ARTICLE IN PRESS

Aquatic Mammals 2021, *47*(5), ###-###, DOI 10.1578/AM.47.5.2021.###

Marine Mammal Noise Exposure Criteria: Assessing the Severity of Marine Mammal Behavioral Responses to Human Noise

Brandon L. Southall,[1,2,3] Douglas P. Nowacek,[3,4] Ann E. Bowles,[5]
Valeria Senigaglia,[6] Lars Bejder,[6,7,8] and Peter L. Tyack[9]

[1]*Southall Environmental Associates, Inc., 9099 Soquel Drive #8, Aptos, CA 95003, USA*
E-mail: Brandon.Southall@sea-inc.net
[2]*Institute of Marine Sciences, Long Marine Laboratory,*
University of California Santa Cruz, Santa Cruz, CA 95060, USA
[3]*Nicholas School of the Environment, Duke University Marine Laboratory, Beaufort, NC 28516, USA*
[4]*Pratt School of Engineering, Duke University, Durham, NC 27708, USA*
[5]*Hubbs-SeaWorld Research Institute, 2595 Ingraham Street, San Diego, CA 92109, USA*
[6]*Murdoch University, 90 South Street, Murdoch, Western Australia 6150, Australia*
[7]*Hawaii Institute of Marine Biology, University of Hawaii, 46-007 Lilipuna Road, Kaneohe, HI 96744, USA*
[8]*Zoophysiology, Department of Biology, Aarhus University, Nordre Ringgade 1, 8000 Aarhus, Denmark*
[9]*Sea Mammal Research Unit, Scottish Oceans Institute, University of St Andrews, St Andrews, Fife KY16 8LB, Scotland*

Abstract

Major progress has been made since the publication of noise exposure criteria by Southall et al. (2007) in addressing the probability and severity of marine mammal behavioral responses to measured noise exposures. New methodological developments for studying behavioral responses have broadened the spatial, temporal, and population scales of potential disturbance studies and expanded scientific data on responses of marine mammals (or lack thereof) to various human noise exposure scenarios. Experimental and observational studies have substantially expanded the resolution, parameters, and contexts for understanding individual and group responses to discrete noise events. The combined data strongly suggest that efforts to derive simple all-or-nothing thresholds for single noise exposure parameters (e.g., received noise level) and behavioral responses across broad taxonomic and sound categories can lead to significant errors in predicting effects that are fundamentally inconsistent with the probabilistic nature of responses. Differences between species, among individuals, across situational contexts, and with the temporal and spatial scales over which exposures occur lead to variability in the probability and severity of behavioral responses. Studies that account for such factors and the variability they cause can provide far more accurate probability functions for predicting effects and can reduce variabilities related to exposure level and response. To that end, several new approaches are developed here for evaluating response severity in laboratory and field conditions in terms of effects on vital rates. These are applied to selected studies of marine mammal behavioral response to demonstrate their application in more consistently addressing acute exposure contexts for individuals or discrete groups. Needs for new approaches and transparent processes are identified for addressing sustained and/or repeated noise exposures on population scales.

Key Words: marine mammals, noise, behavior, response, severity, criteria, vital rates

Introduction

Southall et al. (2007) sought to establish broad exposure criteria to inform management and conservation decisions about the effects of noise on marine mammals. They categorized marine mammal taxa into five functional hearing groups and anthropogenic noise sources into two groups based on their acoustic characteristics (pulse and non-pulse), each with two modes of exposure (single and multiple). They addressed auditory effects and behavioral responses, deriving a novel response "severity scale" to assess behavioral impacts associated with noise exposure for captive and free-ranging marine mammals.

After a decade of ongoing research in each area, several other experts joined the original panel (please see "Acknowledgments" for a complete list of all panelists) to evaluate, update, and improve upon the recommendations of Southall et al. (2007). The first publication of this group updated evaluations of effects of noise on hearing in marine mammals (Southall et al., 2019a). Herein, we present updates for assessing behavioral response severity using novel, modified approaches for studies on different spatial and temporal scales. The primary innovations were to assess how noise can affect vital rates through aligning scores with survival, growth, and reproduction; different means of categorizing sound source types; and differences in how severity scoring may need to be considered to be applicable to studies that evaluate broader-scale, population-level impacts.

Questions regarding behavioral impacts on marine mammals have been addressed using observational and experimental methods for over a half century (e.g., Payne & Webb, 1971; National Research Council [NRC], 2000; Southall et al., 2007; Southall, 2017). Experimental approaches such as controlled exposure experiments (CEEs) can test causal relationships between designed and quantified exposure events and behavioral responses of individuals or identifiable groups of marine mammals (e.g., Nowacek et al., 2007; Southall et al., 2007, 2016). Deliberate escalation of exposure parameters (e.g., received level) within CEEs can identify the lowest exposure condition at which a response is elicited. Many studies for which structured severity assessment have proven the most effective thus far have involved CEE methods. In some well-documented situations (e.g., Moretti et al., 2014), observational studies can document whether a response occurs where exposures are measured and/or modeled. This may require the assumption that the response either occurs coincident with the lowest known exposure condition, if they are variable and uncontrolled, or at an identified point if exposure levels are known to be increasing (e.g., by approaching).

Ideally, severity scales for estimating noise impacts will consider the full range of possible outcomes associated with exposure. If one or more responses are estimated to occur, the basis for calculating any function predicting the associated impact(s) must be clearly specified. Too narrow a focus may lead to an overly broad application of any assessed "score." For example, when assessing workplace accidents, the Occupational Safety and Health Administration (OSHA) (1999) defined an accident severity rating as the number of workdays lost from an accident per thousand worker hours. This definition lumps many different medical or psychological outcomes into a single factor most impactful to the employer—lost workdays—but does not assess impacts of greater importance to the worker such as how disabling the injury was nor chronic effects that do not necessarily preclude work in the short term.

Another approach to evaluating risk involves classifying effect severity along a logical continuum scaled from low to high. Many semi-quantitative risk assessments, particularly for health indicators, assign scores (typically from 0 to some integer) for the severity of each parameter, summing scores for a composite total. For example, anesthesiologist Virginia Apgar developed a scoring system to assess the status of neonates in their first few minutes of life. Five different parameters are scored 0 to 2 and then summed for a maximum value of 10. The scores are self-referential, meaning the health value of one for a specified parameter is not necessarily of equal health consequence as a score of one for a different parameter. Nevertheless, the Apgar scores provide, in broad terms, a means for determining whether an infant requires rapid, critical intervention such as cardiopulmonary resuscitation (Schmidt et al., 1988).

Initial Marine Mammal Response Severity Scoring

The novel response severity scale developed by Southall et al. (2007) described marine mammal behavioral responses to noise in ascending order of presumed consequence. For example, responses such as a "brief orientation" to a noise source were deemed to be low severity (severity score 1) whereas more intense or sustained responses such as "prolonged changes in locomotion" (severity score 5) and "significant separation of females and dependent offspring" (severity score 8) were deemed moderate to high severity responses. Such ordinal scores would be assigned within the context of an experimental or observed noise exposure by informed observers based on the extent to which the observed behavior matched described responses in the severity scoring table ranging from 0 to 9 (p. 450). Southall et al. recommended coalescing severity scores in the 0 to 3, 4 to 6, and 7 to 9 categories into "low," "moderate," and "high" severity responses. Different regulations and/or regulators may apply different thresholds at which a response is regulated; in this case, regulators can use the relevant severity score and ignore lower levels.

The basic unit of analysis in these scales is the unit for which a response was observed—that is, an individual or a group with an observed number of individuals. For social group observations, Southall et al. (2007) proposed, as a precautionary approach, scoring the most severe response by any individual as the response score for the entire group. Where multiple discrete responses are observed from the same individual or group, Southall et al. assigned a discrete severity score for each response but applied a subject-weighting whereby a fractional value of one divided by the number of observations would be applied to each defined exposure for which a response severity was assessed such that the individual (or group) would have the same overall weighting as an individual with one response.

Southall et al. (2007) reviewed the existing literature for different sound types and marine mammal taxa and used their response severity scale to assess the results. Severity scores were judged independently by a subset of the authors and agreed upon (adjudicated) by scorers. Results were tabulated for each hearing group for pulsed and non-pulsed sounds (see Southall et al., 2007, Tables 6-23). Apparent patterns in response as a function of received noise level (sound pressure level) highlighted a number of potential errors in using all-or-nothing "thresholds" to predict whether animals will respond. Tyack & Thomas (2019) subsequently and substantially expanded upon these observations. The clearly evident variability in response is likely attributable to a host of contextual factors, which emphasizes the importance of estimating not only a dose-response function but also characterizing response variability at any dosage. The pooled data showed that some individuals had severe responses at low exposure levels while others had little or no response at high exposure levels. Several species-specific patterns also emerged for particularly sensitive species (e.g., harbor porpoises [*Phocoena phocoena*] and beaked whales), more tolerant species (e.g., humpback whales [*Megaptera novaeangliae*]), and species for which differences in response probability depended upon behavioral states (e.g., bowhead [*Balaena mysticetus*] and blue [*Balaenoptera musculus*] whales). Because of the high degree of variability within and between species and noise types, Southall et al. (2007) did not provide generalized risk probability functions for marine mammal hearing groups spanning all noise exposure types. However, they did provide a descriptive foundation for objective assessments of response severity that was and could be applied in structured assessments of existing literature, along with preliminary categorizations of species' variability and behavioral contexts that potentially affect severity of observed results.

Advances in Marine Mammal Response Severity Assessment

Several empirical studies of behavioral responses to noise exposure have employed the resulting severity scoring methods developed by Southall et al. (2007). Miller et al. (2012) applied the severity assessment to evaluate responses to sonar signals of killer (*Orcinus orca*), long-finned pilot (*Globicephala melas*), and sperm (*Physeter macrocephalus*) whales. Two groups of subject-matter experts independently evaluated individual time-series data from archival movement and acoustic tags and visual observations with known exposures and received noise conditions to identify specific times ("change points") and contexts associated with defined behavioral changes.

