

## Enumerating Sparse Organisms in Ships' Ballast Water: Why Counting to 10 Is Not So Easy

A. Whitman Miller,<sup>\*,†</sup> Melanie Frazier,<sup>‡</sup> George E. Smith,<sup>†</sup> Elgin S. Perry,<sup>§</sup> Gregory M. Ruiz,<sup>†</sup> and Mario N. Tamburri<sup>⊥</sup>

<sup>†</sup>Smithsonian Environmental Research Center, PO Box 28, Edgewater, Maryland 21037, United States

<sup>‡</sup>Western Ecology Division, National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, 2111 SE Marine Science Drive, Newport, Oregon 97365, United States

<sup>§</sup>Statistics Consultant, 2000 Kings Landing Road, Huntingtown, Maryland 20639, United States

<sup>⊥</sup>Maritime Environmental Resource Center, Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, One Williams Street, Solomons, Maryland 20688, United States

**S** Supporting Information

**ABSTRACT:** To reduce ballast water-borne aquatic invasions worldwide, the International Maritime Organization and United States Coast Guard have each proposed discharge standards specifying maximum concentrations of living biota that may be released in ships' ballast water (BW), but these regulations still lack guidance for standardized type approval and compliance testing of treatment systems. Verifying whether BW meets a discharge standard poses significant challenges. Properly treated BW will contain extremely sparse numbers of live organisms, and robust estimates of rare events require extensive sampling efforts. A balance of analytical rigor and practicality is essential to determine the volume of BW that can be reasonably sampled and processed, yet yield accurate live counts. We applied statistical modeling to a range of sample volumes, plankton concentrations, and regulatory scenarios (i.e., levels of type I and type II errors), and calculated the statistical power of each combination to detect noncompliant discharge concentrations. The model expressly addresses the roles of sampling error, BW volume, and burden of proof on the detection of noncompliant discharges in order to establish a rigorous lower limit of sampling volume. The potential effects of recovery errors (i.e., incomplete recovery and detection of live biota) in relation to sample volume are also discussed.

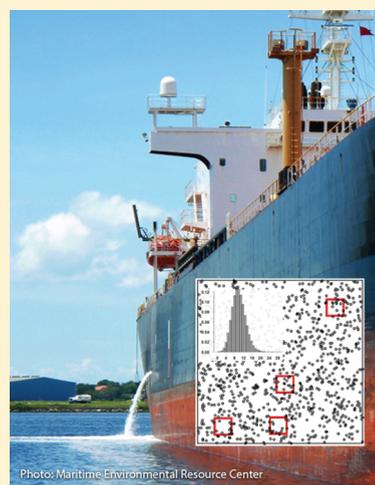


Photo: Maritime Environmental Resource Center

### INTRODUCTION

Maritime transportation is a foundation of the global market. There are well over 50,000 commercial ships which move goods around the world among over 300 major ports.<sup>1,2</sup> However, the ballast water associated with merchant vessel traffic is also responsible for the transfer and introduction of aquatic invasive species to coastal waters where they can cause enormous ecological and economic damage.<sup>3–5</sup>

In an attempt to minimize the risk of BW introductions, the International Maritime Organization (IMO<sup>6</sup>) and U.S. Coast Guard (USCG<sup>7</sup>) have each proposed discharge standards limiting maximum concentrations of living organisms that can be released with BW, including new regulations requiring ship operators to meet those limits. The USCG has proposed to implement regulations in two phases: phase 1 proposes to set standards similar to current IMO standards and phase 2 proposes standards up to 1,000 times stricter. The IMO and USCG phase 1 standards require BW discharged by ships to contain:

- 1 Fewer than 10 viable organisms  $\cdot \text{m}^{-3} \geq 50 \mu\text{m}$  in minimum dimension or smallest measure among length, width, and

height excluding fine appendages such as sensory antenna and setae (the majority of organisms in this size class are zooplankton).

- 2 Fewer than 10 viable organisms  $\cdot \text{mL}^{-1} < 50 \mu\text{m}$  and  $\geq 10 \mu\text{m}$  in minimum dimension. (The majority of organisms in this size class are protozoa, including zooplankton).
- 3 Fewer than the following concentrations of indicator microbes, as a human health standard: (a) toxicogenic *Vibrio cholerae* (serotypes O1 and O139) with  $< 1$  colony forming unit  $\cdot 100 \text{ mL}^{-1}$ ; (b) *Escherichia coli*  $< 250$  cfu  $\cdot 100 \text{ mL}^{-1}$ ; and (c) intestinal *Enterococci*  $< 100$  cfu  $\cdot 100 \text{ mL}^{-1}$ .

To achieve the above discharge standards, technology developers and manufacturers around the world are advancing on-board BW treatment systems<sup>8,9</sup> that use methods such as filtration + UV radiation, deoxygenation, ozonation, and chlorination.<sup>9</sup>

**Received:** August 14, 2010

**Accepted:** March 8, 2011

**Revised:** February 14, 2011

Despite rapid technological advancement, the regulatory framework around BW treatment and discharge is still emerging. In contrast, more mature regulations such as the National Primary Drinking Water Regulations have, for many years, required the use of specific test protocols by certified laboratories for validating treatment efficacy.<sup>10</sup> At present, there are no such codified test procedures designed for validating the effectiveness of BW treatment systems, either on land-based test beds or aboard working ships.

The formulation of standardized BW treatment testing protocols is essential if shipboard BW treatment technologies are to be widely implemented and discharge standards are to be enforced. The success of BW regulations for reducing biological invasions will depend, in large part, on whether (a) approved treatment systems do in fact reduce organism concentrations to the specified standards and (b) individual ships are in compliance with the standards.<sup>11,12</sup> This requires the ability to reliably quantify very few living organisms in large volumes of water.

