Workshop Report on Testing of Ballast Water Treatment Systems:

General Guidelines and Step-wise Strategy Toward Shipboard Testing

(June 14-16 2005, Portland, Oregon)

Prepared by GM Ruiz, GE Smith, & M Sytsma

Aquatic Bioinvasion Research and Policy Institute

A joint program of the
Smithsonian Environmental Research Center (SERC)
& Portland State University (PSU)

Workshop Participants:

Dorn Carlson, Jeffrey Cordell, Fred Dobbs, Richard Everett, John Ferguson, Randy Fisher, Yasuwo Fukuyo, Robert Gensemer, Fred Goetz, Shinichi Hanayama, Penny Herring, Russell Herwig, Michael Holmes, Brian Howes, Lisa Ka'aihue, Henry Lee, Marilyn Leland, Edward Lemieux, Jose Matheickal, Whitman Miller, Kathleen Murphy, Gregory Ruiz, Stephen Phillips, Robert Piorkowski, George Smith, Scott Smith, David Stocks, Bill Stubblefield, Mark Sytsma, Lynn Takata, Mario Tamburri, Craig Taylor, Nick Welschmeyer, and Christopher Wiley.

Submitted to:

Pacific States Marine Fisheries Commission

Alaska Department of Fish and Game

Prince William Sound Regional Citizens Advisory Council

21 June 2006

Preface

Biological invasions in coastal ecosystems have resulted in significant ecological, economic, and public health impacts. Invasions result from the establishment of species beyond their historical geographic range. Invasions have been a focus of increasing attention from scientists, resource managers, and policy-makers, because the observed rate of new invasions in recent time has increased exponentially for coastal habitats worldwide.

In North America, most marine invasions have resulted from transfer of species associated with ships. An increase in the volume and speed of shipping has accelerated shipping related introductions, and the transfer of species by ships may be driving the increase in the rate of new invasions. Concern about ship-mediated introductions has resulted in numerous state, federal, and international efforts to develop and evaluate treatment systems that reduce the number of organisms discharged in ballast water.

Evaluation of the performance of ballast treatment practices/technologies onboard ships, at full scale and under realistic operational regimes, is a common requirement of all ballast management efforts. Such scientific evaluation of treatment systems is challenging and expensive. To date, a diversity of approaches has been used to evaluate ballast water treatment (BWT) systems, often using different treatment conditions, methods, and response measurements.

Past evaluations have met with mixed results. Some have suffered from technical difficulties and others from omissions in the experimental design. Importantly, the different methods used across studies make comparisons across studies (either different technologies or different locations and conditions) difficult. In addition, the diversity of methods also seems to represent some confusion about criteria needed to evaluate treatments and to determine compliance with regulatory requirements.

We held a workshop on 14-16 June 2005 to review strategies for shipboard evaluation of BWT systems. The overall objective of the workshop was to develop consensus recommendations and general guidelines for a standardized approach to shipboard evaluation, including groundwork leading up to fullscale testing.

Workshop participants were selected to include experts in many areas of ballast water research, toxicology, experimental design, ship operations, and biology/ecology of particular groups of organisms, including bacteria, protists, and zooplankton. Participants included mostly research scientists who were familiar and actively involved in the ballast water issue. Most participants were from the United States, but the workshop also included experts from Canada, Japan, Singapore, and UK.

The workshop was not a venue to examine or discuss specific technologies. Speakers were not invited to present technology-specific information or input. Our intent was to be technology-neutral in considering standard approaches to evaluating the performance of BWTs aboard ships.

This report provides a synthesis of information from the workshop presentations and discussions, exploring many key issues in testing BWT. Although the primary focus of the workshop was intended to be shipboard testing, the role of scale and when to conduct shipboard tests was a central theme. Thus, the output of this workshop includes explicit consideration of both scale and experimental approach, in establishing general guidelines for treatment testing.

Various aspects of these guidelines also now exist in some of the emerging programs to test ballast treatment systems. Within the United States, this is seen in the recent documentation for full-scale testing in the Environmental Technology Verification (ETV) Program and the Shipboard Technology Evaluation Program (STEP). More recently, some of these same guidelines have been advanced for shipboard evaluation by the International Maritime Organization (IMO). The parallels between the Workshop synthesis and these programs result from two factors: (a) both processes relied upon some of the same technical expertise to reach similar conclusions, and (b) a two-way exchange of information has occurred between the Workshop and the evolving treatment programs.

This Workshop report seeks to provide a first synthesis of these general guidelines and the underlying rationale. Many of the same ideas are stated or implied in documentation surrounding the ETV Program and STEP, but these have not been integrated into a single coherent document. Thus, the following document serves to provide this integration, explicitly examining the role of scale and outlining an overall strategy for ballast treatment testing, as now being implemented in the United States.

We wish to emphasize that the ideas and concepts in this report result from contributions made by the workshop participants, both in presentations and discussions across the 2-day meeting. Thus, we have attempted to present the key points and consensus viewpoints into a synthetic document, and we gratefully acknowledge the contributions of all workshop participants.

The workshop was supported by Pacific States Marine Fisheries Commission, Alaska Department of Fish and Game, and Prince William Sound Regional Citizens Advisory Council. We also wish thank Randy Fisher, Stephen Phillips, Susan Anderson, and the Pacific States Marine Fisheries Commission for critical logistical support in hosting and organizing the workshop and for their active participation.

1. Introduction

The human-mediated transfer of organisms across the globe is a potent force of change. Once established, non-native populations can become numerically or functionally dominant in invaded communities. Although effects of most invasions remain unexplored, it is evident that some non-native species are having significant and widespread impacts by altering ecosystem processes, impacting economies, and affecting human health (e.g., OTA 1993, Wilcove et al. 1997, Mack et al. 2000, Pimentel et al. 2000, Mooney et al. 2005).

In coastal ecosystems, ships have been a major transfer mechanism (vector) for the movement of species. Ships have contributed strongly to the cumulative number of invasions and also appear largely responsible for a dramatic increase in the rate of known invasions in recent time (Mills et al. 1993, 2004, Cohen and Carlton 1995, Reise 1998, Ruiz et al. 2000a, Hewitt et al. 2004).

Historically, ships have moved organisms associated primarily with their underwater surfaces and ballasted materials, as an unintended result of normal operations. Exposed outer surfaces are colonized by a wide range of fouling organisms, such as barnacles, mussels, bryozoans, and hydroids (Minchin and Gollasch 2003). Organisms are also associated with ballasted materials, which are carried inside of ships. Ballast is used to maintain trim and stability during voyages. Since the late 1800s, ships have increasingly used ballast water, which is pumped into clean, dedicated tanks from the surrounding waters and discharged variously at subsequent ports of call (Carlton 1985). A taxonomically diverse community of organisms is entrained and transported within ballast tanks (e.g., Carlton and Geller 1993, Smith et al. 1999, Hines and Ruiz 2000, Ruiz et al. 2000b, Hulsman and Galil, Drake et al. 2005). Organisms arriving by each mechanism have colonized new areas, establishing self-sustaining populations that often spread geographically.

Concerns about increasing rates of coastal invasions, combined with strong effects observed in several recent instances, have led to numerous efforts to limit the transfer of organisms by ships. Most of these focus on ships' ballast water. In the United States and several other nations, commercial ships arriving from overseas are now required to conduct mid-ocean ballast water exchange (BWE) before discharging ballast water. BWE is used to flush coastal water from the ballast tanks and replace it with oceanic water, reducing the initial concentration of coastal organisms (which are those most likely to colonize coastal habitats surrounding ports).

BWE is currently the only treatment used by ships and is viewed generally as a temporary, "stop-gap" measure to reduce the risk of invasions, pending further requirements for alternative treatments to kill or remove organisms prior to discharge. BWE is a management strategy that many ships can implement immediately and does not require retrofitting or development of new technology. Although it is clear that BWE reduces the concentration of coastal organisms, and thereby risk of invasions, it also has some limitations. First, it is not always possible to conduct an exchange, due to safety conditions (e.g., in high seas where ballast provides critical stability), inherent constraints in the design of some ships, or routes that do not transit a sufficient duration or distance from shore to perform open-ocean exchange. Second, even when performed, ballast exchange still leaves a potentially significant residual of coastal organisms.

Efforts to advance technological alternatives to BWE exist at state, national and international levels. Within the United States, both state and federal regulations currently allow for use of alternative treatments that are as effective as BWE. A significant amount of research is underway to develop and test technologies for treatment, and new programs are taking shape to allow demonstration and testing aboard operational vessels in lieu of BWE. Similar efforts are occurring in other countries.

In 2004, the International Maritime Organization (IMO) reached an agreement to adopt concentration-based discharge standards for organisms in ballast water, intended to implement technological treatment and replace BWE (IMO 2004). This treaty is awaiting ratification and not yet in force, but a similar approach is gaining momentum within individual countries. Thus, it appears that use of technological treatments to meet specific discharge standards will be implemented through several national and international regulations.

Despite considerable research being conducted or considered to test ballast treatment technologies, general guidelines are only now taking shape, and these have yet to be developed in a coordinated fashion within or across nations. To date, the approach used in research and testing of treatments has been largely investigator-driven. This has resulted in a patchwork of studies differing in spatial and temporal scales, operating conditions, experimental design, methods, and taxonomic groups. While providing some useful insights, these differences in approach also limit opportunities for direct comparisons of results across studies, treatments, locations, and conditions. This creates significant challenges in interpreting results, especially where two studies yield dissimilar or even conflicting results. In addition, there is considerable confusion about the approach and criteria needed for rigorous evaluation by technology developers, resource managers, funding agencies, and ship operators.