Based on their observations, Miller et al. (2012) proposed modifications to the Southall et al. (2007) scale, primarily more precise temporal descriptors of behavioral responses. Southall et al. (2019b) applied and expanded the Miller et al. (2012) response severity scale and scoring approach with independent groups of subject-matter experts to quantify effects of sonar on blue whales and compared resulting assessments with those from quantitative statistical methods using designated behavioral indices. Miller et al. (2014) conducted similar comparisons of scored severity and statistical change point detection methods with killer whales. Miller et al. (2012, 2014) and Southall et al. (2019b) thus yield individual instances within a time-series context in which discrete exposures of known received conditions and responses of variable severity were determined. Each method was effective and quite consistent in identifying changes in most subjects. However, some blue whales that were not feeding during the pre-exposure period actually began feeding during noise exposure. These whales were identified statistically as responding, given that those methods are designed to detect changes from the pre-exposure periods regardless of their direction, but not so by independent assessors as feeding onset was not a specified response in the behavioral severity scale. Additional statistical methods have been developed and applied to integrate the results of such responses (or lack thereof) in known exposure conditions to derive species-specific and multi-species exposure-response risk functions using model selection methods (Harris et al., 2016), Bayesian hierarchical models (e.g., Miller et al., 2014; U.S. Department of the Navy, 2017), and recurrent event survival analysis (Harris et al., 2015).

Integrative analyses that yield probabilistic response functions for responses of specified severity are being increasingly applied to assess impacts on individual vital rates and consequences for population-level impacts. Specifically, modeling efforts to quantify population consequences of disturbance from noise seek to build from short-term behavioral and physiological changes to longer-term population-level effects (e.g., Pirotta et al., 2018). Recent efforts (Pirotta et al., 2021) have parameterized these population-level modeled impacts with empirical data from behavioral responses measured in individuals and evaluated with response severity assessments using expert elicitation (Southall et al., 2019b). These kinds of integrative assessments coupling short- and longer-term individual and population-level responses require information on the type, probability, and severity of responses. Further, they require information about how responses affect activities such as foraging or mating, or how responses to threats may influence vital rates of survival and reproduction. Finally, they are strengthened by empirical measures of response at both the individual and group/population level. Few experimental or opportunistic studies have analyzed results from short-term behavior through to life history impacts, but obtaining and integrating results at each level is increasingly relevant and required for efforts to evaluate population-level impacts from discrete and aggregate stressors (National Academies of Sciences, Engineering, and Medicine [NAS], 2017; Pirotta et al., 2019).

In light of research progress, evolution of risk assessment methods, and evolving policy drivers since the publication of the Southall et al. (2007) noise exposure criteria and severity scale, herein we provide several new assessment methods to systematically characterize marine mammal responses. Due to the wide range of spatial and temporal scales of exposure and potential disturbance over which researchers are measuring, we distinguish between the kinds of studies and assessment methods needed for acute noise exposure events with identifiable responses from known individuals or groups vs those involving sustained or repeated (chronic) exposure scenarios and potential responses at population levels.

Most of the early consideration of these issues, including the Southall et al. (2007) exposure criteria and subsequent analytical methods (e.g., U.S. Department of the Navy, 2017) has focused on discrete responses of individuals to known exposures, often tagged individuals in CEEs (e.g., Southall et al., 2016). These kinds of assessments are particularly needed in regulatory scenarios for which responses are considered at the individual level such as discrete behavioral responses deemed to represent a specific level of impact such as specified "takes" of individuals as evaluated under the U.S. Marine Mammal Protection Act of 1972. Herein, we adapt earlier approaches and develop new analytical methods aimed to improve assessments of acute exposures and potential behavioral responses with a more explicit distinction of field and captive scenarios. Further, we abandon the hearing-based distinction of impulsive and non-impulsive noise categories because some source types, such as airguns, may produce impulsive sounds near the source and non-impulsive sounds at greater ranges (for further discussion, see Southall, 2021). Instead, we distinguish more practical operational categories of sources, applying the revised severity assessments to selected studies in each category.

Applying the revised acute exposure severity scale (described in greater detail below) to longer-term (e.g., seasonal, annual) studies with fewer discrete exposures to specific individuals reveals limitations in the applicability of this approach for these scenarios. Some long-term studies have evaluated the impacts on presence and distribution (i.e., at population levels) of aggregate or cumulative exposure to stressors, acoustic and non-acoustic, from human activities such as coastal development, whale watching, and long-term seismic airgun surveys. We highlight the need for additional analytical approaches (beyond the acute severity scale) for broader spatial scale population studies, particularly for regulatory assessments of cumulative impacts and in regulatory paradigms where the focus is at the population level (e.g., European Union Marine Strategy Framework Directive). We explore several new means of evaluating and addressing these broader scales, including new applications of assessment methods for evaluating long-term impacts of whale watching and novel methods to quantify population consequences of disturbance(s).

Studies of both discrete exposure-response and medium- to long-term population-level responses (e.g., distribution, broad-scale acoustic behavior changes) are needed and can be insightful if properly designed. They simply require different analytical methods for results to be integrated more broadly in efforts to predict response type, occurrence, severity, and consequence for application in real-world decision making. Both types of assessment, however, require more comprehensive, detailed, and consistently reported information on exposure contexts, received noise metrics, ecological conditions, and detailed descriptors of individual and/or population response metrics.

Methods

Needed Exposure and Contextual Metrics

There has been increasing focus on the range of exposure and response variables that may be relevant in understanding and describing marine mammal behavioral responses. This includes recognition that different metrics of acoustic exposure may be relevant in different settings (e.g., Madsen, 2005; Southall et al., 2007, 2019a) and the importance of contextual factors (e.g., spatial proximity, behavioral and reproductive state, natural history, ecological parameters) in response probability and magnitude (e.g., Ellison et al., 2012; Southall et al., 2019b).

Consequently, a broader suite of noise exposure conditions other than a single received level (RL) metric should be analyzed and reported in response studies. The relevance of certain exposure and contextual metrics in captive and field exposure scenarios will differ based on species, noise source, context, and temporal scales of analyses, to name a few. Variables related to aspects of sound propagation and spectral/temporal/spatial aspects of sound exposure may also be significant mediating contextual factors (Ellison et al., 2012). Herein, we provide a comprehensive set of recommended metrics that we advocate should be consistently reported, to the extent possible and applicable, to support integrated analyses of exposure-response relationships (Table 1). These are segregated into subject-specific (A), exposure context (B), and noise exposure (C) metrics. Additional details and discussion for each of the elements identified in Table 1 are provided in Appendix 1.

Some of these metrics are more applicable and relevant for acute exposure studies than long-term exposure scenarios where it may be more challenging to characterize individual exposures, but many are relevant regardless of temporal context or may be summarized with ranges or average values for longer-term studies. Many are self-evident and would typically be included in studies, but we advocate for a consistent, systematic reporting of as many of these synoptic data and covariates as possible. Where space is limited due to the constraints of journal length, metrics less essential to the study design can be included in supplementary materials. Not only are the listed metrics important to understanding the outcome of particular studies, but their absence has likely contributed to outcomes of meta-analyses that are difficult to interpret (Southall et al., 2007) or negative (e.g., Gomez et al., 2016).

Subject-specific variables (Table 1A) are intended to identify individual or group features that may be relevant in coding metadata (e.g., species, functional hearing group) and determining appropriate quantitative weighting for individual or multiple exposures (e.g., number of individuals, whether subjects were exposed on multiple occasions). Identifying whether exposure events were "censored" denotes whether responses were observed at the lowest exposure condition (e.g., RL) tested (left-censored) or if no response was measured in any exposure conditions (right-censored). We also call for reporting of social factors (e.g., group size, composition) and behavioral state, which may be important contextual covariates to account for in interpreting responses and/or in pooling results (e.g., Southall et al., 2016, 2019b).

Other variables related to the exposure context are called for as well, including the categorical type of exposure (meaning the four broad categories specified herein along with more specific descriptors) and spatial and relative frequency of similar exposures in the study area (Table 1B). These contextual aspects of exposures are called for because relative proximity, similarity with predator signals, familiarity with exposures, and the presence of other disturbances are important factors affecting the type and probability of response in earlier marine mammal studies (Southall et al., 2007; Tyack et al., 2011; Ellison et al., 2012; Miller et al., 2012; Wensveen et al., 2019). Note that in some cases (e.g., similarity with predator signals), research may be needed to define how such contextual variables should be quantified.

Calls for comprehensive and consistent standardized reporting of the kinds of acoustic exposure metrics called for in Table 1C have strengthened in the last decade (e.g., Ellison et al., 2012; McKenna et al., 2016). Given the diversity of sound types, including impulsive and non-impulsive signals (and those which may transition from the former to the latter), and because subsequent criteria may seek to differentially apply different metrics, we advocate for a complete accounting of multiple sound pressure level (SPL) and sound exposure level (SEL) metrics. As we discovered, these metrics may become important in meta-analyses even though the original authors did not consider them to be the preferred independent variables for study purposes. Accounting for exposure conditions in the context of concurrent noise (e.g., signal-to-noise ratio [SNR]) and subject-specific hearing sensitivity (e.g., sensation level [SnL]) in specified bands has also been called for (e.g., Ellison et al., 2012). In the long run, there will also be value in developing better measures of audibility that can be applied to real-world sounds, although needed psychophysical data are still lacking for animals (e.g., Bee & Micheyl, 2008).

Evaluating Behavioral Response Severity – Discrete Exposures

The Southall et al. (2007) response severity scale has been applied and revised in the course of practical applications using expert elicitation methods (e.g., Miller et al., 2012; Southall et al., 2019b). We propose herein a number of substantial additional modifications for discrete exposure events. Notably, we derive discrete severity scales for captive and field exposure contexts. Further, we segregate responses along different categorical "tracks" of increasing severity. For captive marine mammal contexts, this includes discrete consideration of responses related to untrained and trained behaviors. For field contexts with free-ranging marine mammals, we segregate responses into categories related to foraging, survival, and reproduction, which may differentially affect vital rates.