Enumeration methods are frequently used for quantifying particles and microorganisms in drinking water. Emelko et al.<sup>13</sup> showed that even when using certified sampling and analytical protocols, enumeration of *Cryptosporidium* oocysts in drinking water can yield variable results due to two sources of uncertainty: (1) sampling error and (2) analytical recovery error. The sampling and analysis of BW are prone to the same kinds of error (see Supporting Information for detailed summary of sampling and recovery errors associated with BW discharge analyses). In the absence of standardized sampling and analytical protocols, currently available data are insufficient to create a comprehensive model that quantifies all sources of uncertainty for BW discharge analysis, as has been possible for drinking water.<sup>13,14</sup>

Although we are not yet able to parameterize all potential sources of error, we present a theoretical model that is designed specifically to ascertain the baseline sample volumes required to robustly discern noncompliant zooplankton concentrations under ideal sampling and detection conditions, thereby establishing a rigorous lower limit (or minimum threshold) for sampling effort. This is a crucial first step toward establishing robust sampling procedures for BW regulations that are verifiable and effective,<sup>15,16</sup> as these do not currently exist. Our goal is to provide formal evaluation and guidance on minimum sampling effort to verify BW concentrations, since additional error can never decrease the sampling effort required under the optimized Poisson model presented (which represents the best case scenario). As subsequent studies quantify the various sources of additional error, especially recovery errors, sample volumes should be adjusted to reflect these measures.

In this study, we focus on IMO and USCG proposed phase I standards (hereafter “IMO standard”) and organisms of  $\geq 50 \mu\text{m}$  minimum dimension (hereafter “zooplankton”) to (1) characterize the uncertainty associated with estimating the concentration of organisms in BW due to the stochastic nature of sampling BW (i.e., sampling error); and (2) demonstrate, using specific examples, how various regulatory decisions regarding rates of both type I (i.e., false positive) and type II (i.e., false negative) errors<sup>17</sup> affect the sample volumes needed to verify organism concentrations. In particular, we estimate the statistical power to detect BW concentrations that exceed the current IMO standard of  $<10 \text{ organisms} \cdot \text{m}^{-3}$  using different sample volumes and regulatory scenarios. As discussed above, we focus only on the sampling error expected from BW discharges, since sampling

error should represent a significant source of uncertainty, especially at low concentrations.

## METHODS

Of primary concern is characterizing the sampling effort necessary to quantify live zooplankton concentrations in BW in order to reliably classify BW as noncompliant ( $\geq 10 \cdot \text{m}^{-3}$ ) or compliant ( $<10 \cdot \text{m}^{-3}$ ), with high statistical confidence. Importantly, this sampling effort must be feasible given the realities and logistic constraints specific to BW treatment system testing and ship compliance monitoring. Furthermore, BW verification and compliance testing also require that several decisions be made at the regulatory level, especially if standardized sampling protocols are to be developed.

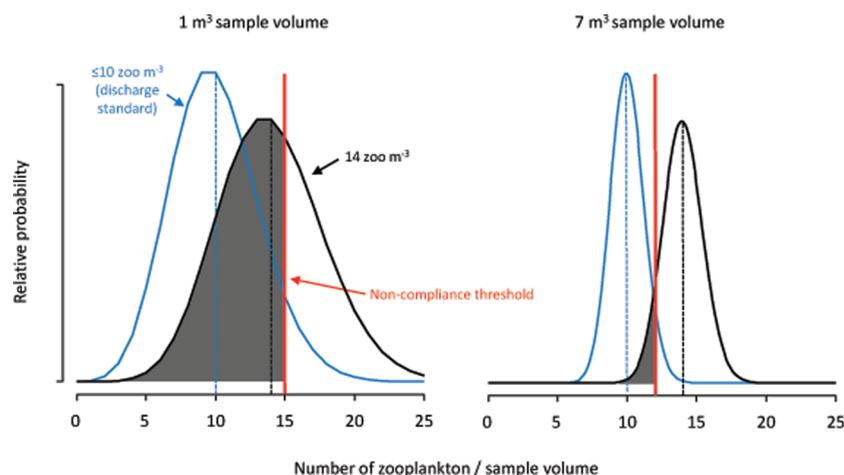
One regulatory decision is how best to handle the inherent uncertainty associated with sampling discharge concentrations, even when using the best sampling protocols. There are at least two philosophies concerning the regulation of BW discharge, which differ according to where the burden of proof is placed. The first is based on the presumption of innocence until proven guilty, which places the burden of proof on the regulator. In this context, a random sample of ballast discharge may contain  $>10 \text{ zooplankton} \cdot \text{m}^{-3}$  and still pass inspection as long as the sample is not *statistically significantly*  $\geq 10 \text{ zooplankton} \cdot \text{m}^{-3}$ . An alternative is to place the burden of proof entirely on the regulated entity, whereby a ship with a measured zooplankton concentration that is not *statistically significantly*  $<10 \cdot \text{m}^{-3}$  is presumed guilty until proven innocent. We use the “presumed innocent” approach in the examples presented in this paper, however, the general methods we describe will apply to other approaches. Given this, treated BW is assumed to have a concentration  $<10 \text{ zooplankton} \cdot \text{m}^{-3}$  until proven otherwise, thus the null hypothesis is as follows:

( $H_0$ ): Concentration of live zooplankton in treated ballast water is  $<10 \text{ zooplankton} \cdot \text{m}^{-3}$ .

At present neither the IMO nor USCG have voiced guidance on which approach will guide regulatory actions, but the approach that is used may depend on the setting and the kind of testing being carried out. For compliance monitoring of individual ships, the “presumed innocent” approach may be preferred. Because a high degree of certainty may be desired for type approval testing of treatment systems (type approval is the process of testing equipment to ensure that it meets technical, safety, and regulatory requirements), it may be reasonable for the burden of proof to be on the manufacturer or ship (i.e., “presumed guilty”).