Here, we present a conceptual framework for testing the effects of ballast water treatment (BWT). Although our primary motivation is developing an optimal strategy for shipboard testing, this includes by necessity an integrated, step-wise approach that utilizes laboratory and mesocosm studies as well. We have outlined a set of guiding principles in several areas, including: (a) characterization of treatment, (b) the role of scale in testing, (c) experimental design, (d) selection of test organisms, (e) viability of organisms, (f) residual toxicity, and (g) selecting ships and routes. In discussing these issues, we have intentionally not outlined a prescriptive approach. Instead, we have highlighted several key issues to consider explicitly for any testing program as a step toward developing standardized elements across programs.

2. Guiding Principles for Testing Ballast Water Treatments

The overall goal of BWT is to reduce the risk of invasion by limiting the delivery of viable propagules, but the specific target or standard necessary to achieve the desired results is still under debate. While it is evident that reduction in propagule supply will reduce the likelihood of non-native populations becoming established, the exact relationship between probability of invasion and particular concentrations of organisms is not resolved for aquatic ecosystems, making it difficult to identify specific science-based targets (Ruiz & Carlton 2003). As a result, three general concepts have served as reference points in the discussion about standards: (a) treatment technologies should achieve propoagule supplies lower than BWE or, in the case of short-duration and coastwise voyages (which may not be able to conduct BWE, as discussed above), current operating procedures; (b) the lower the concentration of organisms, the lower the risk of invasion; and (c) there is no ballast-mediated invasion risk with zero organism discharge.

At the present time, several federal and state laws require that any alternative treatment be "as effective as" ballast water exchange, making this the interim standard (e.g., NISA 1996). The IMO has identified specific standards for permissible organism concentrations in the 2004 convention (see Box A), although this agreement has not yet been ratified and entered into force. The convention allows for individual nations to implement more stringent requirements. Within the U.S., pending regulations and new federal legislation would each set tighter discharge standards for the Nation. In short, specific numerical standards have not yet been established and widely accepted.

Box A: International Convention for the Control and Management of Ships Ballast Water and Sediment, 2004: Numerical Discharge Standard.

The IMO Ballast Water Management Convention (IMO 2004) sets maximum discharge standards for ballast water, replacing the use of ballast water exchange on a phased time schedule that varies with the size (ballast capacity) and construction date of the vessel. The discharge standards outlined are as follows:

- For organisms ≥ 50 micrometers in minimum dimension, not more than 10 viable organisms per m^3;
- For organisms between 10 and 50 microns in minimum dimension, not more than 10 viable organisms per milliliter;
- For indicator microorganisms, as follows:
  (a) Toxigenic *Vibrio cholerae* – 1 colony forming unit per 100 millilter or per gram wet weight of zooplankton;
  (b) *Escherichia coli* – 250 colony forming units per 100 milliliter;
  (c) Intestinital Enterococci – 100 colony forming units per 100 millilter

The IMO convention will not enter into force until 12 months after 30 States (countries) representing 35% of the world's gross tonnage are signatory. Signatories to date (as of 30 April, 2006, per IMO website http://www.imo.org/conventions) include: Maldives, Nigeria, Saint Kitts & Nevis, Spain, Syrian Arab Republic, and Tuvalu.

For more information see http://www.imo.org/home.asp

The lack of specific accepted standards poses some difficulties in advancing development and implementation of technologies. Identification of specific standards is a high priority for many groups, especially those engaged in the development of technological treatments as well as state and federal management agencies. Standards clearly provide the target or goal for treatments, determining which exact measurements and magnitude of effects are required to demonstrate acceptable performance.

However, even when specific standards are determined, it also appears likely that these will evolve through time, reflecting increased understanding and technical capability, just as observed for regulations involving air and water quality. In fact, most pending legislative efforts to establish standards include provisions for reviews at some regular interval to assess the capacity of current technologies to meet standards and also the effectiveness of standards to reduce invasions.

Regardless of any uncertainty about specific standards, the quantitative measurement of treatment effects can proceed. Assessing the efficacy of any treatment requires direct measures of its effect on the concentration and condition (or viability) of organisms released from ballast tanks. The same fundamental research approach applies independent of a potentially dynamic or shifting set of treatment standards. The questions and variables remain the same, albeit with minor adjustments to the specific targets for organism types and concentration thresholds. Below, we outline a common set of guiding principles, which apply to the evaluation of BWTs and are not specific to a particular standard.

Treatment Characterization and Mode of Action

A clear first step for testing a BWT, or any other treatment, is to define as explicitly as possible the mode of action, expected effects, and any likely constraints of the specific treatment, based upon existing knowledge. This initial step is critical to the design of a testing program for each specific treatment, as it helps to define (to various extents) the appropriate scales, conditions, organisms, and measures for analyses.

An extensive literature exists for many types of water treatment processes that should be used to characterize the current state of knowledge about performance. Critical analysis of existing information is necessary to identify the specific mechanism(s) for organism removal / deactivation and the fate of any possible residuals (or by-products, in the case of chemicals). In turn, this helps define expectations for a particular treatment process and address a set of core questions regarding its performance for particular operating conditions, environmental conditions, and types of organisms (Box B). Even where such a literature base is not available, and only a theoretical basis exists to consider a treatment process, every effort should be made to clearly identify expected effects and constraints.

Box B: Core questions to help define the mode of action and expected range of applications for a ballast treatment

- How does this treatment process operate to kill or remove organisms?
- What types, sizes, and life stages of organisms will it affect?
- What organisms are most and least vulnerable to this treatment process, in terms of taxonomic groups, sizes, life stages?
- Is it possible to vary the magnitude (i.e., concentration, power, wavelength, temperature, etc.) of the treatment process, and if so, how does intensity affect the treatment effect?
- Under what environmental conditions, time periods, and (as applicable) dosages will this process be most and least effective?
- Is this process applicable to all ship types, tank volumes, and flow rates?
- What is the potential for toxicity due to residuals or disinfection byproducts, and what is known about potential toxicity effects of treated water in surrounding waters at ballast discharge?

Without a solid understanding about the mode of action and expected outcomes, any treatment system becomes a ‘black box’, whose performance and operational limits are undefined. While testing can certainly be done to evaluate the effects of such treatments on a case-by-case basis, the ability to extrapolate and generalize results to a broader range of ships, geographic locations, species assemblages, and environmental conditions is greatly compromised by this approach. For example, Table 1 illustrates the possible permutations for high and low (plus or minus) values for only four environmental variables. When considering the other additional factors that may affect performance, the potential number of permutations becomes truly daunting.

A rigorous scientific approach evaluates the mode of action (mechanism of the treatment process) to test predictions and define limitations, but questions will always remain about treatment efficacy. In general, all treatments will have limits and a range of conditions for effective operation, whether imposed by the mode of action itself or current design / engineering constraints of application (such as physical scale or rate functions). Thus, a major goal of research and testing is to determine not only treatment efficacy but also how efficacy varies over the range of conditions under which a specific treatment is expected to work.

Table 1. Possible combinations of four environmental factors that may affect the performance of some ballast treatments. Shown are combinations of high (+) and low (-) values, and the circle shows one of 16 possible test combinations

Salinity	+	+	+	+	+	+	+	+
Temperature	-	+	+	+	-	-	-	+
Turbidity	-	-	+	+	+	+	-	-
Biomass	-	-	-	+	-	+	+	+
Salinity	-	-	-	-	-	-	-	-
Temperature	-	+	+	+	-	-	-	+
Turbidity	-	-	+	+	+	+	-	-
Biomass	-	-	-	+	-	+	+	+

Importantly, even if a treatment process cannot meet treatment or discharge standards for all taxonomic groups, ballast practices, or environmental conditions, this does not exclude its utility or application. It is possible, perhaps even likely, that several treatment types will be advanced and implemented for different circumstances. Some may be most effective, economical, or desirable for vessels carrying low ballast volumes, for particular voyage durations, or under specific environmental conditions (routes). Some treatments will be more effective for some biota than others. Moreover, there is already consideration and exploration of using multiple treatments in tandem, on the same vessel(s), to achieve the desired effects.

Scaling Up: When to Conduct Shipboard Evaluation

There is considerable pressure for full-scale shipboard evaluation of ballast treatment technologies. This results from the increasing regulations over time for ships to treat ballast water and also from technology developers striving to demonstrate, market, and sell treatment systems. While evaluation of the performance of ballast treatment technologies onboard ships is necessary at full scale and under realistic operations, it is also challenging and expensive in terms of (a) the installation of treatment technology, (b) the range of conditions that can be tested on any one vessel, which is constrained by design, operations, and route, and (c) the effort involved by the ships' crew and research teams to conduct and evaluate experiments.

A general paradigm for developing a new system includes a step-wise progression from concept and design, followed by testing at very small laboratory scales, followed by intermediate and increasingly realistic scales, prior to full-scale demonstration . The rationale for this approach is compelling: If a system does not work at small scales, it is not going to work at larger scales. Small scale experiments can provide a critical proof of concept, assessing whether the treatment meets or departs from expectations as well as identifying particular limitations. Results from such initial testing may indicate critical flaws or lead to design changes and improvements before scaling up.

Such a step-wise approach should result in an efficient process, allowing many treatment technologies to advance quickly into a testing phase, as the cost and effort at the laboratory scale is not prohibitive. Moreover, small scale laboratory experiments are most conducive for testing performance over the largest range of conditions, establishing a strong baseline of data on efficacy and limitations (i.e., conditions under which a treatment will consistently achieve the desired goals). The number of test conditions are likely to decrease at larger scales, as a practical matter, and testing is geared less at understanding the treatment process (such as dose-response, effects of time, under various conditions) and more toward assuring the engineered system functions as expected (see Box C).

There is clearly a need for full-scale testing, which integrates all sources of variation and interactions in the real world (Diamond 1985), but the ship should be viewed as the platform for final validation, not the venue for establishing proof of concept, or basic relationships between treatment magnitude and efficacy. Land-based and ship-board testing should be sequentially and separately advanced, operating on different scales and addressing different questions (see Box C). Following this approach, only systems shown to perform adequately and predictably at bench and intermediate scale should be tested at full-scale on ships. If the response to treatment is not predictable, it is evident that the mode of action is not well understood, and further research is necessary at bench and intermediate scales to adequately characterize the response(s) prior to shipboard demonstration.