The severity scale derived by Southall et al. (2007) included discrete tracks for responses observed for free-ranging and captive marine mammals in known exposure conditions. This

Table 1. Subject-specific (A), contextual (B), and noise exposure (C) metrics to be reported in captive and field studies of marine mammal behavioral responses to noise

A. Subject-specific (individual or group) variables

Species	Functional hearing group	Subject individual identifier (where applicable)	Subject weighting (subject A for stimulus B for *N* times divide run by *N*)	Censored data? (No or L/R if yes)	Age class (if known)	Sex (if known)	Calf present? (if female)	Group size (single or best estimate of social group size)	Group composition (general sex/age structure)	Behavioral state (e.g., deep/shallow feeding, slow/fast travel, social interaction, calling)

B. Exposure context variables

Exposure type (start of exposure)	Source–animal range (start of exposure)	Source depth (m)	Animal depth (m)	General source movement (relative to subject)	Navigational constraints (is subject confined in any way?)	Exposure novelty (is source type common/rare for area?)	Exposure similar to predator sounds?	Other species present in the area?	Predator species present in the area?	Other anthropogenic presence/noise in area? (type and proximity)

C. Noise exposure metrics

Continuous or intermittent exposure	Interval between exposures (s)	Individual duration (s)	Individual rise time (s)	Total exposure duration (s)	Order if multiple exposures (identify sequence/order)	Harmonics present? (none, few, many)	RMS SPL @ change point or max. if no change (broadband and max. 1/3-octave)	Peak-to-peak RL @ change or max. if no change (broadband)	SEL @ change point or max. if no change (broadband and max. 1/3-octave)	SELcum @ change point or max. if no change (broadband and max. 1/3-octave)	Signal-to-noise ratio (SNR) @ change point or max. if no change (max. 1/3-octave)	Sensation level (SnL) @ change point or max. if no change (max. 1/3-octave)

resulted in both empty cells and unintended presumed equivalences in severity between very different contexts. Herein, we decouple the different contexts of behavioral responses in captive and field conditions. Both the captive and field severity scales describe discrete responses of presumed increasing severity, but they differ in the resolution of ordinal severity sequence, include different overall categories of response, and have specific and dissimilar descriptors of behavioral responses.

Response Severity Assessment Methods (Captive Studies)—The severity assessment proposed herein for captive marine mammals is more narratively descriptive than the field scale and is somewhat more similar to the captive portion of Southall et al. (2007). Inherent in the separation of severity scales for captive and field settings is the observation that captive studies of marine mammal behavioral responses may be limited in their application to free-ranging scenarios given the very different environmental and motivational contexts. In particular, a training paradigm makes it possible to measure the aversiveness of exposures to noise very precisely, and negative responses of captive animals can be observed in greater detail than is possible with individuals observed for short periods with remote monitoring equipment. However, probabilistic response relationships as a function of noise exposure for captive animals that have been reinforced in training to respond to particular signals may be usefully compared to free-ranging animals that experience positive and/or negative reinforcement for responses to ambient signals in the wild (e.g., reactions of resident animals to net alarms where food is present). It may also be useful to compare responses of captive animals to sounds that have not been associated with reinforcement with responses of free-ranging animals to novel sounds or those that lack clear consequences. Habituation may be difficult to measure in free-ranging animals where subjects in CEEs may or may not be familiar with an exposure stimulus or experimental situation, but under captive conditions, repeated exposures over periods of days or weeks can be used as a way of differentiating stimuli that are intrinsically aversive from those that are avoided if novelty arouses defensive behaviors (e.g., Bowles & Anderson, 2012). Such comparisons may help clarify whether animals are responding to acoustic properties of a signal or showing responses that are habitual or conditioned by reinforcement.

Within captive responses, we distinguish between trained and untrained behaviors (Table 2). Both tracks represent increasing severity, but the increase is not necessarily proportional between the two tracks. If a subject exhibited any one of the responses within a severity category, it received that score. Narrative heuristic descriptions are provided for untrained responses to provide some insight into general overall comparisons with field observations of behavioral response. The objective was to develop a practically useful scale (Table 2) for considering studies with heterogeneous methods for comparing different types of responses within untrained and trained behavioral contexts. The scales for captive animals do not incorporate fitness and survivorship-related tracks (e.g., effects on foraging) because they are to be applied in situations where population-level effects are less relevant and certainly less measurable. However, they include details of behavior that might not be measurable in free-ranging animals and include accounting for habituation, which will be unknown or unmeasurable for most free-ranging animals. A direct way of determining whether a behavior is the result of defensive neophobia as opposed to aversion per se is the response over the course of habituation, particularly in a context where subject marine mammals are exposed with a competing positive stimulus (e.g., Götz & Janik, 2011). The captive marine mammal scale differentiates responses that habituate rapidly and completely (severity score 2), responses that habituate slowly and incompletely (severity score 3), and responses that decline little or even amplify over time (severity score 4 – sensitization).

Response Severity Assessment Methods (Field Studies)—We propose a fundamentally different approach from Southall et al. (2007) for evaluating the relative response severity for free-ranging marine mammals. The overall approach retains discrete behavioral categories identified along an ordinal scale of increasing severity with descriptors of response type, magnitude, and duration. The objective is to expand on the original response descriptions and to identify those of increasing severity in more ethological terms along three parallel severity tracks. This framework is thus explicitly relevant to vital rates, defining behaviors that may affect individual fitness and, consequently, population parameters. The three tracks evaluate behavioral responses related to the following:

1. Survival (including effects on defense, resting, social interactions, and navigation)

2. Reproduction (including mating and parenting behaviors)

3. Foraging (encompassing search, pursuit, capture, and consumption)

The presumption is that responses increase in severity along each track, but identical scores across tracks do not imply equivalent severity. There is no expectation nor requirement that subjects would exhibit all the responses within one severity category; a score is assigned for a severity category if a subject displays any response within that category. If it makes several responses from different categories to the same exposure level at the same time, the highest severity score is reflected for that exposure. While there is some small degree of redundancy across these descriptors (e.g., behaviors that arguably relate both to foraging and to survival), the intent is to provide a means of evaluating behavioral responses in a way that facilitates interpreting consequences in terms of vital rates. Herein, we seek to improve the biological and ecological basis for evaluating the severity of responses to discrete exposure events by placing them in the context of individual vital rates (Table 3).

Table 2. Behavioral response severity scale for discrete exposures of captive marine mammals

Response severity score	Untrained behavioral responses	Trained behavioral responses
0	*Heuristic:* No response* No detectable response	No detectable response No evidence of change in coping behavior**
1	*Heuristic:* Just detectable responses—surprise, brief neophobia, and investigation • Momentary detectable behavioral change (e.g., brief orientation, change in swimming, defensive silence) • Short-term orientation and/or investigation of sound source • Short-term and short-distance avoidance during initial exposure, especially if stimulus onset is rapid	• Initial short-term or short-distance movement of head or body while at station or in transit during response (e.g., looking, brief stop, abrupt movement) • No detectable change in performance of trained behaviors • Easily habituates to stimulus (ceases movements with repeated exposure) No or minor changes in coping behavior
2	*Heuristic:* First evidence of aversion and defensive behavior; more than surprise; response habituates • First evidence that a response is negative rather than neophobic or investigative • Brief/short-range persistent avoidance • Isolated or transient defensive behaviors (e.g., bubble clouds, rapid approaches or swimming passes, jaw claps, brief surface activity) • Conservative defensive behaviors—for example, congregation of mothers and calves • Naïve subjects avoid the source, but at close range and < exposure duration	• First evidence that a response is negative rather than neophobic or investigative • Short-term or short-distance avoidance after repeated or habituating exposures • Change in performance begins, especially if the task is difficult but effect small • Experienced subjects begin to show sensitization, but the behavior habituates over time • First evidence of coping behaviors
3	*Heuristic:* Aversion and defensive behavior; exposed animal does not fully habituate • Frequent defensive behaviors (e.g., bubble clouds, rapid approaches, jaw claps) • Moderate avoidance of sound source (> criterion distance; ~duration of exposure) • Decline in response with repeated exposure • Defensive congregation (all age/sex classes) • First evidence of aggression or exclusion directed at other individuals	• Detectable decline in performance • Changes in response to trained behaviors interfere with task (e.g., leaves station frequently, reluctance to return to station, long inter-trial intervals) • Food reward still a sufficient inducement to work in some trials, but experienced subjects begin to avoid the source consistently if there is no food reward • Experienced subjects with a food reward attempt to "game" the reward without getting exposed • Startle is initially negatively reinforcing but will not consistently deter subjects if food reinforcement is available

4	*Heuristic:* Sufficiently aversive that animal makes significant effort to avoid; aversion habituates poorly; may not tolerate exposure even if food reward is available; persistent agonistic behavior • Subjects avoid experimental situation or retreat to refuge area ≥ duration of exposure • Pinnipeds may jump out of the water • Persistent threats, charging, or attacks directed to sound source or displacement objects • Repeated exclusion or displacement of subordinate animals • Negative anticipatory behavior, especially when familiar with the stimulus (e.g., enters refuge as soon as trial begins) • Logging at the surface or bottom of the test pool, especially if the subject does not have the opportunity to leave area	• Breaks in stationing prolonged or station avoided‡ • Stops ongoing activity to charge or attack sound source or displacement objects • Refusal to perform conditioned tasks over time even in the face of food reinforcement • Overt and repeated aggression threatened or directed at trainers, other subjects, or objects • Logging or retreating to the bottom of the test pool even when encouraged to return • If a refuge space is available, retreats into it

*Note whether the criterion is based on observer judgments or statistical comparisons.

**Animals sometimes develop habitual behaviors when a training task becomes difficult. This criterion addresses changes in these behaviors—for example, if they become more exaggerated or if the animal begins to engage in stereotypical movements when it did not before exposure.

‡Strong responses are defined based on the upper limit of response in their experimental situation. The strongest response varied by species, stimulus, and degree of habituation.

Evaluating Behavioral Response Severity – Sustained (Chronic) Exposures

As described (and demonstrated below), the kinds of acute severity assessment methods described above are not necessarily well-suited to evaluate studies on the long-term consequences of non-lethal effects of sounds and concurrent multiple threats (i.e., climate change, other anthropogenic disturbance, etc.). For instance, where the unit of analysis in a longer-term study is a local population and, thus, would be reduced to a single observation, there may be inappropriate or unfair comparisons with severity assessments of acute exposure studies where n known exposed individuals are weighted as N observations. While controlled experimental and observational studies reporting individual-level exposure, context, and response remain critically important, other approaches are needed to consider mechanistic linkages between short-term acute exposures and population effects (e.g., Lusseau, 2014), and to evaluate the conservation impact of acoustic disturbance and chronic stress (e.g., Blickley et al., 2012; Simmonds et al., 2014).

