.A., Savidge J.A., Reed R.N. (2019). "Optimizing Walking Pace to Maximize Snake Detection Rate: A Visual Encounter Survey Experiment," *Herpetologica*, 75(3): 218-223.

Lippiatt, S., Opfer, S., and Arthur, C. (2013). Marine Debris Monitoring and Assessment. NOAA Technical Memorandum NOS-OR&R-46.

Olivelli, A., Hardesty, B.D., and Wilcox, C. (2020) Coastal margins and backshores represent a major sink for marine debris: insights from a continental-scale analysis. Environ. Res. Lett. in press https://doi.org/10.1088/1748-9326/ab7836

Opfer, S., Arthur, C., and Lippiatt, S. (2012). NOAA marine debris shoreline survey field guide. U.S. Department of Commerce.

Parrish J.K., Burgess H.K., Weltzin J.F., Fortson L., Wiggins A., Simmons B. (2018). “Exposing the Science in Citizen Science: Fitness to Purpose and Intentional Design.” *Integrative and Comparative Biology*. 58 (1): 150–160 doi: 10.1093/icb/icy032

Parrish, J K., Burgess H.K. *Coastal Observation and Seabird Survey Team Protocol.* University of Washington, Seattle, WA. 2017.

R Core Team. (2017). R: “A language and environment for statistical computing. R Foundation for Statistical Computing.” Vienna, Austria. URL https://www.R-project.org/

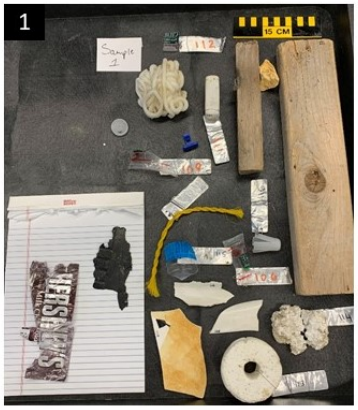
Wiener, E. L. (1964). “The Performance of Multi-Man Monitoring Teams.” *Human Factors*, *6*(2), 179–184. doi: 10.1177/001872086400600207

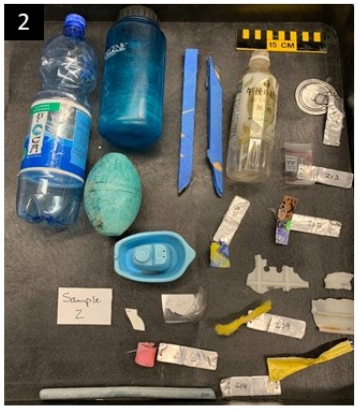
Appendix A. Seeded debris kit contents.

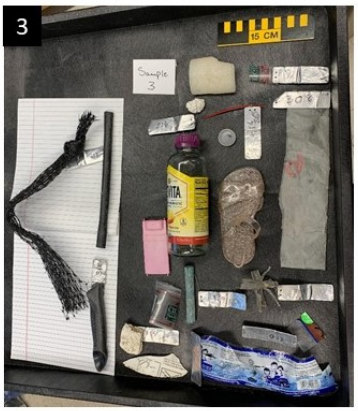
Serial #	Kit ID	Debris Item	Color	Length (cm)	Width (cm)	Height (cm)	Material
101	1	bottle cap	clear	2.7	2.7	0.5	plastic
102	1	unknown fragment	white	8.8	5.8	0.7	plastic
103	1	unknown fragment	white	9.3	4.3	0.3	plastic
104	1	unknown fragment	white	7.5	4.2	0.4	plastic
105	1	rope	yellow	24.5	2	1	plastic
106	1	unknown fragment	multi	7.0	3.4	0.9	plastic
107	1	unknown fragment	white	10.8	8.3	5.6	plastic
108	1	unknown fragment	white	12.7	6.9	0.6	plastic
109	1	unknown fragment	multi	8.8	3.6	0.7	plastic
110	1	car part	blue	2.6	2.6	1.3	plastic
111	1	unknown fragment	white	6.8	2.2	2.2	plastic
112	1	unknown fragment	multi	3.7	2.5	1.2	plastic
113	1	unknown fragment	white	8	8	4.3	plastic
114	1	unknown fragment	white	7.8	7.1	3	plastic
115	1	unknown fragment	multi	4.7	4.6	3.3	plastic
116	1	shotgun wad	white	4	2.9	1.7	plastic
117	1	unknown fragment	black	12.5	6.1	0.1	plastic
118	1	food wrapper	brown	15.5	4.8	0.1	plastic
119	1	lumber	brown	39.4	8.7	3.7	wood
120	1	lumber	grey	21.5	6.3	2.4	wood
201	2	beverage bottle	multi	33.8	8	8	plastic
202	2	Beverage bottle	multi	21.3	6.2	6.2	plastic
203	2	fishing buoy	blue green	12.3	8	8	plastic
204	2	lumber	blue	25	2	1.2	wood
205	2	lumber	blue	25.5	2	1.2	wood
206	2	unknown fragment	multi	4.6	3.9	0.05	rubber
207	2	toy boat	blue	10.2	7.5	6.7	plastic
208	2	aluminum can lid	silver	6.5	6.5	3.5	metal
209	2	rope	yellow	9	1	1	plastic