Regardless of the approach, regulators must also define a standard for how extreme data must be before the null hypothesis is rejected. In statistical terms, this refers to the type I error rate,  $\alpha$ .<sup>17</sup> In the scientific arena, the typical standard is  $\alpha = 0.05$ , however, there is no theoretical reason to assume this should be the default standard for BW regulation, and in fact, this value is often debated in scientific literature. In regard to ballast discharge, if the “presumed innocent” approach is used, then larger  $\alpha$  values will result in more ships being falsely accused of exceeding the limit (i.e., increased false positives). For the examples in this paper we explore how  $\alpha$  values of 0.05 and 0.20 affect sample volume.

Given the statistical framework described above (i.e.,  $\alpha = 0.05$  or 0.20 and a “presumed innocent” approach), we estimated the likelihood of detecting BW with various concentrations that



**Figure 1.** Poisson sample distribution for a population with a concentration that meets the discharge standard of  $< 10 \text{ zooplankton} \cdot \text{m}^{-3}$  (blue curves) and a theoretical test population with a concentration of  $14 \text{ zooplankton} \cdot \text{m}^{-3}$  (black curves) for sample volumes of  $1 \text{ m}^3$  and  $7 \text{ m}^3$ . Gray shading ( $\beta$ ) indicates regions where concentrations cannot be distinguished. Red vertical lines indicate the noncompliance threshold for  $\alpha = 0.05$  (Table 1); random samples that are  $\leq$  noncompliance threshold are classified as compliant with discharge standards based on our definition that ballast is “presumed innocent”. When the concentration of ballast discharge is  $14 \text{ zooplankton} \cdot \text{m}^{-3}$ , nearly 70% of  $1 \text{ m}^3$  sample volumes will result in false negatives (power  $\approx 0.30$  or  $1 - \beta$ ). About 8% of  $7\text{-m}^3$  sample volumes will result in false negatives (power  $\approx 0.92$ ).

exceed the standard. In statistical terms, this is referred to as “power” (Figure 1). To be environmentally protective, regulators must determine the statistical power that is required to adequately enforce BW discharge standards. Low power occurs when the exceedance is small or when sampling is insufficient to yield adequate precision for detecting even a large exceedance.<sup>18</sup> From the vantage point of environmental protection, low power is of great concern because sampling results can falsely suggest that no significant threat is present.<sup>19</sup> Insufficient sampling that yields low power can result in a false sense of security, thereby undermining the intended goals of a testing or monitoring program. To understand which sampling designs maximize power (and optimize sampling effort), we calculated statistical power for a variety of sampling efforts and zooplankton concentrations that exceed the compliance concentration of  $< 10 \text{ zooplankton} \cdot \text{m}^{-3}$ . A power value of 0.80 is frequently considered sufficient to reliably detect statistical differences.<sup>18,20</sup> Nevertheless, Di Stefano<sup>21</sup> argues that the selection of statistical parameters should be based on the respective costs of false positives (i.e., classifying BW as noncompliant when it actually meets the standard) and false negatives (i.e., failing to identify BW that exceeds the standard). We use power values of 0.8 as a reference for comparison among sampling scenarios, but report results from a range of values that correspond to power values ranging from  $< 0.1$  to 1.0.

**Approach.** A two-stage sampling model was applied to a range of hypothetical sample volumes, plankton concentrations, and regulatory scenarios (i.e., levels of type I and type II errors). Power to detect noncompliant discharge concentrations from the proposed discharge standard was calculated for each combination. Stage 1 assesses compliance based on a single sample and is expected to be most useful when the degree of noncompliance is large. Stage 2 combines several independent samples to assess compliance and is expected to improve discrimination when actual concentrations are close to, but still exceed, the discharge standard.

**Assumptions.** If zooplankton are randomly distributed throughout BW discharge (i.e., the presence of one individual does not influence the presence or absence of others), then the Poisson distribution can be used to accurately predict sampling probabilities. This is because integrating a nonhomogenous Poisson process

results in a Poisson distribution which has a mean equal to the mean concentration in the discharge.<sup>22</sup> We employ the following postulates when applying the Poisson distribution to BW discharge: (1) the probability of having some number of organisms in one volume is independent of the number in other discrete volumes; (2) the probability of a single organism in a sample is proportional to the volume of the sample; and (3) the probability of two or more organisms in a very small volume is negligible.

The assumption that biota will be randomly distributed throughout discharge is likely optimistic, since it presupposes that organisms are independent of one another in a BW discharge. Planktonic organisms in BW tanks are known to exhibit complex, yet unpredictable spatial structure owing to diversity of ballast tank design, operation, content, physical mixing that occurs in tanks, and biological interactions and swimming behavior of plankton.<sup>23</sup> Furthermore, some biota are known to aggregate, such as colonial or chain-forming phytoplankton (see Table S1, Supporting Information). Appropriate sampling designs may help ameliorate the effects of aggregation though (see below). Nevertheless, assuming a Poisson sampling distribution will provide the best case scenario with respect to required sample volumes, thereby estimating a lower volumetric limit for what is necessary and sufficient to characterize BW discharge. When organisms are aggregated, estimates of concentrations will be more variable, and consequently larger sample volumes must be taken to obtain reliable estimates of concentration.