Assessing the consequences of non-lethal disturbance is challenging and requires substantial data before and after disturbance, which are rarely available for long-lived marine mammal species. In recent years, ecological modeling and simulation studies have been adopted to overcome the logistical and practical challenges of collecting relevant life-history parameters. For instance, age-structure population assessment allows us to consider different impacts across age class and can be coupled with new technology such as unmanned aerial vehicles (UAVs) for rapid assessment of individual and population health conditions (Booth et al., 2020). Energetic modeling methods have also been developed to link short-term behavioral responses to disturbance to reproductive outcomes (e.g., New et al., 2014). Additionally, Population Consequences of Multiple Stressors (PCoMS) allows the modeling, as the name suggests, of population consequences of multiple and concurrent anthropogenic and environmental disturbances (NAS, 2017).

These new methodologies are still at an early stage of development, however, and are rarely evaluated against empirical data. The assessment approach we use herein for evaluating and systematically assessing population-level studies for chronic and aggregate disturbances is based on concepts first identified by Bejder & Samuels (2003). As a starting point for systematically reporting and assessing longer-term, population-level disturbance studies, we expanded the original approach to specifically address selected studies primarily associated with whale watching. Such studies comprise some of the longest-term marine mammal studies conducted and, thus, the

Table 3. Behavioral response severity scale for discrete exposures of free-ranging marine mammals

Response score	Behavioral changes affecting survival	Behavioral changes affecting feeding	Behavioral changes affecting reproduction
0	No response detected with methods sufficient to identify responses relevant to survival	No response detected with methods sufficient to identify responses relevant to feeding	No response detected with methods sufficient to identify responses relevant to reproduction
1	Identifiable change in behavior indicating vigilance response: • Orientation • Interruption of resting behavior • Listening: Delay in vocal behavior/locomotion/breathing • Detectable change in diving behavior • Minor deviation from typical migratory pathway	Detectable interruption of foraging behavior	Detectable interruption of advertisement and courtship behavior
2	Sustained or multiple vigilance responses		
3	• Individual investigation of potential threat • Recruitment of orienting behavior • Increase in contact or alarm calls to initiate social cohesion • Individual startle response	Behavioral state changes from foraging to other behavior	Behavioral state changes from advertisement and courtship to other behavior
4	• Prolonged silencing or other cryptic behavior to avoid detection • Defensive bradycardia or stillness • Increased interval between surfacing bouts • Reduction in variance of heading • Change in group cohesion • Brief/minor changes in vocal rates or signal characteristics—potentially related to higher auditory masking potential	• Non-foraging state longer than typical • Detectable elevation in energy expenditure (e.g., increase in dynamic acceleration, respiration rate, locomotion, speed) • Brief/minor changes in vocal rates or signal characteristics—potentially related to higher auditory masking potential	• Non-reproductive (advertisement and courtship) state longer than typical • Brief/minor changes in vocal rates or signal characteristics—potentially related to higher auditory masking potential
5	• Onset of avoidance behavior (e.g., heading away and/or increasing range from source) • Recruitment of defensive social behaviors (e.g., rafting, marguerite, vocal threats) • Increase in mother–offspring cohesion (including acoustic signaling and/or mother herding offspring)	• Reduction of foraging success less than typical daily intake requirement (during exposure period) • Detectable change in nursing behavior	
6	• Repeated startle response; abrupt agonistic behaviors (e.g., head thrusting, mouth gaping) • Individual aggressive behavior (e.g., jaw clapping, gnashing teeth, abrupt directed [rush/ramming] movement potentially directed at conspecifics) • Sustained avoidance behavior (e.g., heading away and/or increasing range from source) • Separation of females; dependent offspring exceeding baseline • Group aggressive behavior (e.g., mobbing) • Sustained changes in vocal rates or signal characteristics—potentially related to higher auditory masking potential	• Reduction of foraging success exceeding typical daily intake requirement (potentially extending beyond exposure period) • Energy expenditure exceeds nominal daily baseline • Sustained disruption of nursing behavior • Sustained changes in vocal rates or signal characteristics—potentially related to higher auditory masking potential	• Reduction of advertisement and courtship behaviors potentially sufficient to reduce reproductive success • Disruption of parental attendance behavior • Sustained changes in vocal rates or signal characteristics—potentially related to higher auditory masking potential

7	• Separation of females and dependent offspring sustained for long enough to compromise reunion • Clear anti-predator response (e.g., severe and/or sustained avoidance or aggressive behavior) • Displacement to area of increased predation risk • Failure of vocal mechanisms to compensate for noise (e.g., silencing affects group cohesion/defense)	• Reduction of foraging success sufficient to compromise health and/or reproduction • Failure of vocal mechanisms to compensate for noise (e.g., cessation of acoustically mediated foraging)	• Interruption of breeding behavior • Failure of vocal mechanisms to compensate for noise (e.g., cessation of acoustic advertisement displays)
8	• Disruption of group social structure (e.g., breaking pair bonds/alliances, altering dominance structure) • Prolonged/significant separation of females and dependent offspring with disruption of acoustic reunion mechanisms	• Prolonged displacement to suboptimal foraging habitat • Disruption of group social structure (cooperative feeding groups with specialized knowledge or division of labor)	• Disruption of breeding behavior sufficient to compromise reproductive success (e.g., repeated interruption of mating, disrupting male–female association) • Disruption of group social structure (e.g., breaking pair bonds/alliances, altering dominance structure)
9	Risk that behavioral response leads to serious injury or mortality (predation, outright panic, flight, stampede, stranding, mother–offspring separation)	Disruption of energetic balance sufficient to result in morbidity or mortality	Failure to successfully reproduce during breeding season

most suitable to illustrate direct linkages between behavioural responses to human disturbance (including noise disturbance) and population consequences. Factors and corresponding codes used to identify study elements and distinguish among different exposure and response characteristics based on the approach developed by Bejder & Samuels are specified in Table 4.

We constrained our review to peer-reviewed literature, in preference to "grey" literature, using key words, including whale watching, dolphin watching, long-term effects, impact, ecotourism, anthropogenic disturbance, and other similar terms. All references identified in this search were then evaluated for relevance and, if appropriate, reviewed in full. Of the 16 references included in the review, 11 were directly related to whale watching and related touristic activities. In addition, five papers examined long-term consequences of other anthropogenic disturbances. In some cases, more than one paper was included per body of work or project to provide synergetic results while avoiding overlap between the sources. This is by no means an exhaustive review of these issues but an illustration of an adaptive structured review and assessment of studies in four main topical and temporal categories:

1. Studies evaluating effects of whale watching, encompassing a dataset of multiple years

2. Studies evaluating long-term effects of whale-watching industry, encompassing a dataset of > 10 y and providing long-term measures of the effects of whale-watching activities on population dynamics

3. Studies evaluating long-term effects of anthropogenic activities (other than whale watching), encompassing a dataset of > 10 y or making use of a modeling approach to forecast long-term measures of the effects of anthropogenic activities

4. Studies evaluating long-term effects of multiple concurrent threats, encompassing a dataset of > 10 y or making use of a modeling approach to forecast long-term measures of the effects of anthropogenic activities

We reviewed each study, highlighting study design and analytical approach, data collection platform, and whether the unit of analysis was individuals or groups. Different research methods carry specific strengths and weaknesses. Further discussion on the most appropriate methodology for population-level studies is beyond the scope of this review, but various limitations have been widely considered (e.g., Bejder & Samuels, 2003; Senigaglia et al., 2016; Booth et al., 2020). We specifically considered studies that included both

Table 4. Codes and definitions of factors (derived from Bejder & Samuels, 2003) used to evaluate studies associated with whale watching and other forms of human disturbance in population-level studies evaluated in Tables 9, 10, and 11

Factor	Code	Definition
Short-term measures		
	BB	Changes in behavioral states/activity budgets
	A	Loss of acoustic space/masking (of communication or prey acquisition)
	H	Changes in swim speed, course, and orientation (horizontal avoidance)
	V	Changes in surface, ventilation, and dive patterns (vertical avoidance)
	E	Energetic imbalance due to increased demand (avoidance or increased swimming speed) or decreased intake (less foraging opportunity)
	G	Changes in group size/dispersion/cohesion
	M	Changes in animal motivational state
	BC	Body condition (body mass; blubber thickness)
Long-term measures		
Reproductive success and survival	H	Health status
	CR	Calving rate
	CS	Calves survival
	AS	Adult survival
	PD	Population decline
Ranging patterns and habitat utilization	D	Displacement
	SR	Sighting rate
Study characteristics		
Type	E	Empirical data collected
	M	Modeling approach
	C	Combination of empirical data and modeling forecast approach
Design	CE	Controlled experiments
	O	Opportunistic observations
	H	Historical data
Platform	B	Data are collected from a boat-based platform
	L	Data collected from land platform
Analytics	WE	Within effect comparison
	CI	Control vs impact comparison
	BDA	Before/during/after comparison
Subject	G	Focus on the group
	I	Focus on the individual
Whale-watching industry characteristics		
Duration	S	Short established industry: < 5 y operation
	L	Long established industry: > 5 y operation
Regulatory framework	P	Permits/license legislation
	R	General regulation for cetacean protection
	G	Guidelines (voluntary or official)
Intensity	LI	Low impact: < 3 boats present at one time
	HI	High impact: > 3 boats present at one time
	WW	Whale-watch/dolphin-watch tours
	SW	Swim-with in water encounters with humans
Source of anthropogenic disturbance		
Whale watching	WW	
Pile driving	P	
Seismic survey	S	
Non-targeting boat traffic	BT	
Military exercise (e.g., sonar)	M	
Commercial fishing	F	
Pollution	PC	Chemical pollution
	PA	Acoustic pollution
Climate change	C	
Other	O	(Specific information given on a case-by-case basis)

empirical measurements and modeling approaches to forecast long-term consequences. For whale-watching studies, we also included a description of the history of development of the industry in each location. These activities are frequently implemented with highly variable levels of regulatory oversight and vary greatly in capability, in terms of fleet size and number of vessels allowed per encounter and/or animal, and by type of tour offered, with or without an in-water component. We classified the type of whale-watching regulatory framework in place based on categories provided in Tyne et al. (2014), using methods described in Table 4, which are described and specified in the corresponding text and tables in the "Results" section.