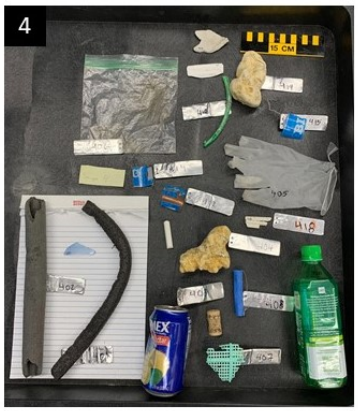
210	2	oyster spacer tube	blue green	23.8	1.3	1	plastic
211	2	unknown fragment	white	8.2	4.5	1.5	wax
212	2	unknown fragment	pink	3.3	2.5	1.5	plastic
213	2	unknown fragment	clear	7	4.5	5	plastic
214	2	food wrapper	multi	12.7	3.9	0.05	plastic
215	2	unknown fragment	white	4.7	3.9	0.3	plastic
216	2	unknown fragment	white	4.8	3	0.3	plastic
217	2	unknown fragment	multi	7.5	2.4	0.05	plastic
218	2	beverage bottle	multi	20.5	9	9	plastic
219	2	unknown fragment	multi	8.6	5.2	4.1	plastic
220	2	unknown fragment	white	4.3	3.1	0.5	plastic
301	3	bottle label	white	27	6.5	0.05	plastic
302	3	unknown fragment	white	7.3	5.8	3	wax
303	3	shoe	clear	6.5	7	6	plastic
304	3	toy bed	pink	9.9	4.5	1.5	plastic
305	3	bottle cap	clear	2.8	2.8	0.8	plastic
306	3	unknown fragment	white	12	4.2	0.3	plastic
307	3	shotgun wad	clear	6.6	6.4	4.1	plastic
308	3	unknown fragment	clear	28.8	7.5	0.1	plastic
309	3	unknown fragment	multi	4.4	3.5	0.05	plastic
310	3	unknown fragment	multi	4.7	4.5	0.05	plastic
311	3	unknown fragment	multi	4.8	2.6	0.5	plastic
312	3	unknown fragment	multi	7.3	4.9	2.5	plastic
313	3	rope	black	50	0.8	0.8	other
314	3	unknown fragment	black	24	1.3	1.7	plastic
315	3	unknown fragment	white	4.9	3.6	1.5	plastic
316	3	unknown fragment	white	3.5	3	2.1	plastic
317	3	bottle cap seal	red	10	0.2	0.1	plastic
318	3	unknown fragment	black	6	2.9	2	plastic
319	3	unknown fragment	blue green	9.4	1.5	1.2	plastic
320	3	beverage bottle	multi	19.3	6.3	6.3	glass
401	4	unknown fragment	black	33	2.1	2.1	plastic
402	4	unknown fragment	gray	28.9	6.7	3.2	plastic
403	4	beverage can	multi	5.5	7.6	12.8	metal
404	4	insulation	yellow	10.9	6.9	5.2	plastic/metal

405	4	glove	clear	23.5	11.1	0.05	plastic
406	4	shopping bag	clear	17.6	16.9	0.05	plastic
407	4	unknown fragment	blue green	6.3	5.9	0.1	plastic
408	4	unknown fragment	blue	7	1.3	1.1	plastic
409	4	unknown fragment	white	4.7	0.9	0.7	plastic
410	4	unknown fragment	green	13.1	0.7	0.7	plastic
411	4	unknown fragment	white	6.7	2	0.7	plastic
412	4	unknown fragment	multi	4	3.8	0.05	plastic
413	4	unknown fragment	blue	3.8	2.8	0.4	glass
414	4	unknown fragment	multi	3.6	3.6	0.05	plastic
415	4	unknown fragment	multi	4.1	3.8	0.05	plastic
416	4	unknown fragment	white	6.5	4.2	0.5	plastic
417	4	lumber	brown	3.6	1.5	1	wood
418	4	unknown fragment	white	4.1	1.7	2	plastic
419	4	insulation	yellow	11.2	7.5	5.5	plastic
420	4	Beverage bottle	green	6.3	5.8	20.7	plastic









Appendix B. Exploratory models of detection rate.

Data cleaning

Item location was calculated for each of the 20 objects deployed across 6 field trial transects (1-3 at Carkeek and PTMSC, Damon Point not used). For each unique item location (n = 120, 6 transects, 20 objects per transect) we calculated measured location for each survey performed at that transect according to distance from Edge-Left. For Edge-Left surveys, measured location was simply the measured distance, whereas for Edge-Right measured location was 500 (i.e. the transect width in cm) minus the measured distance in cm. For midline surveys we looked at the measured locations determined on edge surveys for that object in order to determine whether to calculate measured location for those trials as 250 + measured distance or 250 – measured distance. We then plotted measured location for each transect, and object according to search pattern. For each item, we also calculated an average location as the median location across all observations. The median was used, rather than the mean, because there were several outliers, potentially due to mismeasurement or transcription errors, which would otherwise skew the average location.