Box C: The role of different scales in testing ballast water treatment technologies

A standard engineering paradigm for scale-up of a new system includes a logical progression of step-wise testing that serves to maximize information and minimize both cost and the likelihood of unexpected poor performance at a late-stage and full scale. Following initial steps of defining the mode of action, expectations, and constraints of a particular treatment method (see previous section, Treatment Characterization and Mode of Action), a test plan can be broken down into several different steps, which are designed to address different goals or questions. Ideally, these are implemented sequentially and involve different scales, as follows:

(a) Small (Laboratory) Scale Experiments – Laboratory tests are designed to rigorously and quantitatively define the response of organisms to treatment under controlled laboratory conditions. The scale is small, using bench top containers or microcosms with volumes often less than 5 liters, to minimize the logistics and expense. The purpose is to define the "dose" (chemical and physical treatment characteristics, including magnitude, duration and other key treatment attributes) required to achieve particular results (mortality or removal). Extensive experiments at this scale should include a wide and realistic range of organisms across a wide range of environmental conditions, examining especially the limits of the treatment process to operate under challenging conditions. In addition, the lab-based experiments should include sufficient replication to assure consistent, repeatable, and predictable results based upon the expected mode of action and relative to untreated controls (see next section, Experimental Design)

(b) ***Intermediate (Mesocosm) Scale Experiments –*** Tests at this stage are designed to test for conformity of results to those observed at smaller scales, increasing the complexity to include more realistic conditions such as volume and tank-like characteristics. The scale is intermediate, using volumes of hundreds to thousands of liters. Experiments are conducted on a narrower range of conditions and organisms, including more resistant forms and challenging/limiting conditions. The primary goal here is to confirm treatment performance with increasing scale and engineering design. The intermediate scale experiments should explicitly include the model treatment system and engineering similar to the full scale version, allowing opportunities to test and refine the system at a prototype stage. A successful intermediate stage provides confidence that the treatment system is likely to work, both in terms of the engineering and the effect, at larger scales.

(c) ***Full Scale Experiments*** – A full scale test bed is used to evaluate whether the equipment and treatment meets expected performance objectives or standards. The scale should fall within the range envisioned for actual treatment, using volumes of hundreds of tons and relevant flow rates. The objective here is to demonstrate consistent and predictable performance at full scale, simulating real world application and conditions. Tests should include a range of environmental conditions and organism types, including more resistant forms and challenging conditions. The test bed allows fine-grained analysis of spatial variation associated with "in tank" habitats, including vertical location in the water column, proximity to tank bottom or walls, etc. More broadly, the test bed also provides a platform for extensive work on sampling and analytical methods at full scale. Tests at this scale should explicitly consider and incorporate such spatial variation in analysis of independent and dependent variables (as discussed in next section).

(d) ***Shipboard Evaluation –*** Building upon the previous step-wise sequence, the shipboard evaluation tests whether the treatment technology is operating as expected when fully integrated on the ship. While this includes extensive measurements, the intent is not to demonstrate 'proof of concept' but instead demonstrate acceptable and expected performance, which is consistent with that observed at earlier stages. Since the mode of interaction and the dose-response relationships should have been previously well defined, a considerable focus of work at the shipboard stage is on the engineering design and function. Sampling at this stage is designed to quantify operating conditions and confirm expected performance across the full range of operating conditions intended for use, which may include the interaction of several factors (e.g., vibration, motion associated with sea surface conditions, particular sources of water/biota) that were not included in test bed testing. A key issue to include in shipboard testing is whether performance changes over time in service and if so, what maintenance frequency is required.

Scale step-ups should only follow on the system passing acceptably through carefully laid out tests at the previous, smaller scale. Economy of small scale and ease of manipulating environmental variables and community assemblage at the laboratory and intermediate scales make it possible and practical to estimate if a BWT process and system is likely to be effective over the full range of physical and biological conditions expected in the field; whereas, the same regime on a ship would prove logistically and financially very unwieldy. Thus, smaller scale tests (a and b) demonstrate the treatment's performance and capacity across a wide range of relevant state variables, focusing especially on understanding the treatment process (mechanism), and larger scale tests (c and d) verify that the treatment system performs as expected at a realistic scale but over a subset of test conditions.

Two further aspects of scale reinforce the value of this step-wise approach to testing. Experimental testing at small scales is relatively inexpensive and provides the greatest control of experimental conditions. With increasing scale, the increase in cost and effort is evident. Perhaps less appreciated is the impact of scale on understanding and inference. In general, as scale increases, there exists a diminishing ability to control environmental conditions, to examine interactions across multiple conditions, and to achieve high levels replication (Table 2). All of these factors affect the strength of inference about treatment effects, reducing statistical confidence in characterizing the outcome. At the extreme in this spectrum is full-scale testing on operational ships, where environmental conditions and biotic communities in each ballast tank and voyage differ (independent of ballast treatment), and it is therefore not realistic or feasible to replicate experimental tests under identical conditions at this scale.

In considering this step-wise approach to testing, the entry point for initial tests will depend to some extent on the mode of action and what is already known about the performance of particular treatment systems. While there exists a compelling reason to complete extensive testing and proof of concept at the smaller scales (as discussed above), extensive background and knowledge may already exist for some treatments. For example, the dose-response relationships for some chemicals, and the efficacy of some physical removal (filtration) systems, may already be well defined, and this can be assessed from extensive literature review. If these include the relevant types of organisms and environmental conditions, it may be appropriate to initiate further testing at the mesocosm scale. In addition, some processes (e.g., filtration) may not be workable at small bench-top scale, requiring more extensive testing at the mesocosm scale. Nonetheless, the general approach and goals associated with different scales as outlined in Box C still apply.

Table 2. Trade-offs between scale of the experiment and various attributes of the experiment.

Experimental Attribute	**Scale of Experiment**
	Lab → Mesocosm → Ship
Control of Environmental Conditions	Decreasing
Control of Biotic Content	Decreasing
Ease of Replication	Decreasing
Time to Results	Decreasing
Cost / Effort	Increasing
Logistical Constraints	Increasing
Control of Treatment (dose-response)	Decreasing
Scale Appropriate Results	Increasing
Real-World Variation (ship x voyage x environment x biota)	Increasing

Experimental Design: A Scale-Independent Approach

Each scale of testing involves the same basic issues. First, it is critical to define the treatment characteristics, the response variables (including both the specific organisms and effects assayed), and the specific environmental conditions. These issues apply equally to any stage of testing, as outlined above. Second, general methods for experimental design and statistical analyses are well established for both biological and engineering applications (e.g., Sokal and Rohlf 1981, Chattergee and Rao 2003) and should be employed. Although the question of BWT effects is a specific one, the underlying principals and approaches for ecological and engineering research are directly relevant and necessary to draw robust conclusions.

In this section, we focus explicitly on the elements of experimental design for testing BWT. Although the same fundamental issues apply from laboratory to shipboard testing, the larger scales present some additional challenges associated with the potential variation across ship types, ballast tank configurations (associated especially with structural/spatial heterogeneity), and range of operating conditions. Thus, such sources of variation should be explicitly considered in engineering design as well as testing, as they may clearly affect treatment performance.

Control Treatments. A fundamental element of any experimental test includes the use of one or more control treatments, which control for temporal and spatial changes that are independent of the experimental treatment. For example, it is clear that ballast water communities are dynamic, where many species exhibit large changes in abundance and condition over time. Most studies have found large declines in the abundance of planktonic organisms in untreated ballast tanks over time (e.g., Lavoie et al. 1999, Gollasch et al. 2000, Wonham et al. 2001, Levings et al. 2004, Verling et al. 2005). In the absence of appropriate control treatment(s), such changes in abundance or survivorship of a species may be erroneously attributed to the experimental treatment. In other words, the effect of a treatment must be assessed relative to changes that occur in the control treatment, in the absence of the BWT.

In its simplest form, a controlled experiment to test the effect of BWT at any scale should include at least one control treatment that is run concurrently with one or more experimental treatments. Ideally, the control and experimental treatments should be:

- Identical in all respects (container characteristics, source and characteristics of organisms, environmental conditions, etc.), except for the application of the ballast water (experimental) treatment being tested.
- Initiated to run concurrently to prevent (control for) any variation in response associated with time; although it may not be feasible to run all treatments concurrently, experimental methods, such as block designs, should be used to minimize uncontrolled variation over time.
- Sampled to measure at least the initial and final conditions for the dependent and independent variables (see below). Despite the intent to control for initial conditions across treatments, differences may still exist, especially at larger scales. Thus, initial condition measurements provide confidence and allow direct comparison of the effect of specific treatment conditions on organisms present.
- Replicated to measure variation observed in response to treatment.

Response (Dependent) Variables. The effect of ballast treatment is measured generally by the total number or change in the number (or percentage) of viable organisms following treatment exposure, relative to untreated controls. On first glance, the effect of treatment on survivorship appears a rather straight forward and unambiguous measure, quantifying the number of living organisms that remain under the specific treatment conditions. However, there are several dimensions that require explicit consideration and careful interpretation surrounding (a) selection of species included and (b) measuring viability for particular types and multiple life stages of organisms; these are discussed in subsequent sections below (see Selection of Test Organisms, Measuring Viability of Organisms).

Treatment Characteristics. It is critical to measure the treatment characteristics at the time of testing as the independent variable(s). Although most treatments are based upon some theoretical application of the treatment agent, actual delivery may differ, introducing error in the relationship between treatment and response. This may be especially problematic when operating full-scale or shipboard testing, due to the structural heterogeneity and potential interactions with many uncontrolled variables associated with particular ballast system designs or ship operations. For example, the actual BWT process (such as a chemical) may be unevenly distributed in an actual ballast tank, because of uneven mixing or interaction with particulate (bottom and suspended) and dissolved materials. It is therefore necessary to measure the actual treatment levels delivered at each scale of testing, including consideration of spatial variation especially at larger scales.