In cases of resident populations or animals showing high fidelity to an area, we assumed that the duration since implementation of the industry was a proxy for the potential magnitude of an effect. We thus highlight the number of years from when whale watching was implemented. Mysticetes and odontocetes differ in their life history characteristics. Mysticetes are capital breeders, accumulating energy on feeding grounds and transferring energy to calves in breeding grounds; whereas odontocetes are generally income breeders with less discrete feeding and breeding periods occurring throughout the year (e.g., McHuron et al., 2017). Anthropogenic disturbance studies generally focus on specific habitats within an animal's home range (namely, feeding and breeding grounds, migratory corridors, and areas where populations reside year-round) as this may affect their ability to compensate for disturbances. We consequently separated studies on mysticetes and odontocetes and consider compensatory opportunities on a study-by-study basis. For each temporal category, we provided in-depth analyses of a single selected study per taxonomic group (mysticetes vs odontocetes), when available, representing examples of best practice and/or most significant results. One example for each is given to highlight potential differences (e.g., Wade et al., 2012) or similarities (e.g., Senigaglia et al., 2016) between species in each taxa with different life-history strategies. Results from the remaining studies reviewed are included within the corresponding "Results" tables for the chronic exposure severity assessments, with additional corresponding text included in Appendix 2.

Results

Adjudicated Response Severity Results from Selected Acute Exposure Studies

A full application of the revised behavioral response severity scales for acute (captive and field) exposure scenarios (provided in Tables 2 and 3) to the entirety of marine mammal literature on behavioral responses to all anthropogenic noise sources is well beyond the scope of this article. In an effort to evaluate and illustrate how the modified severity assessments function, multiple assessors independently evaluated a subset of the published literature. We used a structured process both to categorize studies and to select a manageable number (*n* = 20) from over 400 studies published prior to 2018 that were initially identified and considered.

First, we focused primarily on studies of free-ranging marine mammals given the prevalence of such studies in the published literature. Second, we pooled studies by operational source types (specifically, active sonar sources, industrial [continuous] sources, pile driving, and seismic airgun surveys) for all species rather than impulsive/non-impulsive and animal hearing group distinctions (Southall et al., 2007, 2019a). Finally, we assessed all 400+ studies and categorized them (within the four operational source types specified) as having a high, moderate, or low priority for scoring accordingly. *High priority studies* were identified as those having many (> 9) of the exposure, contextual, and response metrics specified (Table 1); papers with less than nine metrics reported were then subsampled for review and evaluation (where > 5 studies existed within each source type) based on the distribution of reporting metrics across categories of animal subject variables, exposure context variables, and exposure metrics. *Moderate priority studies* were identified as those having three to nine of the exposure, contextual, and response metrics given in Table 1. *Low priority studies* were identified as those having fewer than three of the exposure, contextual, and response metrics specified. Within each source type, where more than five studies were identified in the high priority category, a randomization process was used to identify those that would be selected first from that category. If more than five high priority studies were identified within a source type, they were randomly selected, and the selected studies were scored by all three assessors. Where fewer than five high priority studies were selected within a source type, moderate priority studies were selected from the randomized pool until a total of five studies per source type were identified. Low priority studies were not selected or scored.

Following this study selection process, 20 studies (five from each of the four source types) were evaluated relative to the acute (field) severity scale (Table 3) independently by three assessors (authors P. Tyack, D. Nowacek, and B. Southall) with extensive field and analytical experience with marine mammal behavioral response studies. It should be noted that each of these assessors were also lead or co-authors of a number of the studies evaluated. There is extensive benefit in having experts in the field familiar with methods, species, and exposure types involved in studies being assessed within a structured expert assessment, as well as a higher likelihood that the more experienced an assessor, the more likely they may have been involved in some of the most relevant studies. However, a fair criticism could certainly be made that the inclusion of authors of studies being evaluated in this process is not impartial. Subsequent assessments using these methods (e.g., full reviews of all literature for a specified noise type) should consider the relative balances between having very experienced reviewers vs assessors with fewer potential biases within what remains a relatively small field.

Following their independent assessments, an adjudication process (following Miller et al., 2012; Southall et al., 2019b) was conducted in which consensus results summarizing exposure and response occurrence, type, and severity were identified. Syntheses of the adjudicated results, including the level of agreement among the independent assessors, are provided below for each source type, including active sonar sources (Table 5), industrial (continuous) sources (Table 6), pile-driving sources (Table 7), and seismic airgun survey sources (Table 8).

Adjudicated Response Severity Assessment

Active Sonar Sources—Research on the behavioral effects of active sonar on marine mammals has been a very active field, with both observational studies of actual sonar exercises and CEEs to measure individual responses to known exposures (see Southall et al., 2016). Consequently, numerous high priority studies were identified with results amenable to the severity assessment developed here. From these, five were selected: Tyack et al. (2011), DeRuiter et al. (2013), Hastie et al. (2014), Miller et al. (2014), and Isojunno et al. (2016) (see Table 5).

Tyack et al. (2011) combined results from experimental and incidental exposures and responses of Blainville's beaked whales (*Mesoplodon densirostris*) to naval active sonar signals played back through a stationary playback system and to actual naval sonar exercises. Experimental exposures with fine-scale tag sensors are detailed in terms of exposure and response. Longer-term satellite-transmitted tag data provide a longer perspective on avoidance during and following sonar exercises but require assumptions about exposure given the lack of acoustic data and limited information on source proximity (lowest possible RLs were reported). There was extremely consistent scoring of exposure, context, and response severity by all assessors given the details presented. Responses were documented in all individuals, including cessation of foraging and (most commonly) sustained avoidance.

DeRuiter et al.'s (2013) experimental study measured behavioral responses of Cuvier's beaked whales (*Ziphius cavirostris*) to naval active sonar signals and killer whale sounds. Individual exposures and responses were measured continuously using acoustic tags and are reported in detail with some known contextual and exposure metrics specified. Each assessor provided nearly identical assessments of exposure and responses of variable severity, which included cessation of fluking, cessation of echolocation, extended dive duration, shallow ascent, and sustained avoidances. Further, assessors noted the apparent indication of range-dependent response differences and potential RL-dependent effects when considered within exposures.

Hastie et al. (2014) conducted CEEs in a captive setting with gray seals (*Halichoerus grypus*) exposed to two sonar signals; this single captive study was assessed with the field severity scale. Exposure conditions and responses were specified in detail, and assessors reported nearly identical metrics of exposure and response type and severity. Two individuals were each exposed to two different signal types, and RLs were reported within the most sensitive band of hearing. Assessors identified avoidance of the sound source in both individuals to both signals, with a stronger haul-out response identified (and assumed to represent sustained avoidance) for one signal type.

Miller et al. (2014) studied behavioral responses of killer whales (*Orcinus orca*) to several kinds of naval sonar signals. Assessors largely concurred on exposure and responses which were sampled continuously with tags and described in detail, with some brief adjudication required for concurrence on weighting responses by individual. Assessors identified clear responses in tagged individuals in some but not all contexts, primarily involving avoidance (severity score 6 – survival track), but in one case involving female–calf separation (severity score 8 – reproductive track). This contained the most severe responses observed in any study considered.

Isojunno et al. (2016) measured responses of sperm whales (*Physeter macrocephalus*) to several different sonar signals (those used by Miller et al., 2014) and analyzed responses in terms of behavioral state switching. Individual whales were monitored using tags before, during, and after exposure to multiple sonar types, but data were analyzed and presented across individuals within sonar treatment types. Assessors were generally in good agreement in terms of the type and nature of response, although one assessor noted avoidance in one instance but agreed in adjudication it was not sufficiently supported by the data. One score is recorded per sonar treatment type given

Table 5. Adjudicated response severity assessment – Active sonar sources

Study	Study type	Species studied	Assessment summary	Subject	Sonar stimulus	Individual weighting	Adjudicated score	Associated RL at change point or max. RL if no response	
								RMS	SEL
Tyack et al. (2011)	Combined study with experimental and incidental exposures to simulated and real navy active sonar	Blainville's beaked whale (*Mesoplodon densirostris*)	Very good agreement between assessors on exposure, context, and response severity. One individual received two exposures, responding to both. Sustained avoidance was most common.	Md07	Sonar (3.5-4.0 kHz)	0.5	5	138	142 SEL, 152 cSEL
				Md07	Killer whale	0.5	6	98	
				Md08	PRN (3.5-4.0 kHz)	0.0	5	142	144
				Md08	PRN (3.5-4.0 kHz)	1.0	6	142	144
				Satellite tag	3.5-4.0 kHz (incidental)	1.0	6	146	144 cSEL
DeRuiter et al. (2013)	Experimental study with known and documented exposure and fine-scale aspects of individual responses	Cuvier's beaked whale (*Ziphius cavirostris*)	Very good agreement among three assessors in terms of exposure conditions and clear response occurrence and severity. Slight differences among scorers about most severe score, but convergence on sustained avoidance and failure of vocal mechanism.	Zc10_272a	3.5-4 kHz	0.0	1	98	
				Zc10_272a		0.0	2	98	
				Zc10_272a		0.0	6	98	
				Zc10_272a		1.0	7	98	
				Zc11_267a		0.0	1	127	
				Zc11_267a		0.0	2	127	
				Zc11_267a		0.0	4	127	
				Zc11_267a		0.0	6	127	
				Zc11_267a		1.0	7	127	
Hastie et al. (2014)	Experimental study in captive setting; exposures well controlled and quantified.	Gray seal (*Halichoerus grypus*)	Good agreement regarding exposure and response. Slightly different interpretations using field severity scale but adjudicated to different severities for remaining in water (avoidance onset) vs hauled out (sustained avoidance).	Seal 1	BioSonics	0.5	6	160.3 (1/3-octave level)	
				Seal 2		0.5	6		
				Seal 1	Coda Octopus	0.5	5	165.7 (1/3-octave level)	
				Seal 2		0.5	5		

Study	Study type	Species studied	Assessment summary	Subject	Sonar stimulus	Individual weighting	Adjudicated score	Associated RL at change point or max. RL if no response	
								RMS	SEL
Miller et al. (2014)	Experimental study with known and documented exposure and fine-scale aspects of individual responses	Killer whale (*Orcinus orca*)	Very good agreement among three assessors in terms of exposure conditions and response occurrence and severity, with slightly different approaches to scoring multiple exposures/responses readily resolved	oo317a	1-2 kHz	1.0	6	150	158
				oo327s(t)	6-7 kHz	1.0	6	138	143
				oo149a	6-7 kHz	0.3	0	142	166
				oo149a	1-2 kHz	0.3	0	166	176
				oo149a	6-7 kHz	0.3	8	133	133
				oo144a	1-2 kHz	0.3	6	94	94
				oo144a	6-7 kHz	0.3	6	94	95
				oo144a	1-2 kHz	0.3	6	164	171
Isojunno et al. (2016)	Experimental study with documented exposures and responses within individuals. Responses analyzed across individuals within sonar types.	Sperm whale (*Physeter macrocephalus*)	Good agreement about group responses to different treatments. One score per sonar type is presented given how data were pooled. Reduction in foraging was identified for one sonar type; no responses scored for others.	Multiple whales	1-2 kHz	1.0	5	133 (min. of range)	
				Multiple whales	6-7 kHz	1.0	0	158 (RMS; max. of range)	
				Multiple whales	4.7-5.1 kHz	1.0	0	133 (max. of range)	

that multiple whales were included in each treatment type; individual-level data were obtained but could not be scored based on data as reported. A reduction in foraging (severity score 5 – foraging track) was detected for one of three sonar types.