Across the 6 field trial transects there was considerable variation in recorded object location (**Figures 1-3**), however 90% of measured locations were within ± 36 cm of the median location (**Figure 4**).

We then used the median location to infer distances to objects that weren't observed on surveys, and then calculated detection rate as a function of distance using the median distance for all observations ensuring that distance was consistent across all observations.

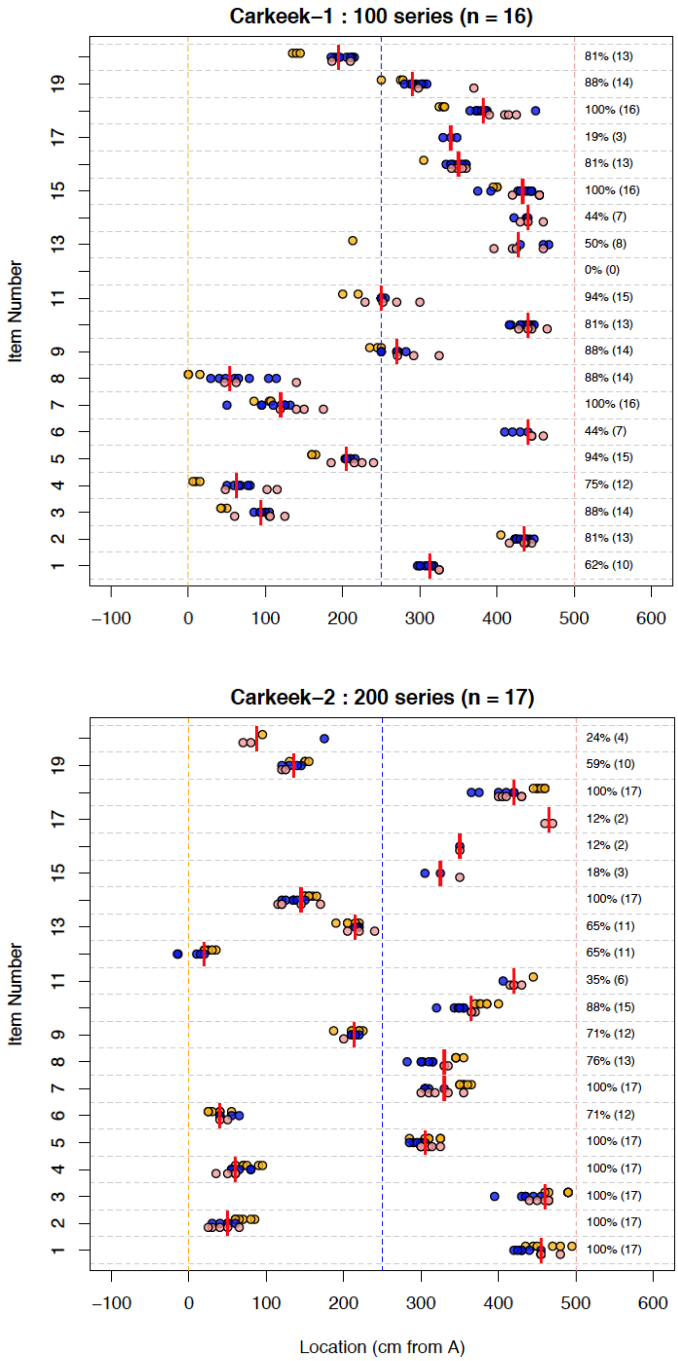


Figure 1. *Locations of debris recorded at Carkeek Park transects 1 and 2 across debris items. Locations are plotted relative to survey Edge-Left (orange dashed line), and points are color coded according to search pattern (orange = Edge-Left, blue = Midline, pink = Edge-Right). Vertical red lines indicate the median location. Detection rate for each item is indicated on the right-hand side.*