Treatments should be characterized by measuring the treatment magnitude (e.g., dose, concentration, etc.) and duration. Time is an obvious key element in any evaluation, simply because effects will vary with duration of treatment. In some cases, the treatment agent itself may change through time. For example, concentration of a biocide may (a) increase over time through addition and mixing with ballast water and (b) decrease over time after addition has ceased. The temporal dynamics of the biocide (concentration x exposure time) may greatly influence the treatment outcome (survivorship). Thus, understanding the effects of many BWTs that affect ballast water chemistry will obviously require sufficient and detailed knowledge of how the treatment agent is distributed temporally as well as spatially.

Where relevant, the creation and fate of chemical by-products that may be toxic should be considered in this context for two purposes. First, some of these by-products may themselves influence survivorship, and tracking their fate may add interpretative power to the results. Second, since the treated ballast water is to be discharged, the potential risk of both the treatment agent and its by-products should be addressed (see Residual Toxicity).

Environmental Conditions. There are many reasons to expect the performance of BWTs to vary with environmental and biological characteristics of ballast water, which will differ among source locations and seasons. For example, performance of a chemical agent or ultraviolet radiation may be influenced greatly by temperature, salinity, pH, turbidity, and biomass. This may result from changes in organism sensitivity across environmental conditions, or interactions between the treatment agent and the environment that affects dosage experienced by the organisms.

Recognizing the potential importance of environmental and biological characteristics to treatment performance, a comprehensive testing program should explicitly address the full range of conditions to be encountered in the field. Obviously, understanding the mode of action and its limitations serve to define the expected range and scope of testing for effective operation of a BWT system.

As a general guide, test conditions should evaluate the effects of high and low levels of several key environmental conditions, which are known to affect sensitivity of organisms as well as the performance of particular treatments. These include temperature, salinity, turbidity / sediment load, organic carbon, and biomass. Each specific treatment process will have its own set of applicable state variables that will affect performance.

Importantly, there is the potential for interactive effects among environmental conditions. For example, a treatment may perform well under many high temperature or high biomass circumstances but not when both conditions exist. The number of permutations among possible environmental conditions (2^n) is daunting, especially with multiple tests associated with different organism types (see Selection of Target Organisms). Any treatment evaluation should consider the potential interactions, possibly choosing particularly challenging combinations based upon the specific technology, including knowledge about its mode of action and perceived limitations.

As with the dependent and independent variables discussed previously, measurement of environmental characteristics (covariates) should be included with experiments. This serves not only to verify and quantify existing conditions at the time of testing, but also to empirically test for expected outcomes and potential interactions.

Selection of Test Organisms

Which organisms should be used to test the effect of treatment? While this remains a topic of much discussion, the overall goal should be to include the full range of organism types present in ballast tanks. Initially, this is often considered as organisms from different taxonomic groups, such as copepods, dinoflagellates, and bacteria. In addition, some consideration should be given to habitat (location within tanks), life history, and behavior of the organisms across taxonomic groups.

The ballast community can be classified in general terms across several different categorical axes, each of which may affect either exposure or response to treatment. Here, we describe briefly these various categories:

- **Taxonomic Group** – Organisms from most major phyla have been reported in ballast tanks, and these groups may respond differently to treatment, due to differences in body size or a number of other attributes.
- **Life History –** Aquatic organisms exhibit a diverse range of life stages and population dynamics, which can influence treatment effects. For example, some species of zooplankton and protists form resting stages or cysts, which can tolerate a wide range of environmental conditions compared to vegetative (non-cysts) stages of the same species. Some organisms reproduce or form resting stages under stressful conditions, and others (such as bacteria and protists) have very short generation times that may allow regrowth during or after treatment. While many of these life history attributes may be restricted to particular taxonomic groups, others are not, and there is certainly a range within each group.
- **Behavior** – Organisms also differ in behaviors that can affect treatment results. Examples of such behaviors include vertical migration in the water column, attraction to surfaces, or burial in sediments. The extent to which such behaviors affect treatments is most likely associated with habitat-specific differences in the distribution of a treatment agent and its interaction with local environment (see Habitats below).
- **Habitat** – The ballast tank can be divided into at least 4 different habitat types, where organisms reside, including (a) water column, (b) at the bottom-water interface, (c) within bottom sediments, and (d) attached to hard vertical or horizontal surfaces. Treatment effects may vary across these habitats and their respective organism types [i.e., planktonic, epibenthic, infaunal, and sessile organisms (including biofilms)]. Variation in treatment responses across habitats may result from uneven distribution of treatment agent, including its

interaction with a local feature (e.g., sediments or surfaces) and environmental conditions, or differences in the tolerance/response of organisms associated with various habitats.

There is considerable interest in using selected model or surrogate organisms that can serve to provide a robust measure of treatment performance, representing the full spectrum of organism types (species x life histories x behaviors) and attributes outlined above. Moreover, if the use of model organisms could be standardized among tests, it would also allow direct comparisons of results across a range of environmental and operating conditions, both within and among treatments.

An obvious challenge in advancing this strategy is in identifying species that can be considered "representative" of such a broad range of organisms, including all types that are available for transport in coastal waters throughout the world. Some programs are currently conducting research to identify such surrogates, exploring the use of taxa that are particularly hardy and tolerant across a wide range of conditions (Hunt et al. 2005, more refs). An additional premium exists for organisms that have the following attributes: (a) easily and economically cultured in large numbers, (b) demonstrate a reliable and consistent response to treatment across culture batches, and (c) are hardy enough to withstand routine handling associated with addition to treatment systems, residence in control tanks, and routine sampling. Overall, the use of surrogates assumes that these organisms are among the least susceptible to treatment, such that if they succumb then most other organisms would be eliminated as well. Obviously, this assumption must be validated before any specific surrogates can be used.

The use of surrogates is a pragmatic and necessary strategy to address the overwhelming diversity of organisms available in ballasted water, but we also urge caution in relying on a very narrow range of organisms and extrapolating results. Clearly, there is more research needed to identify surrogates for ballast technology testing. Even as this progresses, there are additional steps that should be implemented to assure robust results:

- <u>Several taxonomic groups, species, and life stages should be included as surrogates during treatment testing</u>. This protects against the likelihood that there is an interaction between treatment type, environmental condition, and surrogate species performance. In other words, selected surrogates will be more or less susceptible to different treatments, and it is possible that the same species is not appropriate across all treatment types.

- <u>Multiple geographic regions should be included as a source for surrogates and testing.</u> Just as environmental conditions may affect treatment performance, such conditions will certainly affect the response of organisms to treatment due to (a) acclimation and condition of organisms at the time of treatment and also (b) the unique evolutionary histories and physiological tolerances of organisms from different regions. .Presumably the use of surrogate taxa in laboratory settings would control for acclimation and condition, possibly using standardized laboratory cultures of an identical strain. Yet, the expected range of responses may differ geographically. For example, treatment effects may differ greatly between tropical and temperate organisms, Pacific and Atlantic ocean basins, etc.

Thus, despite a strong rationale for using surrogate taxa, the knowledge required to confidently select particular taxa and source regions that are sufficiently robust indicators is incomplete. Adding some redundancy in surrogate taxa and source regions is one approach to minimize the risk of spurious results, whereby effects on one set of surrogate taxa misrepresent results under certain conditions or groups of organisms. When considering surrogate taxa, there are a large number of potential combinations to consider. For this reason, extensive testing should be performed at the laboratory and mesocosm scales to include a wide range of surrogate species, life histories, habitats, source regions and environmental conditions, as outlined above (see Scaling Up). This argues for the routine use of several different laboratories, which are distributed geographically to provide access to different sets of organisms for these early stages of testing.

Although full scale test bed facilities are intended to focus on a narrower range of organisms and the system engineering design, these facilities also serve an important purpose in further evaluation of effects on selected organisms. First, using highly controlled conditions and full scale, it is possible to examine effects of behavior and life history (population growth or cyst formation) in treatment performance. Second, it is also possible to examine fine-grained habitat differences in the distribution of a treatment agent and/or its effects across in-tank habitats, as well as effects of flow variation on in-line treatments

As with treatment testing at smaller scales, it may also be necessary to establish full scale test bed facilities in multiple global regions, providing access to a wider diversity of organisms and environmental conditions than is cost-effective at one location. Much discussion surrounding such test bed facilities has focused on examination of surrogate taxa, providing an important level of repeatability, control, and replication not possible with communities from the field (wild). Distributed test facilities could serve the purpose of testing such surrogate taxa from different regions, thereby preventing the potential accidental introduction of surrogate organisms to ecosystems where they do not naturally occur. Importantly, use of different sets of surrogate species at different test sites will require careful "calibration" between the various regional sets of species, to minimize any confounding factor and maintain as great a degree of comparability among test sites as possible. Regional test facilities could also provide a first opportunity for whole community "challenge experiments" at a full scale, which would test how robust the system is to an entire wild community. Comparability of efficacy between surrogates and ambient organisms would provide added confidence that (a) surrogate taxa are broadly representative and (b) there are no unexpected interactions when including uncontrolled variables from the field.

Shipboard testing provides a critical step in evaluating real-world performance of test systems, including the full spectrum of ballast organisms. This also serves the purpose of "challenge experiments" using wild communities, including all normal vessel operations. While this may seem to obviate the need for challenge experiments in full scale test bed facilities, it is also much more complex, time-consuming, and expensive to conduct such experiments aboard ships. Further, test beds provide more opportunities for control and monitoring of state variables. and will likely provide higher-resolution data than is obtained on ships, particularly those engaged in commercial activities. Thus, there is clear value in time, expense, and detail to having some initial challenge experiments at the test bed level, providing quick feedback and confirmation of system function, in addition to any shipboard measures.