Industrial (Continuous) Sources—Behavioral responses of marine mammals to other continuous sources such as offshore drilling, construction, or vessels has also been the subject of extensive research for decades (see Southall et al., 2007; NAS, 2017). During our study, we found a large number of papers that failed to provide sufficient details about noise source parameters and operations and were thus deemed low or moderate priority studies. We selected and scored five studies for assessment from the few high priority studies identified—Malme et al. (1986), Gordon et al. (1992), Nowacek et al. (2004), Holt et al. (2009), and Kendall et al. (2013)—and several were deemed moderate in terms of the number of exposure and response parameters provided (see Table 6).

Malme et al. (1986) conducted experimental playbacks of recorded vessel and drilling noise and seismic airguns to migrating gray whales (*Eschrichtius robustus*). The industrial (continuous) noise exposures are considered here, and the seismic airgun surveys are scored separately (below). This published technical report provides extensive details at the individual level of exposure(s) and response. It was noted that the context of and relatively low source levels for drilling sounds played back relatively closer to animals differ from full-scale operational sources. There was very good agreement between reviewers given extensive details on individual exposure and response with some slight differences between reviewers in scoring onset of avoidance (where clear movement away from the noise source is first observed) vs sustained avoidance (where directed movement is maintained during and especially following exposure). This was adjudicated to the former.

Gordon et al. (1992) conducted an observational study of potential responses of sperm whales to whale-watching vessels in different contexts. Multiple observations were made of groups with different compositions of individuals and presumed familiarity with such vessels. Individual acoustic and diving behavior was observed but reported at the group level. There was moderately good agreement among assessors on the nature of responses, with some discussion required regarding exposure metrics and weighting the observations. Because it was challenging to assess the relative differences between animal groups in experience or familiarity with sources with certainty, assessors adjudicated to scoring a single observation for this study. Assessors concurred on the presence of detectable changes in diving and in subsurface interval/time, the latter being the more severe response and the effective score assigned. Insufficient information about exposure was provided to report explicit corresponding RLs with the observed responses. It was noted that this type of study would be better assessed within the context of longer-term severity assessment.

Nowacek et al. (2004) conducted CEEs with North Atlantic right whales (*Eubalaena glacialis*) using a variety of industrial (vessel) and experimental (alarm) stimuli. Given the low source level of all stimuli, they were projected at relatively close range for all exposures. Extensive and detailed individual exposure and response parameters were measured continuously using tags supporting scores for all individuals (weighted by number of exposures as needed). Assessors were in complete agreement for individual responses and individual weighting, as well as exposure and response conditions, given the detail provided. Some discussion was required to adjudicate minor differences in severity of foraging changes and inclusion of minor changes in diving.

Holt et al. (2009) studied vocal responses of killer whale groups to vessel presence. Group observations of one feature of vocal characteristics (estimated source levels) were observed in different conditions of vessel exposure. Assessors generally agreed on both the group weighting (a single observation for the study) and the nature and severity of response being brief/minor changes in vocal rates or signal characteristics. Insufficient details were provided on noise exposure conditions at the group level to assign a RL (or range); maximum levels given are not applicable since they are not associated with the response that was detected.

Kendall et al. (2013) studied acoustic responses of beluga whales (*Delphinapterus leucas*) to construction noise in an observational study of group vocal (clicking) behavior. Group observations were made in relation to construction noise, but details on noise exposure conditions or ranges for groups were not reported. There was good agreement about the nature of response and reporting weighting as a single observation. A reduction in click rates was detected, which could either reflect movement from the area or reduction in vocal output. Since this trend of reduction was not significant, it was considered a potential vocal response but of the least possible severity. It was again noted that this type of study is also likely more appropriately considered with population-level assessments.

Pile-Driving Sources—Extensive research has been conducted on the potential effects of noise associated with the installation of offshore facilities, notably the noise associated with pile

Table 6. Adjudicated response severity assessment – Industrial (continuous) sources

Study	Study type	Species studied	Assessment summary	Subject	Noise stimulus	Individual weighting	Adjudicated scores	Associated RL at change point or max. RL if no response	
								RMS	SEL
Malme et al. (1986)*	Controlled exposure experiments with recorded vessel and drilling noise. Report extensive detail on individual exposure(s) and response.	Gray whale (*Eschrichtius robustus*)	Very good agreement between assessors on weighting individual exposures. Some adjudication required between onset of avoidance (5) vs sustained avoidance (6); adjudicated to former.	Q	Vessel noise	0.5	5		
				Q	Drill ship #2	0.5	0	110	
				M	Vessel noise	1.0	5		
				W	Drill ship #3	1.0	0	106	
				Not specified	Drill ship #5	1.0	5	108	
				K	Vessel noise	0.5	5		
Gordon et al. (1992)*	Observational study of potential group acoustic and diving responses to vessel presence	Sperm whale	Moderate agreement on response and manner of reporting. Some discussion of segregating familiar and unfamiliar whales but adjudicated to a single observation with detectable changes in diving (1) and changes in subsurface interval/time (4); insufficient data to report received levels (RLs).	Groups of sperm whales	Vessel noise (multiple vessels)	0.0	1	N/A	
						1.0	4		
Nowacek et al. (2004)	Experimental study with well-described individual exposures and responses reported. Vessel noises were familiar, but alarm stimulus was likely unfamiliar to whales.	North Atlantic right whales (*Eubalaena glacialis*)	Very good agreement between assessors in individual weighting, exposure parameters, and response severity. Some discussion required to adjudicate minor differences in foraging response severity of foraging changes and inclusion of minor changes in diving.	02_213g	Alarm	0.3	4	148	
				02_213g	Vessel	0.3	0	135	
				02_213g	Vessel	0.3	0	142	
				02_220f	Alarm	1.0	4	143	
				Eg2350	Alarm	1.0	4	137	
				Eg3109	Alarm	0.5	4	137	
				Eg3109	Vessel	0.5	0	133	
				02_232d	Alarm	0.3	4	133	
				02_232d	Vessel	0.3	0	136	
				02_232d	Vessel	0.3	0	132	
				02_233a	Vessel	1.0	0	136	
				Eg3103	Alarm	0.5	0	134	
				Eg3103	Vessel	0.5	0	129	
				Eg2145	Vessel	1.0	0	133	
				Eg1142	Vessel	1.0	0	139	

Study	Study type	Species studied	Assessment summary	Subject	Noise stimulus	Individual weighting	Adjudicated scores	Associated RL at change point or max. RL if no response	
								RMS	SEL
Holt et al. (2009)	Observational study measuring potential changes in group vocal response with variable levels of vessel presence	Killer whale	Good agreement between assessors about nature of exposure and response. Concurrence on single observation, nature of brief vocal response, and absence of sufficient data to assign group-specific RL associated with response.	J-pod	Vessel noise (multiple vessels)	1.0	4	N/A	
Kendall et al. (2013)	Observational study of potential group vocal behavior in relation to construction noise	Beluga whale (*Delphinapterus leucas*)	Good agreement among assessors about nature of response and individual (group) weighting as a single observation given details provided. Changes were detected in click rates but considered the least severe possible responses. Insufficient details were provided to associate change with RL.	Groups of beluga whales detected acoustically	Construction noise/ presence	1.0	1	N/A	

*In some instances, multiple different types of noise sources were used in some studies, which are described parenthetically.

driving, which is impulsive at the source. Many of the published studies on the effects of pile-driving noise are observational rather than experimental in nature, and many focus on the local population rather than the individual level, making application of our assessment for response type and severity more challenging as it is not designed for assessing population responses to long-duration activities. We selected from identified high and moderate priority papers as described above and herein provide assessments for Blackwell et al. (2004), Brandt et al. (2009, 2011), Tougaard et al. (2009), and Thompson et al. (2010) (see Table 7).

Blackwell et al. (2004) conducted an observational study of responses of ringed seals (*Phoca hispida*) hauled out or in the water in the course of construction (pile driving) and operations (helicopter noise) associated with an offshore energy facility. Individual responses were reported, but individual identity was not known across days, meaning that individuals could have been exposed on multiple occasions. Multiple stimulus types were considered, some occurring simultaneously, with aerial and underwater noise conditions reported. Maximum levels were given at the position of the closest seal but were not reported for all individuals. Assessors were in moderate agreement, with some different interpretations of including individual-level data and how to report RLs associated with scores. Assessors agreed on the nature of responses, which included mild orienting responses with a single incident of a seal abandoning a haulout, and they concurred upon adjudication to report individual incidents separately as presented. Maximum RLs were determined here only for individuals at or within 10 m distance reported for closest seal.