The land-based testing centers that are currently evaluating ballast treatment systems circumvent this problem by using in-line sampling of the ballast discharge pipe to collect a representative sample of the entire discharge.<sup>24</sup> In this case, the Poisson distribution can theoretically be used to accurately predict sampling probabilities, but the sample must be well-mixed if an additional subsampling step is performed. For ship-board testing, time-integrated sampling of the entire discharge is probably not possible; however, the problem of aggregation may still be mitigated by sampling at several time points during discharge. Alternatively, if only a single discrete sample is taken from the discharge pipe, it may be indicative of the instantaneous concentration of discharge, but will not necessarily accurately estimate the

**Table 1. Noncompliance Threshold Values for  $\alpha = 0.05$  and 0.20; If Sample Counts or Concentrations Exceed the “Noncompliance Threshold” the Discharge Is Statistically Unlikely To Be Compliant with the IMO Discharge Standard ( $<10$  zooplankton  $\cdot m^{-3}$ )**

sample volume ( $m^3$ )	noncompliance threshold			
	$\alpha = 0.05$		$\alpha = 0.20$	
	count (N)	concentration (zoo $\cdot m^{-3}$ )	count (N)	concentration (zoo $\cdot m^{-3}$ )
1	15	15.0	13	13.0
3	39	13.0	35	11.67
7	84	12.0	77	11.0
14	160	11.43	150	10.71
21	234	11.14	222	10.57
28	308	11	294	10.50
35	381	10.89	366	10.46

mean concentration of the entire BW discharge. More empirical research is necessary to determine how the aggregation of organisms in BW affects sample estimates. In the examples that follow, we assume that organisms are randomly distributed throughout BW or that sampling protocols that eliminate or mitigate this problem are used, and thus can be modeled using the Poisson distribution. Our assumptions include the following:

- 1 The BW sample is time-integrated and proportional to the discharge flow to control for any underlying spatial/temporal structure of organism distribution.
- 2 The total BW sample volume is processed.
- 3 All live organisms  $\geq 50 \mu m$  are captured and detected (i.e., recovery error is negligible; Table S1, Figure S1, Supporting Information).

*Equations. Poisson Probability and Statistical Power*<sup>25–27</sup>

$$P[X = x] = \frac{e^{-m} m^x}{x!} \quad x = 0, 1, 2, \dots, m > 0 \quad (1a)$$

$$P[X \leq c] = \sum_{x=0}^c \frac{e^{-m} m^x}{x!} \quad (1b)$$

$$p = P[X > c] = 1 - P[X \leq c] \quad (1c)$$

where  $X$  = a random variable taking values  $x$  where  $x$  = non-negative integer (i.e., 0, 1, 2, ..., where  $X$  represents the count observed in a sample taken from a population);  $e$  = base of natural logarithms;  $m$  = mean of Poisson distribution (i.e., true concentration of organisms in discharge);  $c$  = count of organisms at the noncompliance threshold for a given  $\alpha$  and sample volume (Table 1);  $p$  = the probability of exceeding  $c$ . In this application,  $p$  is the false positive rate ( $\alpha$ ) when the BW is compliant and is power when BW is noncompliant with the discharge standard, i.e.,

$$P[X > c | \text{null hypothesis true}] \doteq \alpha \quad (2a)$$

$$P[X > c | \text{alternate hypothesis true}] = \text{power} \quad (2b)$$

*Applying Sampling Statistics.* Using the Poisson distribution,<sup>25–27</sup> we modeled the probability that a random sampling unit of ballast discharge will contain a specific number of

organisms (eqs 1a and 1b). For example, if the true concentration of ballast discharge is 5 zooplankton  $\cdot m^{-3}$ , the probability that a random sampling unit ( $1 m^3$ ) will contain 0 organisms is 0.0067 (eq 1a). Alternatively, the probability that a sampling unit will contain  $\leq 3$  organisms is 0.265 (eq 1b). (See Supporting Information 1 for example calculations.) The units for this parameterization of the Poisson distribution equal the number of organisms per sampling unit. To convert to concentration, the total count is divided by the total sampling unit volume.

*Stage 1 (Single Trial Analysis).* Inherent uncertainty around sampling data is reduced by sampling larger volumes.<sup>17</sup> We determined how increased sample volume improves the ability to identify sample concentrations that exceed the IMO standard of  $<10 \cdot m^{-3}$ . We compared the sampling distribution of a zooplankton concentration of  $<10 \cdot m^{-3}$  to sampling distributions obtained from theoretical populations with concentrations  $\geq 10$  zooplankton  $\cdot m^{-3}$  in order to calculate statistical power, based on the Poisson distribution described in eq 1a.

In our framework, a sample of ballast discharge must be statistically significantly  $\geq 10$  zooplankton  $\cdot m^{-3}$  to be classified as noncompliant. The noncompliance threshold represents the maximum number of organisms that are likely to occur in a sample if the concentration does not exceed the standard (Figure 1) given our predetermined  $\alpha$  values. These noncompliance threshold values (Table 1) were determined (eq 1b) by summing the probabilities of obtaining counts from 0 to  $x$ , given a true concentration of  $10 \cdot m^{-3}$ , until the cumulative probability just exceeded 0.95 ( $\alpha = 0.05$ ) or 0.80 ( $\alpha = 0.20$ ). Statistical power was calculated for each  $\alpha$  value to determine how reliably population concentrations ranging from 10 to 20 zooplankton  $\cdot m^{-3}$  could be discriminated from populations of  $<10$  zooplankton  $\cdot m^{-3}$  (eq 1c) for sample volumes of 0.1, 1, 3, and 7  $m^3$ . Single trial analyses may be the only tractable sampling approach available on working ships, and best suited for detecting large exceedances of the discharge standard.