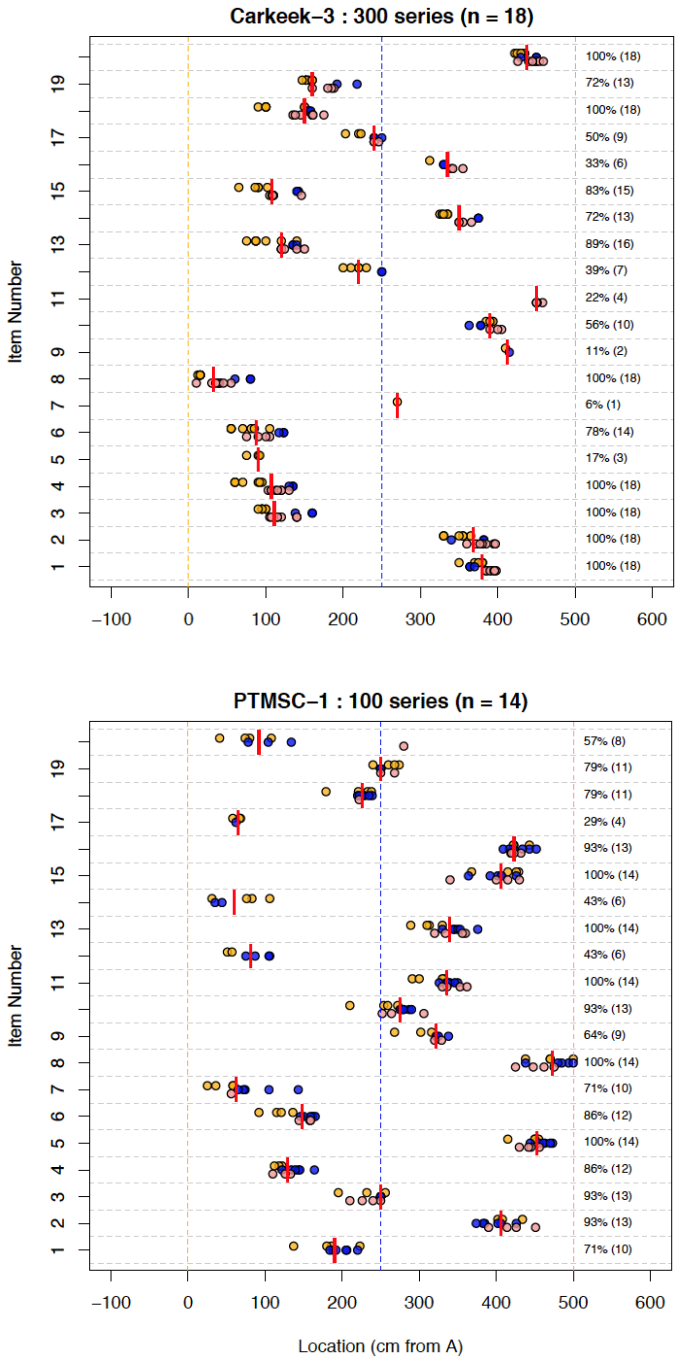


Figure 2. *Locations of debris recorded at Carkeek Park transect 3 and PTMSC transect 1 across debris items. Locations are plotted relative to search pattern Edge-Left (orange dashed line), and points are color coded according to search pattern (orange = Edge-Left, blue = Midline, pink = Edge-Right). Vertical red lines indicate the median location. Detection rate for each item is indicated on the right-hand side.*

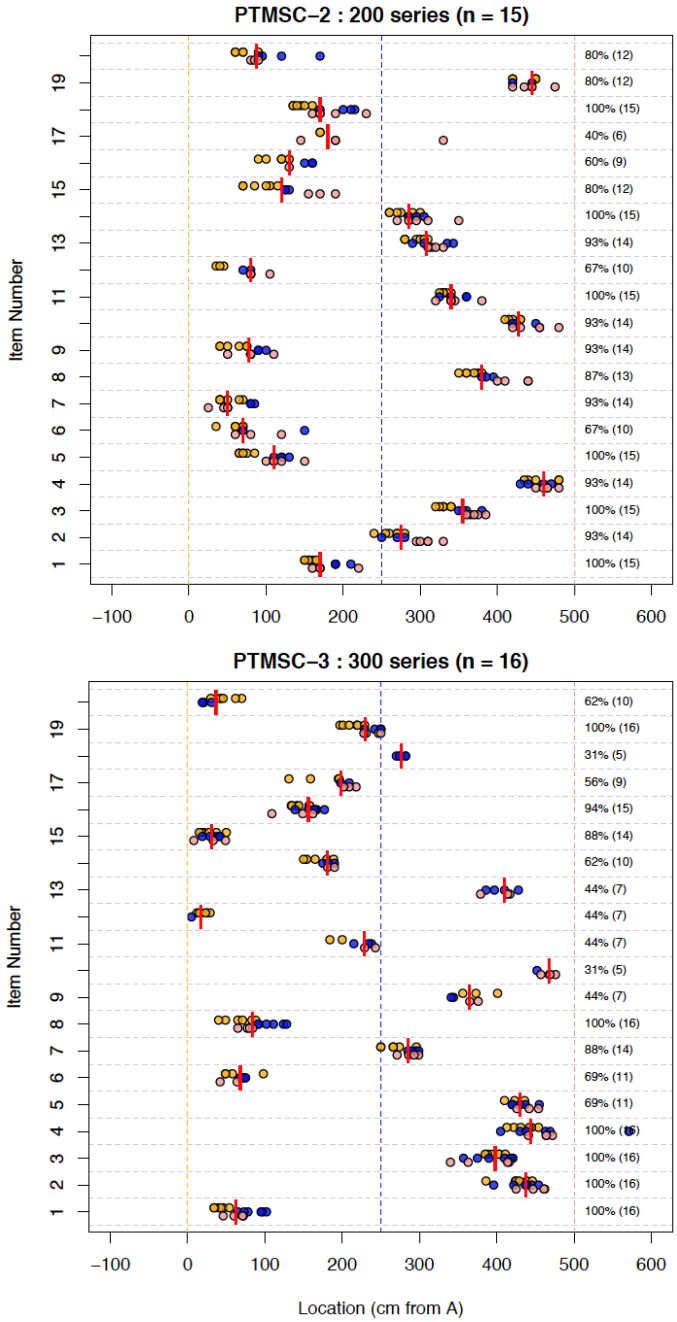


Figure 3. *Locations of debris recorded at PTMSC transects 2 and 3 across debris items. Locations are plotted relative to search pattern (orange dashed line), and points are color coded according to search pattern (orange = Edge-Left, blue = Midline, pink = Edge-Right). Vertical red lines indicate the median location. Detection rate for each item is indicated on the right-hand side.*