Brandt et al. (2009) is a technical report documenting potential responses of harbor porpoises in the area to the construction of an offshore wind facility; Brandt et al. (2011) is the peer-reviewed publication that followed. Both were scored here in part to evaluate relative differences in information provided and assessments drawn from a more expansive technical report and a peer-reviewed publication. Assessors had slightly different interpretations of the results from the same study published in separate formats. In each, group vocal responses (changes in clicking behavior) are reported, but it is difficult to distinguish whether the reported reduction in clicks represented reductions in foraging or avoidance of disturbed areas or both. For each publication, assessors agreed on reporting a single value given the unit of observations, which was not specific to an individual or a group. Maximum RLs reported were not included with a severity score since response results were pooled across multiple recorders, not all of which included measured levels. Assessors identified a minor reduction of vocal output and possible sustained avoidance in Brandt et al. (2009) as changes, adjudicating to include both scores given ambiguity on the nature of this response; whereas changes reported in Brandt et al. (2011) were scored as reduced vocal mechanisms and possible sustained avoidance. Both scores are reported with a differential weighting proposed for Brandt et al. (2011) as both scores reported had the same severity, though it is noted that effectively one observation with a severity score of 6 was determined for each publication.

Tougaard et al. (2009) conducted a similar observational study with group vocal responses of harbor porpoises to pile-driving noise. There was good agreement among assessors regarding use of a single observation for this group-level study. Aspects of the study suggest habitat avoidance in addition to vocal reduction; assessors scored both but rated a lesser severity to vocal reduction (severity score 4). Source-level estimates and some RL measurements are given, but assessors found these challenging to assign to a specific exposure and to relate to the severity scores. Assessors noted that this was an important and clear study but, given the broad spatial extent of the study and the lack of a clearly defined response onset and cessation, it would likely be more appropriately evaluated by using methods for assessing severity on broader spatial and longer temporal scales.

Thompson et al. (2010) observed group vocal activity for three odontocete species (harbor porpoise, common bottlenose dolphin [*Tursiops truncatus*], and common dolphin [*Delphinus* spp.]) in the presence of offshore wind power installations. The nature of these observations was similar to those in Brandt et al. (2009, 2011) with similar challenges in determining whether reduced click rates reflect changes in vocal rates or habitat avoidance. This study required substantial discussion and adjudication among assessors, primarily centered around whether and how to report data by species. Assessors ultimately determined not to report scores for *Tursiops* or *Delphinus* given challenges in distinguishing among delphinid clicks. Scores for *Phocoena* were reported and equally weighted as sustained avoidance and vocal reduction. Insufficient data were provided to assign specific RLs in terms of minimum levels associated with onset of response.

Seismic Airgun Survey Sources—Some of the earliest quantitative analyses of human noise impacts on marine mammals concerned the potential effects of seismic airguns on whales (see Richardson & Würsig, 1997). In recent decades, this has remained an area of active research involving experiments wherein the investigators

Table 7. Summary assessment for selected scored papers – Pile-driving sources

								Associated RL at change point or max. RL if no response	
Study	Study type	Species studied	Assessment summary	Subject	Stimulus	Weighting	Adjudicated scores	RMS	SEL
Blackwell et al. (2004)	Observational study of responses in the course of several categories of known operations. Individual responses reported, but individual identity not given.	Ringed seal (*Phoca hispida*)	Moderate agreement among assessors on response type and severity, with some differences regarding whether to combine into a single observation or split out individuals as they could not certainly be distinguished individually. Adjudication was to report individual incidents separately. Exposure conditions were not reported individually, so decision was to provide max. RLs only for individuals at or within 10 m of value reported for closest seal.	Not specified	Piling + helo	1.0 (water)	1	151	145
					Piling + helo	1.0 (water)	1		
					Piling + helo	1.0 (water)	1		
					Piling + helo	1.0 (water)	1		
					Piling + helo	1.0 (water)	1		
					Piling + helo	1.0 (water)	1		
					Piling + helo	1.0 (water)	1		
					Piling + helo	1.0 (water)	0		
					Piling + helo	1.0 (air)	5		
					Piling + helo	1.0 (air)	1	96	90
					Piling + helo	1.0 (air)	1	96	90
					Pile driving	1.0 (water)	0	151	145
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0 (water)	0		
					Pile driving	1.0	0		
					Helicopter	1.0	0		
Brandt et al. (2009)	Observational study of group vocal responses to pile-driving noise stimulus. This is the technical report which precedes Brandt et al. (2011) (see below).	Harbor porpoise (*Phocoena phocoena*)	Good agreement on approach to weighting (single value given group nature of observations) and nature of response (minor reduction of vocal output) and possible sustained avoidance given vocal reductions as changes. Max. RLs were not included here since response results were pooled across multiple recorders.	Harbor porpoise local population	Pile driving	0.0	4		
				Harbor porpoise local population	Pile driving	1.0	6		

Study	Study type	Species studied	Assessment summary	Subject	Stimulus	Weighting	Adjudicated scores	Associated RL at change point or max. RL if no response	
								RMS	SEL
Tougaard et al. (2009)	Observational study with well-defined group acoustic responses of harbor porpoises to pile driving given nature of observations for which individuals were not identified and RLs were not known	Harbor porpoise	Good agreement among assessors regarding single observation for this group-level study; this could be more appropriate to consider with group-focused severity assessment. Some minor differences regarding inclusion of either vocal reduction or sustained avoidance. Adjudicated to report both, but to include less severe vocal reduction and sustained avoidance; the latter is the effective score. Source-level estimates for RLs and some measurements were given, but assessors found it challenging to assign to a specific exposure.	Harbor porpoise local population	Pile driving	0.0	4		
				Harbor porpoise local population	Pile driving	1.0	6		
Thompson et al. (2010)	Observational study of group vocal activity for three cetacean species in context of offshore wind energy installation	Harbor porpoise Bottlenose dolphin (*Tursiops truncatus*) Common dolphin (*Delphinus* spp.)	Poor initial agreement among assessors. Through adjudication, agreement that given challenges in distinguishing delphinids based on clicks, scores for *Tursiops* or *Delphinus* would not be reported. Scores for *Phocoena* reported as both sustained avoidance and vocal reduction. Insufficient information reported to assign specific RLs.	Harbor porpoise local population	Pile driving	0.5	6		
				Harbor porpoise local population	Pile driving	0.5	6		
Brandt et al. (2011)	Observational study of group vocal responses to pile-driving noise stimulus. This is the peer-reviewed publication from the Brandt et al. (2009) technical report (see above).	Harbor porpoise	Good agreement on approach to weighting (single value given group nature of observations) and nature of response. However, given the details and explanation provided here, reduced vocal mechanisms and possible sustained avoidance given vocal reductions (equal severity) were identified as changes. Max. RLs were also not included here since response results were pooled across multiple recorders.	Harbor porpoise local population	Pile driving	0.5	6		
				Harbor porpoise local population	Pile driving	0.5	6		

direct sound sources or monitor ongoing operations. We selected experimental studies in the high priority category for this source type, scoring five studies: Madsen et al. (2002), Miller et al. (2009), Dunlop et al. (2015, 2017), and Gailey et al. (2016) (see Table 8). We also scored the seismic airgun exposures reported in Malme et al. (1986).

Madsen et al. (2002) was an experimental study with sperm whales using visual observations at the individual level accompanied by acoustic monitoring. Two different exposure scenarios were reported with multiple individuals at different distances from sources. No observable responses were reported for any individual (i.e., right-censored data as defined above); maximum levels are thus appropriate to associate with no response. The assessors generally agreed on relevant details of exposures and lack of observable response(s). Assessors differed as to whether to report an overall response and weight by number of individuals, but ultimately agreed that sufficient detail was provided to support three separate scores, each with associated response scores and received noise levels.

Miller et al. (2009) exposed sperm whales to known and controlled exposures to a full-scale seismic survey. Responses were measured and reported at the individual level with archival movement and acoustic tags, and augmented with visual observations. There was very good agreement among assessors regarding exposures and response at the individual level. Behavioral changes were scored for almost all individuals classified as either vigilance and/or reduced foraging. A slight difference among assessors initially regarding responses as vigilance or reduced foraging was adjudicated to the more severe response based on reduction in foraging rate.

Dunlop et al. (2015) conducted experimental exposures of tagged individual humpback whales to a single seismic airgun. Behavior was characterized from passive acoustic recordings as well as observational tracking at individual and group levels. Responses were considered and reported within age classes, which were assumed but not known to be unique individuals since the study was in the context of migratory behavior. Received noise conditions were not quantified at the individual level, though broad ranges of exposure conditions were given. Assessors did not initially agree in terms of segregating groups/individuals into separate exposure instances or in terms of severity of response. Upon adjudication, it was agreed to report a single value for the study given limited information about individual exposure conditions, mixed contexts across conditions/age classes, and some level of uncertainty on whether individuals were observed repeatedly. The single observed score was identified as a minor deviation in the typical migratory pathway.

Gailey et al. (2016) studied behavior of gray whales incidentally exposed to an ongoing seismic survey. Individuals were observed and tracked during operations; two focal case studies were reported. Extensive multivariate data were presented on movement before, during, and after exposure to seismic airgun signals, but limited information was provided on individual exposures and responses. Assessors were in good agreement that it was not possible to conduct a severity assessment for this study given information available in the paper. It was noted that while this was a useful study with an extensive, expensive, and complex effort, when results were compiled within complex multivariate models without individual exposure details provided either within the paper or supplementary materials, exposure and response are challenging to assess for severity scoring.

Dunlop et al. (2017) conducted CEEs with tagged individual humpback whales exposed to a seismic airgun array. While individual exposure and response were known, they were reported and analyzed collectively, making it impossible to score individual responses. Exposure-response contexts of RL and source proximity were both used in a multivariate analysis. Avoidance responses were observed for some individuals, and these were scored as a group response by assessors. There was generally good agreement among assessors for reporting a single score given the nature of the analyses. Assessors noted that for this type of study where individual details were obtained and known—whether provided in supplementary materials or linked to the paper—additional individual scoring with this severity scale would be possible. Given the expense and difficulty of conducting this kind of study, detailed reporting of the individual cases would be advisable, enabling future analyses.

Malme et al. (1986) conducted experimental playbacks of recorded drilling noise and vessel noise (described above), as well as operational seismic airguns, with migrating gray whales. There was very good agreement among assessors given extensive detail on individual exposure and response parameters. Responses included minor deviations from the migratory pathway, changes in locomotion, and brief avoidance.