*Stage 2 (Multiple Trial Analysis).* An alternative approach for gauging the efficacy of a treatment system is to pool the results from multiple independent ballast trials and to examine them simultaneously. The simplest, and arguably most powerful, approach for evaluating multiple tests relies on the fact that Poisson distributions are additive and generate a summed Poisson distribution.<sup>22,27</sup> For example, the total number of zooplankton from two 4- $m^3$  trials would be summed and compared to a Poisson distribution where mean and variance = 80 (i.e., the expected count for a 10 zooplankton  $\cdot m^{-3}$  discharge standard and total sample volume of 8  $m^3$ ). To determine how summing the results from multiple trials affects statistical power, we calculated the probability of identifying noncompliant concentrations of 11–14 zooplankton  $\cdot m^{-3}$  for 1–15 independent trials, using 7- $m^3$  sample volumes. For each total sample volume (7–105  $m^3$ ), we calculated a noncompliance threshold value, based on the upper probable count expected in samples with concentrations of 10 zooplankton  $\cdot m^{-3}$  ( $\alpha = 0.05$  in this scenario). Power was calculated by determining the predicted proportion of samples with counts greater than noncompliance threshold values (eqs 1a–1c). Multiple test trials may be most feasible on land-based test beds, which have fewer logistical constraints than ships, and allow for more controlled and repeated sampling and analysis.

*Application of Model to BW Treatment Test Results.* To demonstrate the potential practical utility of this statistical approach, we applied our analysis to discharge data from tests

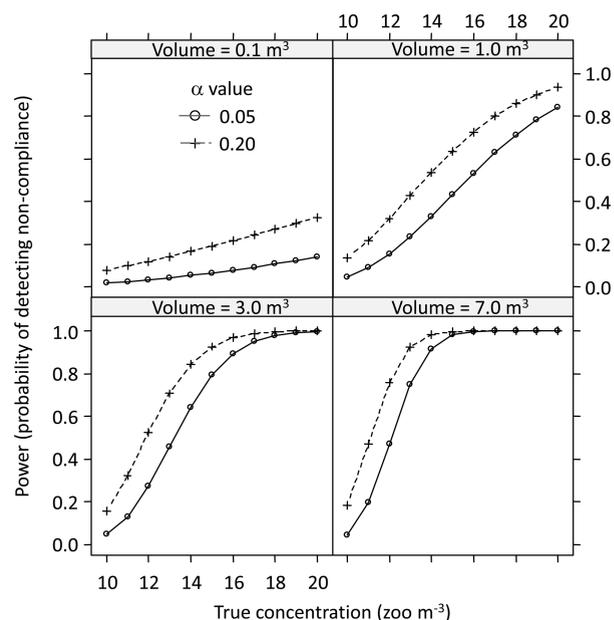
of three BW treatment systems. The tests were conducted at the Maritime Environmental Resource Center (a test facility at the Port of Baltimore, Maryland, USA) to evaluate compliance with the IMO discharge standard. For each treatment system, tests occurred in 4–5 replicate trials, and all live zooplankton were enumerated from 5-m<sup>3</sup> time-integrated samples for each trial. Using the zooplankton counts, we analyzed per-trial results and composite results using the summed Poisson method.<sup>28–30</sup> Importantly, while actual BW treatment system data are used as examples to test our model, it was not our goal to draw conclusions on the performance of any particular system or approach.

## RESULTS AND DISCUSSION

**Single Trial Analyses.** For sample volumes of 1, 3, and 7 m<sup>3</sup>, the noncompliance threshold concentrations are 15.0, 13.0, and 12.0 zooplankton · m<sup>-3</sup>, respectively, if  $\alpha = 0.05$ , and 13.0, 11.7, and 11.0 if  $\alpha = 0.20$  (Table 1). When zooplankton concentrations (10–20 · m<sup>-3</sup>) were modeled under the Poisson distribution (eqs 1a and 1b) at four sampling efforts (0.1, 1, 3, and 7 m<sup>3</sup>), we observed substantial increases in power to discern statistical differences between noncompliant and compliant (<10 zooplankton · m<sup>-3</sup>) concentrations (eq 1c) with larger sample volumes (Figure 2). When  $\alpha = 0.05$ , for a 1-m<sup>3</sup> sample volume, zooplankton concentrations must be  $\geq 20 \cdot \text{m}^{-3}$  before the statistical power of the test to correctly identify a noncompliant tank exceeds 0.8. Increasing  $\alpha$  to 0.20 effectively reduces the “benefit of doubt” that ships are afforded; in this case, for a 1-m<sup>3</sup> sample volume, zooplankton concentrations must be  $\geq 18 \cdot \text{m}^{-3}$  before statistical power exceeds 0.8. For  $\alpha = 0.05$ , when sample volume is increased to 3 m<sup>3</sup>, zooplankton concentrations of 15 and 18 · m<sup>-3</sup> can be differentiated from the discharge standard with power = 0.8 and 0.98, respectively. Further power gains are achieved when sample volume is increased to 7 m<sup>3</sup>: power = 0.92 for a concentration of 14 · m<sup>-3</sup> and near certain detection is expected for concentrations above 15 · m<sup>-3</sup> (Figure 2). Not surprisingly, further increasing sample volumes provides greater precision and confidence; however, additional gains in precision with incremental increases in volume diminish beyond 7 m<sup>3</sup> (Table 1) and the likelihood of nontreatment effects (i.e., increased mortality) with extended sampling and analysis is expected to increase.

In a single trial, if zooplankton concentration exceeds the non-compliance threshold, one can reliably infer (with high statistical confidence) that the mean concentration of the discharge exceeds the standard (see Table 1). As discharge concentrations approach 10 zooplankton · m<sup>-3</sup>, it becomes progressively more difficult to differentiate compliant from noncompliant samples. Since single trial volumes cannot be increased indefinitely, it becomes necessary to combine trials for further gains in statistical power.