Measuring Viability of Organisms

The goal of BWT is to greatly reduce the number of viable organisms that are discharged with ballast water that can colonize receiving waters, and thereby minimize the risk of invasions. Some treatments work by physically removing organisms or causing lethal physical destruction, and others cause mortality without conspicuous physical damage. Regardless of the mode of action, quantitative measures of treatment effects require classification of residual organisms into viable and non-viable portions. Thus, the need for such viability assays is common to all treatment types.

In the absence of conspicuous physical damage that is clearly lethal, assays of viability can be difficult or cumbersome across a wide range of taxonomic groups. This is manifest in a variety of ways across different groups:

- For many zooplankton (including both holoplankton and meroplankton), it is possible to classify individual as live or dead, based upon movement or response to physical stimuli ("poking"). This is a very cumbersome and time consuming approach, requiring that samples be analyzed immediately upon collection, since non-treatment related mortality can begin almost immediately upon collection (removal) from the ballast tank. Such approaches to assaying viability can easily be a limiting step in the scope of experiments and associated replicate sampling. Further, this procedure is inherently difficult to standardize and requires calibration to verify that individuals scored as dead are really so.
- For vegetative stages of dinoflagellates and other protists, it can be very challenging to classify live versus dead individuals. Analysis of physical integrity of preserved cells and internal structure under a compound microscope can distinguish compromised from intact cells. This provides a minimum estimate of treatment effects, but there is often uncertainty about the viability of intact cells.
- For bacteria, viruses, as well as resting stages / cysts of multiple groups, viability can also be challenging to assess. Direct counts can provide a quick and quantitative measure, but this measure can clearly underestimate treatment effects, as some non-viable organisms are included.

Growth is sometimes used as an assay of viability for bacteria, protists, and resting stages. While this can be very useful in demonstrating viability, it does not allow adequate classification for the residual organisms that do not grow. Many of these organisms require specific conditions for growth, and some can remain relatively inactive despite the presence of seemingly appropriate conditions. For example, a large proportion of bacteria in marine waters are thought to be in a "viable but non-culturable" state. Although there is uncertainty about the fate of these bacteria, it appears that some remain viable and able to undergo population growth, even though they are not readily cultured under laboratory conditions. Thus, absence of growth in laboratory conditions does not demonstrate absence of viability.

Vital stains can be used to provide additional information for some groups of organisms. For example, it is possible to stain for the presence of DNA or other attributes. While absence of a critical structure or attribute can identify non-viable (dead) individuals, providing a minimum estimate of treatment effects, some recently dead organisms may appear similar to viable cells. At the present time, there is considerable uncertainty about the use of stains to distinguish live from recently dead (and undamaged) zooplankton, protists, and bacteria. Certainly some stains are used in this fashion, especially for cellular-based assays to characterize the integrity of cell walls or the presence of particular enzymes. However, these are very time consuming, and a robust validation of performance (efficacy) of stains is often lacking for many taxonomic groups.

Assays of metabolic activity are sometimes used to characterize the status of organisms such as protists, but these are usually bulk measures applied to community assemblages that include many species. Examples include respiration, uptake of isotopic carbon or nitrogen, or activity of photosynthetic systems for autotrophs. While these are very informative measures for particular purposes, they have not been standardized to function as bulk measures of viability, normalized to density and biomass for single species, much less for multiple species assemblages (the latter of which likely exhibit different rate functions).

As with stains, many molecular techniques are designed to measure the presence or absence of particular characteristics but may not be very informative about viability. For example, polymerase chain reaction assays (PCR) can be used to detect particular DNA sequences, and techniques for quantitative PCR are emerging. While this may be especially useful for detecting copy number of a particular sequence, and may be related to number of organisms present (especially for known species or surrogates), the technique does not distinguish between viable and recently dead (non-viable) individuals. Fluorescence in-situ hybridization also offers the ability to quantify reasonably intact (whole) bacteria and protists, but this measure may not be diagnostic of viable organisms.

There are clearly many issues to consider in evaluating the number of viable organisms associated with ballast treatment experiments, and there is a critical need to improve upon current methods available. Certainly the selection of surrogate taxa may be influenced by the feasibility for assessing viability, and advancement of methods may focus on these surrogates.

Despite existing limitations, it is still possible to collect meaningful information, so long as significant attention is given to clearly identify and interpret within the limits of particular methods. In many cases, this would result in a count of potentially viable organisms, whereby reasonably intact (whole) organisms with key characteristics are considered viable but include some non-viable fraction that is recently dead. This is a precautionary approach and provides a conservative (minimum) estimate of treatment effects, which will undoubtedly increase as methods improve the capacity to discriminate viable from non-viable populations. In contrast, grow-out and culture assays may produce a maximum estimate of treatment effects, by estimating a minimum number of viable organisms, and we urge particular caution in interpreting these results.

Sampling, Variation, & Replication

Core components of any experimental test program are sampling design and the methods for sample collection, including especially the level of replication both within and among experiments. Previous sections have focused on general experimental design and sample analysis, but they have not addressed sample methodology or allocation of effort, which are of critical importance to adequately characterize the response of organisms to treatment.

The primary goal of sample collection is to obtain representative samples that characterize the dependent or independent variables in an experiment. The combination of sampling method and sample distribution should be designed to estimate the local conditions and spatial variation in those conditions, respectively, for each time period of sampling. For example, the water in a large tank (e.g., full scale test bed or ballast tank) may be sampled by pump through a hose at multiple locations in the tank. The individual sample is sieved to provide a local estimate of organism abundance and number viable, and physical and chemical characteristics of the water sample are also measured. During the same sampling time period, this is replicated in multiple locations to measure variation of these attributes within the tank, providing a quantitative estimate of average conditions and spatial variation.

Obviously, the sampling strategy (method and distribution) should be designed explicitly to provide representative samples and test for potential spatial variation that may exist. In the example above, a pump sampling method must assure that the suction is sufficient to overcome any avoidance or escape response by motile organisms, otherwise bias may exist in the proportion of dead organisms sampled relative to the actual population. The dispersion of replicate samples should be stratified vertically and horizontally to include areas that may differ in organism abundance and status (live versus dead) or associated chemical/physical environmental conditions. Additional consideration should be given to organisms that may be attracted to structure (tank walls or beams) or colonize the bottom of tanks.

In similar fashion, samples collected from a ship's ballast water system (pipe) during discharge must be done in a manner that assures a representative sample is drawn from the passing water. The orientation, position, size, material, and shape of such an in-line sampling port, as well as the pressure at which samples are drawn, may affect the nature of samples. Instead of spatial variation in samples, such in-line sampling should include temporal replication, allowing discrete samples to be drawn at different time points during the discharge process. It seems likely that any spatial differences that exist within the tank may be manifest as differences in the discharge characteristics over time.

Although we have focused on a large scale for illustration, and it is likely that any spatial differences are more pronounced at larger scales, the same issues deserve explicit consideration at smaller scales (i.e., laboratory and mesocosm scales). Bias associated with sampling method and spatial variation can exist at any scale. It is often assumed that tanks are well-mixed, especially at small scales, but there are many reasons why this need not be so, including behavior of organisms or an uneven distribution of the treatment agent. Some bacteria and protists are known to exhibit strong phototaxis or geotaxis, such that populations that are well-mixed initially can change very quickly.

As part of implementing the overall experimental design (see Experimental Design), issues of replication, independence, and inference must be considered carefully and explicitly. In general, each test conducted in an individual tank, mesocosm, or beaker is the level of experimental replication. For example, an experiment that consists of a treatment and a control (untreated) tank represents one replicate measure for each condition. Multiple samples taken within each tank (within or among times) cannot be considered independent measures and are used solely to measure variation and estimate mean conditions within a tank. Results can certainly be compared among tanks, as planned, but these cannot be attributed to treatment effects without replicating the experiment. Replicate experiments provide a measure of variation observed in the relative treatment effects, allowing a test of whether consistent effects exist statistically and the level of variation that exists among independent tests.

Ideally, control and treatment tanks are run concurrently, to remove any effect of time on responses. Another approach is to alternate these two treatments, replicating the experiment multiple times, to remove time as a factor. When two different tanks are used (control and treatment), replicate experiments should be performed by alternating treatments among tanks, to control for effects due to tank differences (rather than treatment differences). Importantly, if treatment involves the use of chemical that may persist in the tank following ballast discharge, this will greatly complicate the experimental design unless the tanks can be adequately cleaned or chemicals deactivated between experiments. In some cases, it may be difficult to adequately clean tanks (neutralize treatment agents) between experiments. The use of replicate tank pairs (control and treatment) can generally provide a useful approach to assure that observed effects result from a treatment instead of tank-specific differences or artifacts, which may exist independent of treatment. For this reason, replicate tank pairs should be considered wherever feasible.

There is no set rule about the number of replicate experiments required at any scale (laboratory to shipboard testing). In general, while 3-5 replicate experiments are often deemed sufficient to evaluate performance under each specific set of test conditions, where high variation is observed, more replication will likely be required to both provide a better measure of central tendency and to understand factors that may contribute to such variation. Moreover, at the shipboard testing stage, further replication may be especially important, because (a) there are many uncontrolled variables (i.e., unique sets of conditions) that occur on each voyage, and (b) any change in treatment operation over time is itself an important issue for a fully integrated system (see Other Considerations below).

Residual Toxicity

In addition to demonstrating that a BWT performs successfully in eliminating organisms, there are also regulations to consider for any system that will introduce chemical residue to the environment with discharges of ballast. Thus, proposed use of any chemical must often include evaluation at some level of the potential toxicity of discharged treated water to organisms in harbors, bays, and other receiving waters. These requirements will vary by state or country and require careful attention, as they may delay or preclude the use of some treatments.