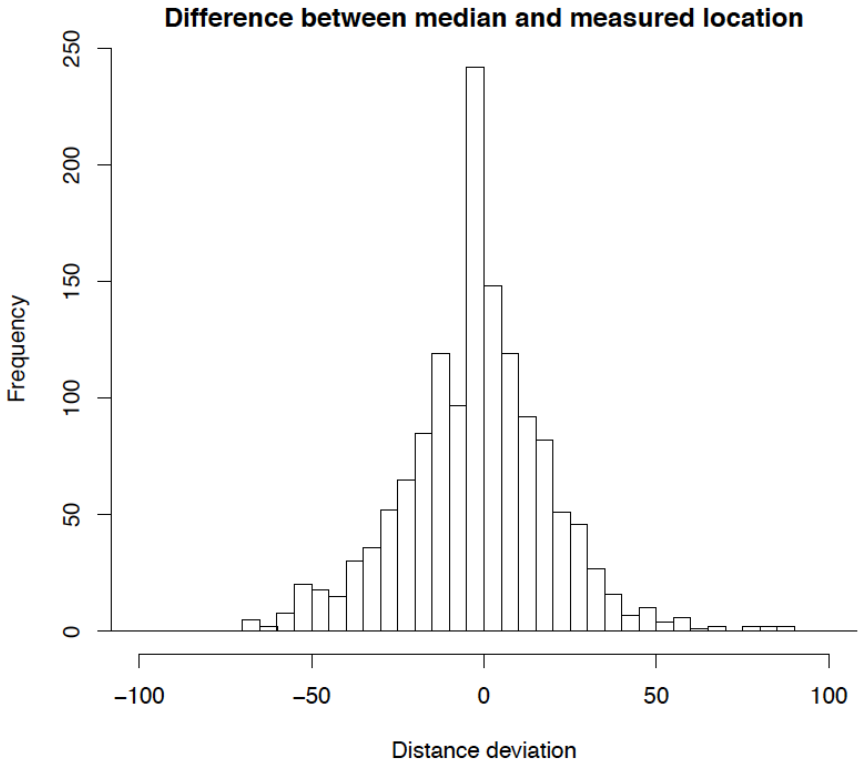


Figure 4. *Difference between measured location (across all passes) and median location. Quantiles: -38cm (5%), -28cm (10%), -17cm (20%), 0cm (50%), 15cm (80%), 25cm (90%), 33cm (95%).*

Initial models of detection rate

To begin examining differences in detection rate we created a series of binomial GLMS using binary detection (1=detected, 0=not detected) as the response and object characteristics (size, color) and distance as predictors.

Our modelling followed a two-stage process. The initial phase was used to identify how to group colors as initially there were too many color levels, and more complex models with random effects didn't converge due to the number of color levels. Although reflectivity may be important, only 3 of the 60 items were recorded as shiny, and so we use color as our primary determinant of object detection.

Our initial models consisted of binomial GLMs with predictors of median distance and object characteristics: color and object area (length * width) as a measure of object size, which was also categorized into small, medium and large (**Table 1**).

Table 1. *Predictors used in the first phase of analysis*

Predictor	Description	Range/Levels
med.dist	Median distance of object from observer	0 – 500 cm
color	Primary object color	clear, white, multi, blue, yellow, silver, red, black, brown
obj.area	Maximal object surface area (length * width)	2 – 347 cm^2
size.class	Categorized size class	Small (< 25 cm^2), medium (25 – 70 cm^2), large (> 70 cm^2)

Because object area and size class are different classifications of the same phenomena (object size) we trialed them against each other in parallel model selections and compared among results to identify which of the two provide the best results. Within the analysis, object area was square-root transformed to provide a more uniform spread or values.

Our modelling approach consisted of fitting a full model with the following structure

$$y_i \sim Binomial(\widehat{p_i})$$
$$\widehat{p_i} = \beta_0 + \beta_D Dist_i + \beta_C Color_i + \beta_S Size_i + \beta_{SD} Size_i \times Dist_i$$

Where β's are model coefficients. The full model includes main effects of distance, color and size, and an interaction term of size and distance, allowing for different intercepts for different colors, and continuous effects of distance, which can be modified according to object size. This allows for different detection rate curves for objects of different sizes.