Although we have chosen to concentrate exclusively on sampling error in order to help define the lower limits of sample volume, analytical recovery errors can introduce uncertainty that will influence enumeration and the required sample volume.<sup>13,14</sup> Recovery errors are expected to result in under-counting rather than over-counting (i.e., sample bias, Table S1). Although existing BW testing data are insufficient to accurately parameterize recovery errors, we investigated how hypothetical rates of zooplankton recovery (100, 90, 75, and 50%) strongly affect the power to detect noncompliance. As expected, the putative effect of incomplete recovery is most pronounced for smaller sample



**Figure 2.** Power of the Poisson one-sample test to detect noncompliance with a discharge standard of  $<10$  zooplankton · m<sup>-3</sup> as a function of sample volume (0.1, 1, 3, or 7 m<sup>3</sup>), discharge concentration (10–20 zooplankton · m<sup>-3</sup>), and  $\alpha = 0.05$  and 0.20.

volumes and concentrations that are near the discharge standard (Figure S1).

**Multiple Trial Analyses.** Using repeated, independent trials of a BW treatment system provides a more robust test of performance than a single trial for multiple reasons. Repeated measures are needed to test consistency in performance under a range of conditions. Less appreciated is the potential use of a summed Poisson analysis, whereby integrative sampling allows zooplankton counts from multiple trials to be added together, providing a cumulative probability based on total volume sampled (Table 1). This approach can overcome many critical limitations of volume and handling time for single trials. Using this summed Poisson technique, statistical power exceeded 0.8 when comparing concentrations of 14, 13, and 12 zooplankton · m<sup>-3</sup> (with 1, 2, and 3 trials respectively; 7 m<sup>3</sup> per trial;  $\alpha = 0.05$ ) to the discharge standard (Figure 3). Nearly 100% power was achieved for all three test concentrations with 7 trials (total volume = 49 m<sup>3</sup>). As concentrations approach the discharge standard, more trials are required before power exceeds 0.8. When the 11 zooplankton · m<sup>-3</sup> concentration was examined, 10 trials (70 m<sup>3</sup>) were required to attain a power of 0.8 when  $\alpha = 0.05$  (Figure 3). When small sample volumes are used, there is a high probability of mistakenly attributing observed counts to a compliant concentration due to extensive overlap of concentration distributions, with either a single trial or the summed Poisson approach. For example, with a sample volume of 0.1 m<sup>3</sup> the power to detect a moderate exceedance (14 · m<sup>-3</sup> or 40% above the IMO standard) is very low ( $\sim 0.05$ ). Even when ten trials are completed, power to detect exceedance is still low ( $\sim 0.35$ ) (Figure 2). However, increasing sample volume from 0.1 or 1.0 to 7 m<sup>3</sup> enables robust differentiation (power  $> 0.9$ ) of non-compliant zooplankton concentrations of 14 · m<sup>-3</sup> and greater from the IMO standard.

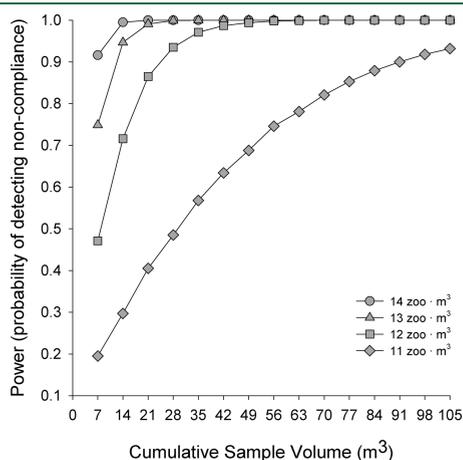
The application of the summed Poisson approach is simple and can be applied iteratively as test results become available. If

sample volume per trial is set at 7 m<sup>3</sup>, then compliant and noncompliant tests will often be apparent after a single test. In cases where results are very close to compliance thresholds, multiple trials may be necessary before success or failure can be fully assessed.

When the summed Poisson method was applied to test data from three different BW treatment systems, results were readily interpreted at the per-trial and multiple trial levels. Although two systems yielded mixed results in which some trials positively rejected the null hypothesis and others did not, when summed Poisson was applied, noncompliance with the discharge standard was unequivocal (Table 2).

In addition to land-based testing currently underway, BW treatment systems on ships will require shipboard evaluations to (1) verify initial performance, and (2) ensure that treatment consistently meets the standard throughout the vessels' lifetime.<sup>15</sup> The summed Poisson method permits rapid and robust analyses of results and can, in some instances, provide extremely prompt performance feedback.

In all probability, identical or similar systems will be installed on multiple vessels. Because discharge standards are concentration-based, they apply to all vessel types, regardless of the environmental conditions of operation. If sampling protocols are standardized across vessels and meet the assumptions described above, then results from multiple vessels might also be considered as independent tests of the same treatment system.



**Figure 3.** Power analysis of the summed Poisson method for identifying BW concentrations that exceed a discharge standard of 10 zooplankton · m<sup>-3</sup> using multiple, 7-m<sup>3</sup> sample volumes from independent trials,  $\alpha = 0.05$ .

Under these circumstances the summed Poisson approach allows individual installations to be assessed separately, thereby providing specific information about the performance of specific installations (single vessels) across time. Alternatively, the fleet can be assessed as a whole, yielding more generalized performance information on the treatment system across platforms.