In the United States, there are several federal regulations that may apply to discharge of chemical residuals or by-products associated with BWT. The Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) created a regulatory framework for biocide distribution, sale, and use, including for experimental purposes. The U.S. Environmental Protection Agency (EPA) has oversight of FIFRA and has established rigorous requirements for approval of biocides, including extensive testing to evaluate environmental fate and lingering toxicity (http://www.epa.gov/pesticides/regulating/index.html). This can be a lengthy and expensive process, and therefore should not be overlooked in developing and scheduling a test plan for any treatment.

In addition, as authorized by the Clean Water Act, the EPA also regulates point source discharges to the environment more broadly (beyond the scope of FIFRA) through the use of National Pollution Discharge Elimination System (NPDES) permits (http://cfpub.epa.gov/npdes/). In March 2005, a ruling in U.S. District Court concluded that EPA inappropriately excluded ships' ballast water from considerations under the NPDES system. The court's remedy solution is pending and, depending upon the outcome, potentially has significant implications for BWTs. In general, the NPDES permit process requires demonstration of "no adverse effects" as evaluated through chemical specific parameters and standardized Whole Effluent Toxicity (WET) testing, experiments designed to evaluate any residual toxicity on a variety of organisms.

In general, treatments that involve a chemical mode of action must consider federal and state regulations that may apply to ballast discharge and the associated risk of exposure for organisms in the receiving waters. While one component in such analysis is the chemical concentration upon discharge, exposure risk will also depend upon dilution, circulation, and degradation of the active chemical agents (and by products). Thus, a complete analysis would include when, where, and how much chemical can be discharged to avoid deleterious effects.

Selecting Vessels and Routes

Ships exhibit a tremendous range in physical and operational characteristics, such that not all ships are equal with respect to testing BWTs. Ships are often classified among many different types, including general categories such as tankers, containers, bulk carriers, passenger vessels, and many others. These general categories can be further subdivided into many designations, based upon the size, design, and type of cargo transported. Some schemes for classification include well over 50 such vessel categories and subcategories. Importantly, there is considerable variation both within and among these vessel types in the physical capacity of ships to carry ballast water, and this is manifest in the number, shape, location, and volume of ballast tanks as well as the size and rated capacity of ballast water pumps, the extent and size of ballast system piping, and the manner in which ballast water is taken on or discharged

In addition to differences in ballast water capacity, ships also display a tremendous range of operational patterns, which greatly affect the actual quantities of ballast water used as well as amount and frequency of discharge. Figure 3 highlights quickly some of these differences in ballast capacity and management among several vessel categories, including:

(a) a large (several fold) difference exists in the mean ballast water capacity across vessel types;
(b) on average, ships in ballast do not operate at capacity, as the mean total ballast on board (TBOB) is considerably lower than the total ballast capacity in all cases;
(c) mean discharge volume is lower than mean TBOB, indicating that ships do not discharge all of their ballast upon arrival.

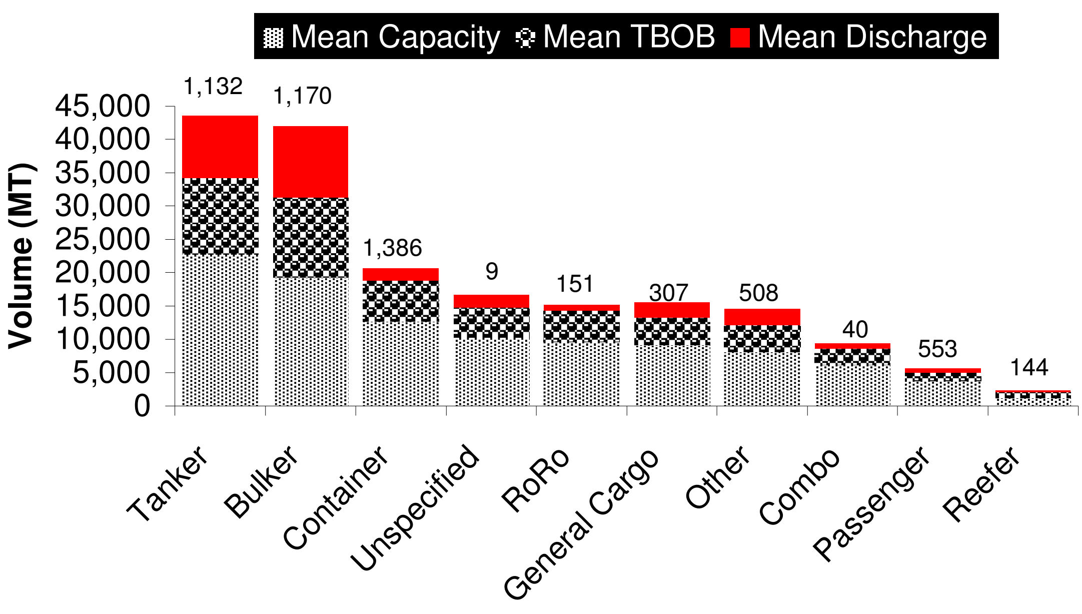

Figure 3. Mean Per Vessel Volume Measures by Ship Type for Arrivals to U.S. Ports from Overseas. Shown by vessel type are the mean ballast water capacity, mean total ballast water on board (TBOB), and mean ballast water discharged upon arrival. Volumes are shown in metric tons (MT) as reported to the National Ballast Information Clearinghouse from July-December 2004 (n = 5,299 arrivals by discharging vessels, and sample size by vessel type shown above each bar; AW Miller, unpublished data).

Although Figure 3 serves to illustrate some coarse level differences among vessels, it does not fully characterize the observed variation in ballast water management and operations. First, there are large differences in ballast water management associated with route and port. For example, oil tankers arrive to Port Valdez, Alaska at near-capacity in ballast water to be discharged, but carry and discharge very little ballast from Alaska to their primary destinations in California and Washington (Hines et al. 2000). A similar pattern exists for bulk carriers arriving in ballast to Chesapeake Bay from the northeast Atlantic, carrying mostly cargo in the opposite direction (Smith et al. 1999). Second, port visits vary greatly among ships, both in terms of residence time as well as predictability of route and schedule. For example, containerships often have relatively low residence times in port (often 12-24 hours) and high predictability. Some ships may spend days in port and also routinely identify or alter next port of call only days before arrival. Third, vessels obviously exhibit a wide range of routes, affecting the sources of ballast and voyage duration as well as organism content and survivorship (Lavoie et al. 1999, Verling et al. 2005).

The fact that such major differences exist in ballast water management among ships is perhaps no revelation, and has been discussed in several publications (Carlton et al. 1995, Minton et al. 2005, Verling et al. 2005), but this has important implications for evaluation of BWTs. The attributes of physical and operational constraints define what is feasible on particular vessel types, in terms of what technologies may be most appropriate. For vessels with relatively low ballast volumes (e.g., passenger vessels or container ships), some treatment technologies may be possible that are not feasible on vessels with larger volumes. Likewise, some short duration voyages may preclude the use of treatments that require more time than available to perform properly and achieve desired results.

Some vessels are also simply more amenable for testing particular treatment technologies, even if a treatment is broadly applicable to a wide range of conditions. Criteria for vessel selection include not only permission to work on a vessel but also conditions that maximize the quantity and quality of data from testing, including:

- regular and predictable ballast operations;
- predictable route(s) for planning and implementing tests;
- routes with weather patterns and conditions that can accommodate a work plan;
- sufficient port residence time and voyage duration to implement work without compromising replicate sampling, time course, etc;
- access to multiple tanks, allowing comparison of at least treatment to control tanks;
- ready access for tank and/or inline sampling with sufficient replication;
- adequate space for equipment and personnel;
- good communication (including ability to overcome language and cultural differences) and coordination with the vessel owner, captain, and crew.

While these criteria seem very straight forward, we are aware of many cases where vessel operations or conditions have greatly limited the intended scope of testing. This stems in part from difficult logistics associated with the selected vessels (in terms of route, time, or access) as well as the priorities and constraints for commercial vessels. Vessels have as their first and foremost priority the movement of cargo in an efficient, timely, and safe manner. For the ship, there is a clear premium on minimizing the time in port and transit. Ships are also often short-handed, allowing little time to assist with ballast treatment tests. Thus, although vessel operators may be highly supportive of research and testing, it is not their first priority and is accommodated to the extent possible. For this reason, every effort should be made to minimize the constraints for performing ballast treatment research and testing aboard vessels, by carefully selecting the most conducive platforms.

3. Current State of Ballast Treatment Testing in the U.S.

At the present time, no BWTs have been approved for use other than ballast water exchange, and development and testing of treatment technologies is an active area of research. The approach to testing has been highly variable, limiting the opportunity for comparisons among different technologies and regions. In addition, it is evident that some of the studies have provided only limited insight, due to problems with design or implementation.

In 2000, an audit of four large scale and shipboard tests of BWTs in the U.S. found that *"the experimental designs and analyses had fundamental problems whose extent and gravity undermined any conclusions that could be drawn about the treatment performance of any of the BWT systems"* (US Coast Guard 2004). This review found significant problems with experimental design, use of controls, replication, the type and method of measurements for dependent variables, and quality control. This does not appear unusual to these particular studies, as evidenced by presentations of other BWT research given to scientific meetings, funding agencies, and informal groups. The Coast Guard report concludes by recommending the development of general guidelines for scientifically sound test programs, including the role and use of experiments at different scales.

To some extent, the problems with past large-scale testing resulted from a confluence of several factors. Many early projects were advanced too quickly, sometimes by vendors or inventors with little experience in scientific design and biological research, without sufficient consideration of the best work platform and experimental design. In some cases, the treatment systems were installed before developing a specific test plan or consulting biologists, constraining what could be done. In other cases, full-scale treatment had not been tested before installation on ships, causing breakdowns or malfunctions in the treatment systems. For some, there may also have been a lack of appreciation about the full range of environmental conditions, organism types, global regions, and ship operating conditions that are relevant to this issue. In particular, a misunderstanding about the complexities may have caused some to seek a 'quick answer', where one does not exist, possibly limiting the necessary groundwork and scope of inquiry required.