Using this model as our base, we then evaluated all possible model combinations of these predictors and ranked them based on AICc. This process was performed separately for midline surveys, whereas edge surveys were pooled as detection curves may differ between midline and edge surveys.

The best models for each search pattern, and size classification, included size, color and distance, but didn't include the interaction term of size and distance (**Table 2**). Comparatively, models with object area (sqrt transformed) as a predictor outperformed those with object size class (**Table 2**). The best fitting models for both survey types (A|C vs B) included negative effects of distance and positive effects of object area, indicating that larger objects have a higher detection rate, and more distant objects have a lower detection rate (**Table 2**).

Table 2. *Model selection tables. Model parameters and statistics are given for the top-3 models in each case. Pred. weight is the summed Akaike weights for that predictor.*

Model parameters						Model statistics			
Rank	**Int**	**color**	**dist**	**size**	**dist:size**	**df**	**AICc**	**delta**	**weight**
A|C Surveys: Size as small, medium, large									
1	2.58	+	-0.0044	+		12	1218.7	0	0.842
2	2.84	+	-0.0052	+	+	14	1222	3.3	0.158
3	3.17		-0.004	+		4	1283.8	65.1	0
Pred. weight		1	1	1	0.158				
A|C Surveys: Size as sqrt(area)									
1	-0.7	+	-0.0045	0.34		11	1197.5	0	0.544
2	-1.21	+	-0.0029	0.44	-0.0003	12	1197.8	0.35	0.456
3	-1.68	+		0.32		10	1269.1	71.67	0
Pred. weight		1	1	1	0.456				
B surveys: Size as small, medium, large									
1	1.62	+	-0.0055	+		12	686.2	0	0.713
2	0.84	+	0.0000098	+	+	14	688	1.86	0.281
3	0.96	+		+		11	695.8	9.6	0.006
Pred. weight		1	0.994	1	0.281				
B surveys: Size as sqrt(area)									
1	-0.47	+	-0.0059	0.2		11	674.7	0	0.504
2	0.22	+	-0.01	0.08	0.00082	12	674.8	0.04	0.494
3	-0.9	+		0.18		10	685.9	11.19	0.002
Pred. weight		1	0.998	1	0.494				

To examine which colors were similar/different and could therefore be pooled, we used our best model to predict detection rates for small objects (20cm2) at a distance of 200 cm from the observer across the range of colors and then plotted the predicted rate and its 95% confidence interval (**Figure 5**).

The brighter colors of blue, yellow and red had higher detection rates, whereas clear and white colored objects had intermediate detection rates (**Figure 5**). Silver and brown objects had highly variable detection rates, likely due to the low sample size of objects with these colors (silver, n=1; brown, n=2) (**Figure 5**). Black objects had the lowest detection rates, whereas multi-colored objects were intermediate between black and white objects (**Figure 5**).

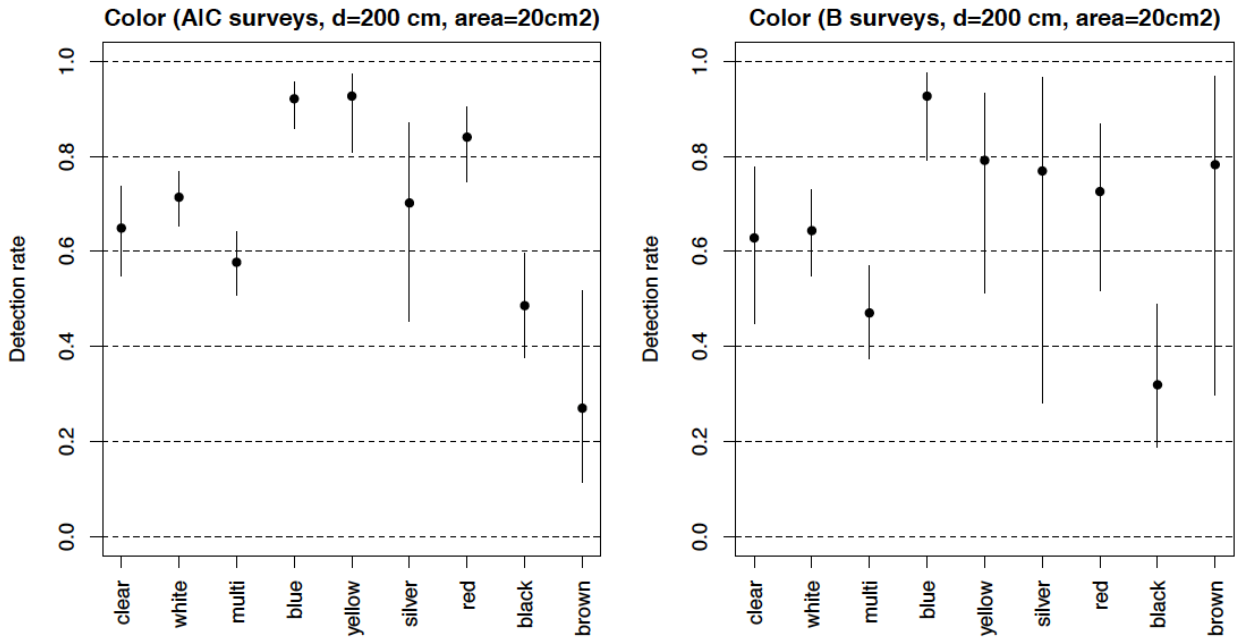


Figure 5. *Fitted detection rates for objects of different colors according to A|C surveys and B surveys. Distance and object size were held constant at 200 cm and 20 cm^2, respectively.*

Based on these results, we grouped colors as: bright (blue, yellow, red), dull (black, brown, multi), reflective (white, clear, silver). We then refitted the models according to the methods outlined above, but with the re-levelled color-scheme. Model results are shown in **Figures 6-8**.