Although detailed sampling and analysis may not be feasible for frequent, routine, or continuous compliance monitoring of operational BW treatments systems,<sup>15</sup> there will likely be a need for targeted, comprehensive biological assessments of high-risk vessels entering ports. Results from the present analyses indicate that 7-m<sup>3</sup> time-integrated samples may provide a reasonable balance of statistical power and logistic achievability when applied to zooplankton discharge. When applied to actual BW treatment test facility results, the summed Poisson approach provided clear-cut results, even at sample volumes of 5 m<sup>3</sup>. Given the apparent power of this testing protocol, one course of action would be to conduct selected but infrequent biological assessments of BW interspersed with continuous, automated monitoring of treatment system mechanical operations and indirect measures of treatment performance, such as changes in BW physical or chemical conditions.<sup>15</sup> Our approach is well suited for discharge testing of zooplankton (biota  $\geq 50 \mu\text{m}$  in dimension) at the IMO discharge standard. In theory, the same basic statistical treatment should apply to organisms in the regulatory size class ( $\geq 10$  and  $< 50 \mu\text{m}$  in minimum dimension—but admonishments concerning colonial or chain-forming phytoplankton on aggregation must be considered, see Table S1 and references therein), with sample volume and threshold lowered to account for higher concentration allowed in the discharge standard ( $< 10$  viable organisms · mL<sup>-1</sup>), and assuming viable organisms can be as readily detected and differentiated from dead.<sup>31,32</sup> Phase 2 discharge standards proposed by USCG are effectively up to 1000 times more stringent than phase 1, and if implemented, will clearly require protocols (and sample volumes) that differ from what is presented here. Indeed, more sophisticated technologies for use in BW sampling, biological detection, biological viability analysis, and enumeration may be necessary for compliance testing at the USCG phase 2 standard level. Furthermore, while other sources of error must be addressed to identify proper sample volume thresholds (see Supporting Information) regardless of the discharge standard, this likely becomes even more important as the discharge standard becomes more stringent. There are various criteria that must be considered in establishing robust sampling protocols and methods. However, the statistical approach that is ultimately used to enforce ballast water discharge standards will influence

**Table 2. Summed Poisson Analysis Applied to Three Treatment Technologies<sup>a</sup>**

	Trials (no. zooplankton)					Total Zoo	Total Vol (m³)	Noncompliance threshold values:	
	#1	#2	#3	#4	#5			Zoo/5m³	Zoo/Tot Vol
<b>Tech #1</b>	30	40	2,580	135	nd	2,785	20	62	224
<b>Tech #2</b>	285	200	75	110	25	695	25	62	276
<b>Tech #3</b>	7	1	2	4	0	14	25	62	276

<sup>a</sup> All trials employed 5-m<sup>3</sup> time-integrated sampling from discharge pipe. All technologies were evaluated based on individual trial results and the combined trial results. Red shading indicates noncompliance and green indicates compliance with IMO discharge standard for zooplankton ( $\alpha = 0.05$ ).

the ecological and economic outcomes of these regulations. Consequently, it is imperative that the statistical aspects of the sampling protocols be defined. For example, it will be necessary to identify the thresholds used to classify ballast discharge as compliant or noncompliant based on the chosen  $\alpha$  value and enforcement approach. Thus, if all the organisms in 1 m<sup>3</sup> of BW are counted, and a “presumed innocent” approach is used with  $\alpha = 0.05$ , then a ship would be classified as compliant if  $\leq 15$  organisms were counted. However, if a “presumed guilty” approach is used with the same parameters, then a ship would be classified as compliant if  $\leq 4$  organisms were counted. Currently our understanding of how large an inoculation must be to achieve a successful invasion remains coarse.<sup>12</sup> A firmer comprehension of dose—response relationships and invasion success could inform us about which regulatory approach is most appropriate, as well as whether it is crucial to differentiate concentrations that are very close to, but still exceed proposed discharge concentrations. Unfortunately, such biological information is difficult to collect and strong generalities remain elusive. Given the profound influence that these variables can have on regulatory outcome, the consequences of regulatory decisions must be described clearly.

In the end, it is necessary for regulators to determine the level of environmental protection that is acceptable in accordance with scientific evidence and societal needs and desires. In the case of BW-borne biota, the scientific component of decision-making includes a specific set of target discharge standards as well as guidance about the required stringency of tests and/or monitoring procedures to provide sufficient confidence that discharge standards are achieved. Scientific analyses can inform policy makers about the levels of uncertainty associated with testing and monitoring protocols, but regulators must determine how much uncertainty is acceptable.

## ■ ASSOCIATED CONTENT

**S Supporting Information.** Discussion of potential sources of error, both sampling and recovery errors, in BW discharge analysis; table containing descriptions of error types, the expected effects on sample volume, and possible remedies; simple sensitivity analysis describing the effects of under-counting on statistical power according to sample volume and  $\alpha$ ; example calculations of both discrete and cumulative Poisson probabilities. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*Phone 443-482-2439; e-mail: millerw@si.edu.

## ■ ACKNOWLEDGMENT

Support was provided in part by the Maryland Port Administration (508923) and U.S. Maritime Administration (DTMA1H0003) to the Maritime Environmental Resource Center; USCG/National Ballast Information Clearinghouse for A.W.M. (HSCG23-06-C-MMS065); as well as by a U.S. EPA postdoctoral fellowships to M.F. (AMI/GEOSS EP08D00051 and NHEERL). We thank the International Council for the Exploration of the Sea (ICES), Intergovernmental Oceanographic Commission (IOC) and International Maritime

Organization (IMO) Working Group on Ballast and Other Ship Vectors, as well as G. F. Riedel, H. Lee II, D. A. Reusser, R. A. Everett, M. S. Minton, and K. J. Klug for their comments and suggestions on our analyses. This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