In the intervening time, guidelines for ballast treatment testing have begun to take shape, particularly at large scales. Two different efforts are underway that together help drive a more standardized and formal approach to ballast treatment testing. First, the U.S. Coast Guard has entered into a partnership with the Environmental Technology Verification (ETV) Program of the U.S. Environmental Protection Agency to develop protocols for evaluating the performance of BWT systems. Second, the U.S. Coast Guard has developed the Shipboard Technology Evaluation Program (STEP), which establishes a formal process for evaluation of BWT systems on ships. We discuss each of these programs separately below.

The ETV Program

The ETV Program for BWT was initiated in 2001and has focused primarily on the development of protocols for testing technology at full-scale test bed facilities (see Tanis et al. 2004, Hunt et al.

2005). The immediate goal of this program as stated by Tanis et al. (2004) is to:

> "*verify the performance characteristics of commercial-ready treatment technologies with regard to specific verification factors, including biological treatment performance, predictability/reliability, cost, environmental acceptability, and safety*."

The ETV Program does not certify technologies but instead is intended to provide objective and high-quality data that can be used by all interested parties to evaluate performance. To date, the ETV program has focused on developing rigorous protocols for a full-scale, land-based testing facility. These protocols are still under active development by an extensive panel of technical experts and consider explicitly experimental design, environmental conditions, sampling methods, surrogate taxa, and viability measures (Hunt et al. 2005).

One goal of the ETV Program for ballast treatment testing is to create test centers, where vendors can utilize standardized and accepted protocols for third-party verification. These centers are intended to serve as standardized test platforms, which provide comparable conditions among test events and across facilities (locations), removing uncontrolled variables and idiosyncrasies that occurred across many previous experiments (see above).

The first ETV-related test facility has been developed at the U.S. Naval Research Laboratory (NRL) in Key West, Florida (http://ans.nrl.navy.mil/whyarewehere.asp) through a collaborative partnership between NRL and U.S. Coast Guard This prototype facility is actively being used to establish protocols, from sampling methodology to use of particular organisms and viability measures. In addition to implementing technology testing at this center (to begin in 2006), the facility serves as a pilot platform to advance and ground-truth methods that would be adopted at other test centers, and for ship-board testing as well.

Other test facilities are under consideration, but none have advanced beyond a general conceptual stage. Within the U.S., there is interest on the part of federal and state agencies to implement additional test centers, including some funds that are now devoted to this effort. For example, the National Sea Grant Program is supporting the design phase for such test centers, with the intention of subsequently supporting full-scale development of some centers through a peer-reviewed, competitive process (http://www.seagrant.noaa.gov/funding/rfp2006.html).

To our knowledge, development of similar facilities for testing BWT systems has not progressed as far overseas. In Norway, the Norwegian Institute of Water Research (NIVA) and the independent foundation Det Norske Veritas (DNV) are developing a large-scale test facility for BWT systems (http://www.niva.no), but little information about this effort is yet available. More broadly, there is surprisingly little coordination of ballast treatment testing at the international level. While extensive and on-going discussions occur at the IMO about developing ballast treatment standards and their implementation, there has been limited progress in developing an international framework or consortium to verify treatment performance (see next section).

Thus, while there is tangible progress in developing full-scale test protocols and one test center now exists, the ETV Program for ballast water testing is still at the pilot stages. The current test facility has provided a much-needed controlled and quantitative approach to identifying sources of error and improving protocols, some of which will no-doubt transfer directly to ship-based testing. Yet, full implementation will require additional work at this facility and the creation of other test centers, both nationally and internationally.

To date, the ETV activities have placed a premium on using land-based test platforms, to provide a high level of control for many of the variables (mechanical, biological, weather, design) that can affect performance. While a greater level of control can be achieved on land-based facilities, especially when designed explicitly for this purpose, there is also interest by some groups to add floating platforms (ships or barges) to serve as additional test bed facilities. Past experience on ships has shown some significant constraints exist, especially in terms of space, engineering, sampling design and environmental fluctuations. Thus, any such proposal to work on a floating platform must carefully evaluate the tradeoffs in data quality, repeatability, and costs.

The STEP Program

The U.S. Coast Guard established the STEP Program in 2004 and has updated accompanying information and application materials (Navigation and Vessel Inspection Circular No. 01-04; http://www.uscg.mil/hq/g-m/mso/step.htm). The purpose of this program is to assess the performance of BWT systems aboard operating vessels. Vessel owners/operators may apply to the STEP program, which allows the ship to operate an experimental ballast treatment system in lieu of ballast water exchange after being accepted to the program. The applicant must agree to conduct specific measures of system performance, including an initial experimental testing phase (experimental evaluations during the 1st and 5th years of operation) and monitoring phase (for the life of the treatment system). If accepted into the STEP Program, participating vessels are considered to meet any existing or subsequent ballast water discharge regulations up to the life of the vessel or the life of the treatment system.

To participate in the STEP, the operator must submit an application that includes very detailed information, including:

- Letters of commitment by all principals;
- System description that describes in detail the engineering, control, and expected performance;
- Environmental compliance assurances, indicating the treatment and discharge meets all existing environmental regulations;
- Class conformity, indicating the treatment system is acceptable to the classification society or Marine Safety Center;
- Prior experimental results, demonstrating the system's potential or performance to date;
- Study plan, outlining the experimental design, sampling methods, measures, etc. for the experimental and monitoring phase;
- Vessel characteristics, indicating the physical and operational attributes of the specific vessel;
- Route and Operation, indicating the usual route, ports, and ballast operations of the vessel.

A key aspect of this application process involves both the study plan and prior experimental results. The study plan relies upon a formal and explicit experimental design, using controls and including details of sampling, analyses, and replication. Moreover, the study plan for the 5-year experimental phase includes a sufficient time horizon and explicit directions to examine variation due to season, route, environmental conditions, and organism types. Another key aspect of the study plan examines any degradation in the system performance over time.

Although the focus of the STEP is on shipboard evaluation, there is an emphasis in the application materials on prior results at smaller scales. In fact, the supporting documentation underscores the importance and utility of extensive testing at small (laboratory or bench-top) scales to understand and optimize the treatment process, scaling up to a full-scale test bed before shipboard work. Although the STEP does not require specific protocols for prior experimental results at smaller (pre-shipboard) scales for consideration, the model provided clearly encourages a transparent series of controlled, quantitative, replicated experiments.

Thus, there is strong concordance with the guidelines outlined above (see Guidelines) and those illustrated in the STEP application materials. Indeed, the overall goals with experimental design and performance metrics are largely scale-independent. Both the STEP and ETV Program are guided by the same general approach, with an implicit intent to have the two programs (and stages of testing) working in concert, as end-members in the step-wise progress of increasing scales for treatment testing.

The STEP is now accepting applications, which will undergo review by an external expert panel, and expects to review up to six applicants per year. Although ships are not required to participate in the STEP for testing technologies, the program does provide some important incentives, including both (a) the granted equivalency (exemption) to ballast water discharge standards and (b) a formal evaluation process that includes both guidelines and external review process. The STEP clearly represents an important milestone in advancing shipboard testing, establishing benchmarks and a formal process to provide much needed high-quality data for evaluation by scientists, agencies, and policy-makers.

4. Conclusions & Recommendations

BWT evaluation is still at an early stage of development. While testing of some treatment processes and systems has occurred over the past decade, there have often been significant problems associated with experimental design, measurements made, treatment performance (especially at large scales), and strength of inference. These problems may have resulted partly from a rush to market (sell) treatments, as there is a strong commercial incentive involved, without full appreciation of the level of scientific rigor required in evaluation and the associated complexities. Because testing of treatments has often occurred at large scales (and aboard ships) before much of the groundwork has been established at smaller and land-based scales, there are significant gaps in understanding about the modes of action, treatment performance, and engineering design of the treatment system.

There is a clear need for standardized approaches to evaluating treatments that explicitly consider the experimental design, environmental test conditions, sampling and analytical methods, and replication. This serves several purposes, including (a) establishing minimum criteria that must be met to adequately explore the performance of treatment systems across a full range of conditions and (b) creating the opportunity to compare performance across systems. Such standardization need not be prescriptive, as the basic components should include general principles applied routinely in ecological and biological experimental research.

In addition, there is a strong incentive to implement a step-wise approach as part of a standard approach to testing BWTs. This graded-entry allows initial testing to proceed quickly and inexpensively at a small (laboratory) scale, providing the critical empirical basis to evaluate performance of the treatment method under various conditions. It is at this scale that a solid understanding of the mechanism (mode of action) should emerge, defining the interactive effects of dose, time, and environmental conditions on various types of organisms. Not only is a scaled approach more efficient in time and resources, but performance at larger scales (including shipboard) will also be enhanced by refinements and knowledge of system engineering that result. Thus, while there is often a drive to place treatment systems on ships and initiate an advanced stage of testing as soon as possible, an assessment should first be made to critically evaluate whether this is the most efficient and appropriate level of testing or whether smaller-scale testing would be more advantageous based upon the pre-existing empirical data.

We have outlined general guidelines for ballast treatment testing, outlining several key issues for consideration in experimental testing and their application through a step-wise series of scales, including laboratory (bench-top), mesocosm, full-scale test bed, operational ships. Although there was a clear and strong consensus among our group in this approach, we surmise that this framework and underlying rationale has not been presented to a broader audience, including those involved in testing ballast treatments within the U.S. and overseas.

It is evident that U.S. Coast Guard has adopted this framework in advancing the ETV Program and the STEP. Much of the same rationale about the need for a rigorous experimental approach is reflected in the program descriptions. In addition, the utility of a step-wise strategy to scaling up in tests is also highlighted in the STEP application process, which places a strong emphasis on the value of empirical data at small and land-based scales as a precursor to shipboard testing. While the role of scale is implicit in the STEP application materials, and was discussed in previous audits of early shipboard testing projects (USCG 2004), we intend the current document to more explicitly formalize the sequential nature of tests at multiple scales as a general framework for ballast treatment testing.