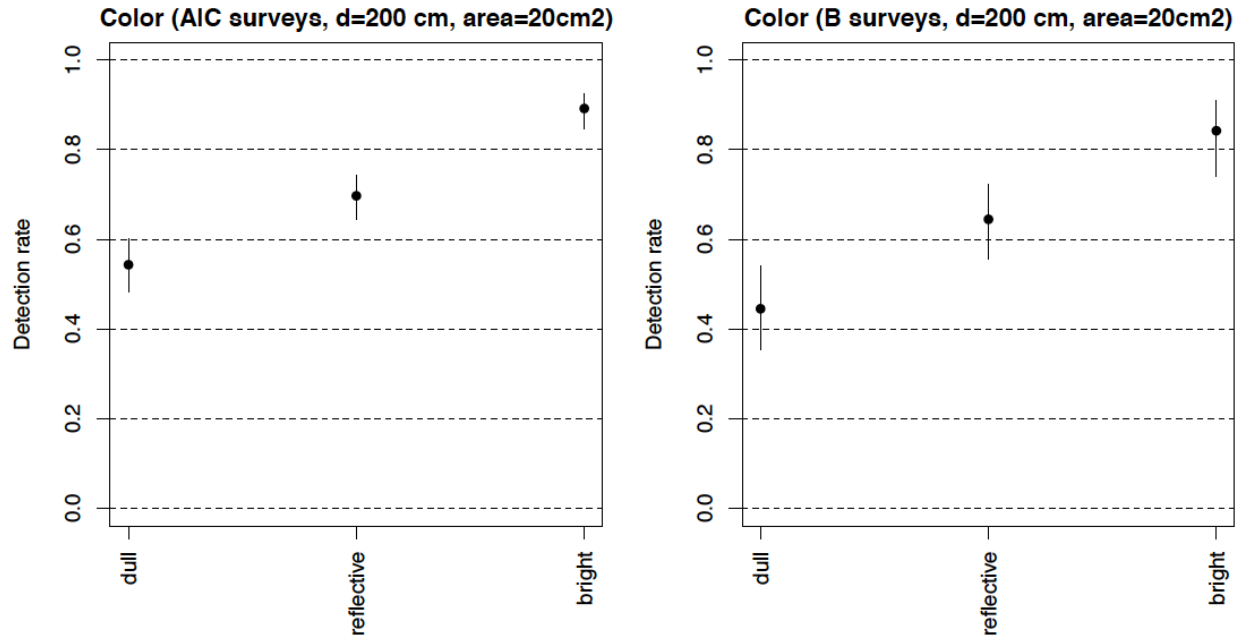


Figure 6. *Fitted detection rates for objects of different color types (bright: red, yellow, blue; reflective: white, clear, silver; dull: brown, black, multi) according to A|C surveys and B surveys. Distance and object size were held constant at 200 cm and 20 cm^2, respectively.*