## ■ REFERENCES

- (1) UNCTAD. *Review of Maritime Transport*; United Nations Publication: Geneva, 2009.
- (2) EQUASIS Statistics. *The World Merchant Fleet in 2007*; Equasis: France-Ministry for Transport – DAM/SI, 2007.
- (3) NRC. *Stemming the Tide: Controlling Introductions of Nonindigenous Species by Ships' Ballast Water*; National Academy Press: Washington, DC, 1996.
- (4) Ruiz, G. M.; Fofonoff, P. W.; Carlton, J. T.; Wonham, M. J.; Hines, A. H. Invasion of coastal marine communities in North America: Apparent patterns, processes, and biases. *Ann. Rev. Ecol. Syst.* **2000**, *31*, 481–531.
- (5) Fofonoff, P. W.; Ruiz, G. M.; Steves, B.; Carlton, J. T. In ships or on ships? Mechanisms of transfer and invasion of nonnative species to the coasts of North America. In *Invasive Species: Vectors and Management*; Ruiz, G. M., Carlton, J. T., Eds.; Island Press: Washington, DC, 2003.
- (6) International Convention for the Control and Management of Ships' Ballast Water and Sediments; Doc. IMO/BWM/CONF36, 16 February 2004.
- (7) U.S. Department of Homeland Security, Coast Guard. Notice of Proposed Rulemaking, 28 August 2009, Standards for Living Organisms in Ships' Ballast Water Discharged in U.S. Waters, 33 CFR Part 151 46 CFR Part 162. *Fed. Register* **2009**, *74* (166), 44632–44672.
- (8) Dobroski, N.; Scianni, C.; Takata, L.; Falkner, M. Update: Ballast Water Treatment Technologies for Use in California Waters; California State Lands Commission, Marine Invasive Species Program, October 2009.
- (9) Lloyd's Register. Ballast Water Treatment Technology: Current Status. February 2010.
- (10) U.S. EPA. National Primary Drinking Water Regulations, 40 CFR Part 141; 2002.
- (11) Minton, M. S.; Verling, E.; Miller, A. W.; Ruiz, G. M. Reducing propagule supply and coastal invasions via ships: Effects of emerging strategies. *Front. Ecol. Environ.* **2005**, *3* (6), 304–308.
- (12) Bailey, S. A.; Velez-Espino, L. A.; Johannsson, O. E.; Koops, M. A.; Wiley, C. J. Estimating establishment probabilities of *Cladocera* introduced at low density: An evaluation of the proposed ballast water discharge standards. *Can. J. Fish. Aquat. Sci.* **2009**, *66*, 261–276.
- (13) Emelko, M. B.; Schmidt, P. J.; Reilly, P. M. Particle and microorganism enumeration data: enabling quantitative rigor and judicious interpretation. *Environ. Sci. Technol.* **2010**, *44*, 1720–1727.
- (14) Schmidt, P. J.; Emelko, M. B.; Reilly, P. M. Quantification of analytical recovery in particle and microorganism enumeration methods. *Environ. Sci. Technol.* **2010**, *44*, 1705–1712.
- (15) King, D. M.; Tamburri, M. N. Verifying compliance in ballast water discharge regulations. *Ocean. Dev. Int. Law.* **2010**, *41*, 1–14.
- (16) Lee II, H.; Reusser, D. A.; Frazier, M.; Ruiz, G. M. *Density Matters: Review of Approaches to Setting Organism-Based Ballast Water Discharge Standards*; EPA/600/R-10/031; U.S. EPA, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Western Ecology Division, 2010.
- (17) Sokal, R. R.; Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research*, 2nd ed.; W. H. Freeman and Company: New York, 1981.
- (18) Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; L. Erlbaum Associates: Hillsdale, NJ, 1988.

(19) Peterman, R. M. The importance of reporting statistical power: The forest decline and acidic deposition example. *Ecology* **1990**, *71* (5), 2024–2027.

(20) Park, H. M. *Hypothesis Testing and Statistical Power of a Test*; Working Paper; The University Information Technology Services (UITIS) Center for Statistical and Mathematical Computing, Indiana University, Bloomington, IN, 2008.

(21) Di Stefano, J. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Funct. Ecol.* **2003**, *17*, 707–709.

(22) Cox, D. R.; Isham, V. Point Processes. In *Monographs on Applied Probability and Statistics*; Chapman and Hall: London, 1980.

(23) Murphy, K. R.; Ritz, D.; Hewitt, C. L. Heterogeneous zooplankton distribution in a ship's ballast tanks. *J. Plankton. Res.* **2002**, *24*, 729–734.

(24) Richard, R. V.; Grant, J. F.; Lemieux, E. J. *Analysis of Ballast Water Sampling Port Designs Using Computational Fluid Dynamics*; Report CG-D-01-08; U.S. Coast Guard Research and Development Center, 2008.

(25) Bolker, B. M. *Ecological Models and Data in R*; Princeton University Press: Princeton, NJ, 2008; 408 pp.

(26) Elliot, J. M. *Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates*; Scientific Publication No. 25; Freshwater Biological Association, 1971.

(27) Patel, J. K.; Kapadia, C. H.; Owen, D. B. *Handbook of Statistical Distributions*; Marcel Dekker, Inc.: New York, 1976.

(28) Maritime Environmental Resource Center. *Land-Based Evaluations of the Siemens Water Technologies SiCURE Ballast Water Management System*; MERC ER02-10, [UMCES] CBL 10-038; 2010.

(29) Maritime Environmental Resource Center. *Land-Based Evaluations of the Severn Trent De Nora BalPure BP-1000 Ballast Water Management System*; MERC ER01-10, [UMCES] CBL 10-015, 2010.

(30) Maritime Environmental Resource Center. *Land-Based Evaluations of the Maritime Solutions, Inc. Ballast Water Treatment System*; MERC ER02-09, [UMCES] CBL 09-138, 2009.

(31) Tang, Y. Z.; Dobbs, F. C. Green autofluorescence in dinoflagellates, diatoms, and other microalgae and its implications for vital staining and morphological studies. *Appl. Environ. Microbiol.* **2007**, *73*, 2306–2313.

(32) Drake, L. A.; Steinberg, M. K.; Riley, S. C.; Robbins, S. H.; Nelson, B. N.; Lemieux, E. J. *Development of a Method to Determine the Viability of Organisms >10  $\mu\text{m}$  and <50  $\mu\text{m}$  in Ships' Ballast Water: A Combination of Two Vital, Fluorescent Stains*; 6130/1011; U.S. Naval Research Laboratory, 2010.