While the creation of the ETV Program and the STEP represent significant progress in defining a clear process to evaluate and approve BWT systems, as endpoints in the step-wise testing framework, this effort is just now being implemented. Only one test center exists in the ETV Program, and it is just now about to initiate its first tests on a treatment system. The STEP has received applications but has not yet approved a treatment system for a ship. This represents clear forward momentum, but the rate of progress will depend greatly upon continued work to resolve several fundamental issues in testing as well as expansion of the program to include (a) additional ETV test centers and (b) international participation.

We wish to underscore the benefits of advancing an international strategy to BWT testing. With the ballast water management regulations that now exist in several countries, the recent advance of the IMO convention (see Box A), and the global scale of shipping, a concerted international effort geared at testing would pay many dividends. A shared network of testing facilities would serve to (a) distribute the cost and effort, (b) guarantee that results are comparable and additive toward the common goal, (c) facilitate development of diverse surrogate taxa and testing sites from different global regions, thereby increasing the robustness of results, and (d) increase the number and scope of possible treatments being examined, overcoming the limited output that exists a the single test facility. The science capacity exists in many countries to implement such a concerted effort, which would greatly accelerate the advancement of promising treatment systems and analytical methods, to mutual benefit. Thus, international coordination and collaboration deserve serious attention, may provide important strategic opportunities, and should be pursued to the extent possible.

Some initial steps toward developing an international framework to evaluate treatment systems now exist. In July 2005, IMO adopted guidelines for approval of ballast water management systems, which closely track those of the STEP program. These guidelines currently include two approved components, referred to as G8 and G9 (IMO 2005a,b). G9 outlines an approval process for the use of active substances, such as biocides, to characterize the risks associated with their use and discharge into the environment, providing basic approval for use. G8 outlines a process to evaluate the full scale land-based and shipboard performance of ballast treatment system to remove organisms and meet discharge standards under the IMO Convention.

Despite emergence of these IMO guidelines, there is a surprising lack of international cooperation in testing BWT systems. The U. S. Coast Guard and the Naval Research Laboratory have initiated preliminary discussions with test organizations in several other countries, including the Netherlands, Norway, and Singapore, but as yet no formal arrangements for collaboration or coordination have been established (R. A. Everett, personal communication.). We are presently unaware of any other activities among countries to advance internationally coordinated test facilities or testing programs.

Beyond creating a broader framework for evaluation of ballast treatment systems, a significant amount of research is still needed to examine the biology and sampling surrounding test organisms. Among the highest priorities, we recommend further research in the following areas:

- Viability of organisms. It is evident that reliable measures of viability are not well established or accepted for several types of organisms, yet measurement of viable organisms is the key dependent variable required to assess treatment performance. The use of live/dead stains has the most appeal, if these could be applied to samples and allow for subsequent enumeration without significant time pressure, as exists for analysis of live samples.
- Surrogate organisms. This is still an active area of research with important questions about which organisms are most "representative" and under which conditions. Moreover, it appears important to include surrogate organisms from different environments and global regions, as may have different tolerances. As a minimum, surrogate organisms should include taxa from high and low salinity environments, and temperate and tropical climates. The use of diverse and possibly different surrogate taxa among several test facilities also has important implications for standardization, creating the need for inter-calibration among different surrogate species and test facilities.
- Sample collections. While there are methods for sampling most organism types from open tanks, there remain significant gaps in considering sample collections from ballast tanks, including various habitats (e.g., water column versus benthic communities) and sufficient internal replication. Such partitioned sampling is likely to be critically important for understanding and refining in-tank treatment processes. In addition, a significant premium may be placed on in-line sampling of ballast water upon discharge or intake. Such in-line sampling may be the most efficient way to obtain replicate samples that integrate space and time, and it provides a direct measure of discharged materials, which is the basis for regulations. A significant amount of work remains to be done on the engineering design, modeling the fluid dynamics and behavior of particles (organisms) of different types, and validation of performance of in-line sampling mechanisms, to provide independent and representative samples.

There are also still several programmatic decisions to make in implementing shipboard ballast treatment evaluation and compliance testing. We have outlined a scale-independent approach to experimental design, including the use of controls on shipboard testing, as this provides the best scientific approach to estimate the effect (efficacy) of treatment. However, there are practical issues that are advanced as counter-arguments against this approach. First, it may be restrictive to maintain control and treatment tanks on ships, due to operational constraints of treatment systems and also the ability to handle untreated ballast water that cannot be discharged near ports. Second, compared to smaller scale experiments, shipboard testing is less about measuring effects on biota than measuring the delivery of a particular treatment agent with a known mode of action and established dose-response relationships (Box A). Third, use of controls and biological measurements on shipboard tests greatly increases the effort and complexity.

Instead of using control tanks, one alternative approach for shipboard testing is to measure concentrations of viable organisms in treated ballast water at discharge. Although this approach can certainly assess whether particular treatments meet numerical discharge standards, it cannot measure treatment effect or change due to treatment (as there is no control). This may be a reasonable trade-off, with sufficient replication and extensive background information prior to shipboard testing (as outlined above). A key issue here is confirming that performance is indeed robust across a full range of challenge conditions, including especially those cases where large numbers of organisms are otherwise delivered in untreated tanks. As a minimum, it would be useful to collect initial and final data from treated tanks, which would serve to (a) characterize challenge conditions and (b) identify conditions where the treatment did not meet standards.

Another approach is to simply measure the successful delivery of the treatment agent, as a proxy for treatment effect. While this may initially seem an appropriate extension of scaling up, it also assumes perfect knowledge about the mode of action. Identifying dose-response and mode of action under various challenge conditions is the goal of experimental tests at small scales. However, it is important to realize the inherent limitations of these tests, as it is impossible to include all field conditions in laboratory, mesocosm, or test bed experiments, and the interaction of some conditions during shipboard applications (i.e., in the field) may produce unexpected results. Thus, use of such a treatment proxy as a sole indicator during an experimental testing phase for shipboard treatment systems is problematic without verification that the expected performance (effects on organism concentrations) is achieved.

We suggest that shipboard testing of ballast treatment systems occur in multiple phases, to balance the need for verification of performance against the effort required. A similar phased approach is outlined by U.S. Coast Guard in the documentation for the STEP, including two phases: Experimental and Monitoring. While we concur with the general strategy of phasing, we also suggest some changes in approach, as outlined below:

1. Experimental Phase. In the first (Experimental) phase, shipboard tests employ a controlled experimental approach, as outlined previously in the general guidelines. As a minimum, samples are collected (in replicate) before and after treatment for at least one control and one treated tank. This is the approach now outlined by US Coast Guard in the current STEP documentation, and serves to (a) establish that the treatment system meets the expected level of performance based on pervious developmental research and (b) provide an estimate of efficacy during ship board trials. .

The Experimental Phase would include at least one tank pair (treatment and control) that is challenged across different seasons, experiencing variation in environmental conditions, species composition, and organism concentrations. We recommend a minimum of 8 separate experimental tests, including 2 per season (fall, winter, spring, and summer). Ideally, this would occur in one year, but more time may be required in some circumstances. As discussed earlier, particular attention should be given to selecting the specific ship and route, which have important implications for the logistics of implementing this intensive experimental phase and also the applicability of the treatment system to a range of conditions.

2. Monitoring Phase.

In the second (Monitoring) phase, shipboard treatment includes all ballast tanks. During this phase, a suite of standardized measures is monitored through time. These are selected to measure the engineering performance as well as the residual biota in discharged ballast. Ideally, at least one "core" ballast tank would be selected for routine measures, and one or more additional tanks would be selected on each sampling event; these additional tanks would be selected by a prescribed schedule to include all tanks over time. The selection of a core tank would be based upon ballast operations and access aboard the ship. Measures would be made on the initial ballast water (environmental and biological) conditions and upon discharge/arrival to a subsequent port. These should be repeated at least quarterly, again to include the full range of seasonal variation in environmental and biological (and operational) conditions.

3. Compliance Phase.

After 5 years of operation, a third (Compliance) phase would begin. The Compliance Phase would rely primarily upon measures (preferably automated) of engineering performance of the treatment system, as developed under the Monitoring Phase. The data collected under the Monitoring Phase, combined with all previous work at smaller scales, establishes the effect of various treatment levels on biota. The Compliance Phase routinely tests whether this treatment level is achieved and sustained over time. This is done on all voyages.

In addition, some biological measures are collected on a specified timetable for the Compliance Phase, to determine the continued performance and lifetime of the treatment system. Since all treatment systems are new to ships, the expected longevity is not established for any system. Thus, we propose biological measures on at least two voyages per year, where biological data are collected as outlined in the Monitoring Phase. These voyages should be selected to meet the most challenging conditions, based upon prior data collected in the monitoring phase or knowledge of the source ports of ballast water. Also during the Compliance Phase, as elsewhere in the shipboard testing process, ships (and ballast systems) are subject to both regular and impromptu inspection.

This strategy differs from the current STEP guidelines in further defining and streamlining the measurement process. In particular, the number of sampling events is restricted to a finite number per year and the scope of the sampling effort is graded, declining over time to a low level that is maintained through the life of the treatment system. This approach is intended to minimize effort without compromising data needed for full evaluation.

In summary, allocation of effort in the evaluation of BWT systems is of paramount importance in achieving desired goals. There is a clear logic in implementing a standardized experimental approach that utilizes multiple scales in a step-wise progression and allows development of a coordinated international framework. Recent progress is evident in these areas, but additional aspects of testing protocols are still needed. We have outlined some of the current gaps in the biology and sampling of test organisms as well as the allocation of effort to shipboard evaluations. While multiple strategies are possible, there is now a premium on addressing these gaps quickly to advance an efficient, well-defined, and international approach to testing of BWT systems.