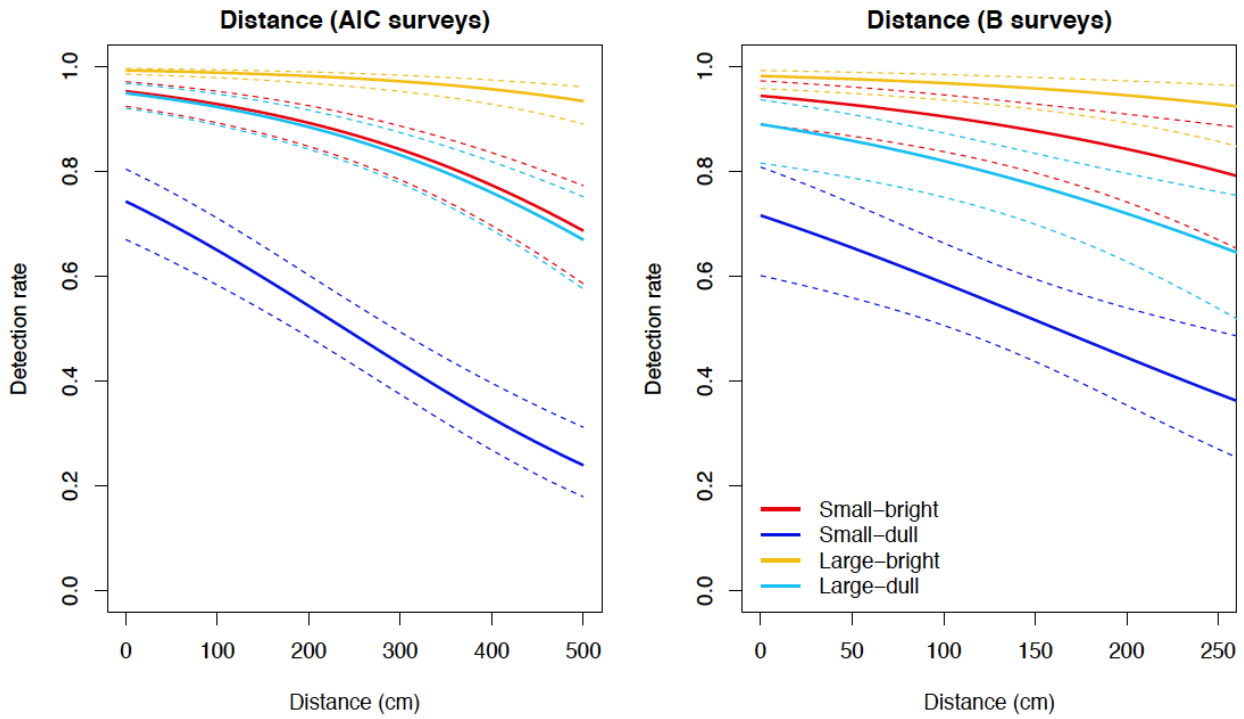


Figure 7. *Fitted detection rates for objects at different distances, size (small: 20 cm^2; large 100 cm^2) and color type (bright: red, blue, yellow; dull: black, brown, multi) according to A|C surveys and B surveys.*

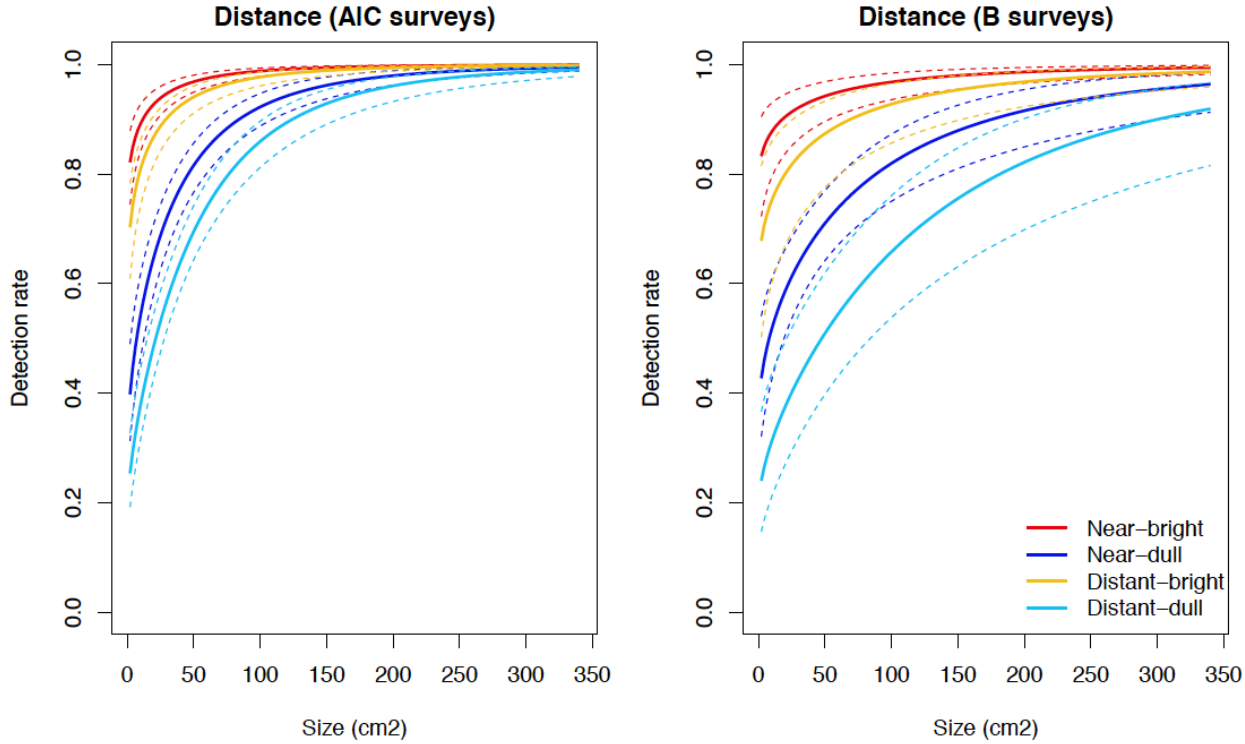


Figure 8. *Fitted detection rates for objects of different sizes for set distances (near = 100 cm, distant = 250 cm) and color type (bright: red, blue, yellow; dull: black, brown, multi) according to A|C surveys and B surveys.*