Population-Level Study Results

Multi-Year Whale-Watching Studies—A number of studies have empirically evaluated multi-year effects of whale-watching disturbance on vital rates or predicted long-term consequences for both mysticetes and odontocetes. We evaluated selected

Table 8. Summary assessment for selected scored papers – Seismic airgun survey sources

Study	Study synthesis	Species studied	Assessment summary	Subject	Stimulus	Weighting	Adjudicated scores	Associated RL at change point or max. RL if no response	
								RMS	SEL
Malme et al. (1986)	Experimental playback of recorded drilling noise and experimental seismic airgun noise to migrating whales. Detailed technical report with extensive detail on individual exposure and response parameters.	Gray whale	Airgun exposure–response exposures reported here; drilling and vessel noise addressed in Table 6. Very good agreement between reviewers given extensive individual details on exposure and response.	E	Single airgun (100 cu in)	0.5	5	149	
				E		0.5	1	172	
				A + E		1.0	4	154	
				B		1.0	5	165	
				B		1.0	5	176	
				K		1.0	4	160	
				L		1.0	4	170	
				N		1.0	5	174	
Madsen et al. (2002)	Experimental study with individual visual observations and acoustic monitoring. Two different exposure scenarios were reported with multiple individuals at different physical ranges.	Sperm whale	Good agreement among assessors regarding exposures and lack of observable response. Some initial differences about reporting an overall response weighted by number of individuals, but ultimately provided three separate scores, each with associated scores and responses.	Sperm whale	Sleeve airgun array with four sub-arrays of 10 guns	1.0	0	123	
						1.0	0	130	
						1.0	0	130	
Miller et al. (2009)	Experimental study with known and controlled exposures and responses measured and reported at individual level. Individual exposure and response were obtained from tag data and augmented with visual observations.	Sperm whale	Very good agreement among assessors regarding exposures and nature of response at the individual level. Some initial differences as to responses as vigilance or reduced foraging; adjudicated to more severe response and agreed; one individual did not respond.	sw253a	Seismic airgun array	5.0	1	144	127
				sw254a		5.0	1	140	123
				sw254b		5.0	1	135	118
				sw254c		5.0	1	139	123
				sw164a		5.0	1	146	129
				sw165a		5.0	1	146	130
				sw165b		5.0	1	147	130
				sw173b		0.0	0	147	131

Study	Study synthesis	Species studied	Assessment summary	Subject	Stimulus	Weighting	Adjudicated scores	Associated RL at change point or max. RL if no response	
								RMS	SEL
Dunlop et al. (2015)	Experimental study with tagged individuals and controlled exposures to single seismic event. Reported behavior from archival sensors and visual tracking.	Humpback whale (*Megaptera novaeangliae*)	Poor agreement among assessors on this study, both in terms of effective number of observations and in severity of response. A single score was ultimately generated given limited information about exposure conditions, contexts spanning age classes, and some uncertainty on whether repeat individuals observed. Minor deviations in migratory pathways were observed.	Local population of migrating whales	Single seismic airgun (20 cu in)	1.0	1	N/A (SNRs all > 10 dB)	108
Gailey et al. (2016)	Observational study of behavioral responses of two focal whales to an ongoing seismic survey. Extensive movement data but limited individual exposures and response data.	Gray whale	Assessors in good agreement; it was not possible to apply severity assessment given information available in the paper. Response results are compiled within multivariate models without individual detail making it challenging to assess in terms of exposure-response severity.	Two case studies	4-D seismic airgun survey – array (volume not given)	N/A	N/A		
Dunlop et al. (2017)	Experimental study with individuals tagged and known; controlled exposure experiments. Exposures and responses reported at population level.	Humpback whale	Good agreement among assessors on reporting a single score given nature of analyses and on type and severity of response (sustained avoidance in some individuals). Assessors assigned single overall RL representative of most exposures here.	Local population of migrating whales	Seismic airgun arrays (20 and 140 cu in)	1.0	6		140 (overall est. as change point; hard to assign to individual)

studies which specifically investigated whale-watching effects on vital rates and which were comprised of data collected in impact and control situations over multiple years (Williams et al., 2002, 2006; Lusseau, 2004, 2005; Lusseau et al., 2006; Christiansen & Lusseau, 2014, 2015). For each of these studies, we provided a summary assessment for a selected example based on the structured review process described in Table 4 and a short synthesis of each paper (Table 9); additional details for other studies are given in Appendix 2. Overall, the studies examined reports that whale-watching activities influence cetacean behavior that, in turn, can have repercussions on vital rates. However, the impact of such changes largely depends on individual cumulative exposure through time as well as concurrent environmental stressors and additional anthropogenic activities.

Christiansen & Lusseau (2015) quantified the effect of whale watching on the energetic expenditure of minke whales (*Balaenoptera acutorostrata*) on a feeding ground using 4 y of data. The whale-watching industry in Flaxafoi Bay, Iceland, started in 1991 (Martin, 2012) and now coexists with the whaling industry, reinstated in 2006 in waters neighboring the whale-watching area (Bertulli et al., 2014). Christiansen & Lusseau (2015) quantified individual exposure to vessels as a function of duration and encounter rate and documented seasonal and yearly variation reflecting the industry's seasonal variation and inter-individual differences in whale behavior. They documented that whale watching can affect body condition in minke whales, but the low cumulative exposure of individual minke whales to vessels in the area indicated that levels of disturbance at that time did not have significant long-term effects on vital rates. However, individual variation in estimated exposure to disturbance increased with increasing whale-watching prevalence. Moreover, whales' use of the area is strongly influenced by the presence of sand eels. In years of low prey density, whales spent more time in the area and, thus, were subjected to greater exposure to whale-watching disturbance. Adaptive management strategies in years of low sand eel density could therefore be adopted to mitigate interactive effects of higher disturbance during periods of reduced prey availability.

Williams et al. (2002) investigated the effect of whale watching on northern resident killer whales (NRKWs) in Johnstone Strait, Canada, documenting sex differences in responses to disturbance. Williams et al. (2006) subsequently investigated the energetic cost of anthropogenic disturbance on NRKWs in the presence and absence of whale-watching vessels. Activity budgets were then converted to energetic cost to calculate the energetic consequences of disturbance. Whales reduced their amount of time feeding and increased traveling; and while the associated energetic cost of active avoidance is estimated to be just 3 to 4%, the concurrent loss of feeding opportunities resulted in an estimated 18% decrease in energy gain. However, the 5 y of data collected were insufficient to measure population fluctuations. Disturbance in Johnstone Strait at that time was apparently sufficiently limited such that NRKWs were able to balance the energetic cost of avoiding boats without detrimental consequences, although concerns regarding auditory masking effects that could decrease prey detectability were identified. Similar findings of behavioral disruption have also been reported for the endangered southern resident killer whale (SRKW) population (Williams et al., 2006, 2009b). These highly social odontocetes rely on echolocation to locate fish prey and to live in high-traffic, prey-depleted areas (Williams et al., 2011), with high levels of whale-watching disturbance that have been identified as factors related to population decline (Williams et al., 2006).

Decadal-Scale Whale-Watching Studies—A small number of studies have evaluated decadal-length effects of whale-watching disturbance on vital rates and linked these to long-term consequences. We evaluated selected studies with sufficient available data (Bejder & Samuels, 2003; Bejder et al., 2006; Weinrich & Corbelli, 2009; Filby et al., 2014; Senigaglia et al., 2019) and provide a summary assessment and synthesis of each in Table 10, with selected examples discussed in detail below and others considered in Appendix 2.

We identified a single study that quantified disturbance in terms of population dynamics for mysticetes. Weinrich & Corbelli (2009) measured calf production and survival of humpback whales in southern New England where the whale-watching industry began in 1960 and presently includes over 15 companies on Stellwagen Bank alone. Data from 1980 to 2006 were used to explore calving rates and calf survival rates at ages 1 and 2 y in relation to exposure (measured as total exposure time and total number of boat–whale interactions). The exposure of female humpback whales to whale watching during conception and pregnancy was compared against successful calving events while prey density (mean number of sand lance) was accounted for as a possible confounding factor during putative pregnancy years. There was no evidence of reduced female reproduction or calf survival. Prey availability was not correlated with calving rates but influenced calf survival at both ages 1 and 2 y. Weinrich & Corbelli concluded that, at that time, the industry was deemed sustainable for humpback whales.

Bejder et al. (2006) documented a decrease in relative abundance of bottlenose dolphins due to increased whale-watching pressure in Shark Bay,

Table 9. Summary of studies evaluating multi-year effects of whale-watching disturbance on vital rates or predicted to cause long-term consequences. Codes of study characteristics (based on the key components defined in Bejder & Samuels, 2003), source of human disturbance, whale-watching industry characteristics, and short- and long-term measures refer to factors listed in Table 4. Definitions of area use type (year-round habitat, breeding ground, feeding ground, migratory corridor, multiple use, etc.) followed Weinrich et al. (2010). Target species are listed with abbreviation of their scientific names and include minke whales (*Balaenoptera acutorostrata*; Ba), Indo-Pacific bottlenose dolphins (*Tursiops aduncus*; Ta), and killer whales (*Orcinus orca*; Oo). N/A = data not available.

Study characteristics	Data collection length (y)	Location	Area use type	Source of anthropogenic disturbance	Whale-watching industry characteristics	Target species	Exposure measure	Short-term measures	Long-term measures	Detailed measured or modeled results	Reference
E O B WE I	4	Iceland	Feeding ground	WW	L LI WW G	Ba	Encounter rate; encounter duration	E BB V	N/A	Whale-watching causes decrease in energy acquisition of 42.1% and increase in energy expenditure by 27.6%. Low cumulative exposure and most impacted individuals would lose 0.66% energy needed to carry on pregnancy.	Christiansen & Lusseau, 2014, 2015
E O B CI G	3	Pacific Ocean, New Zealand	Year-round habitat	WW	L LI WW P	Ta	Exposure time (h/d)	BB V	PD	Decreased resting behavior. The effect was stronger in Doubtful Sound where socializing behavior was also strongly impacted. Dive interval for females linearly increased with the number of violations committed during an interaction, causing a 18.6% increase in dive interval when one violation occurred and a 37.1% increase for more than one violation. Drastic population decline in Doubtful Sound forecasted within 30 to 70 y with current level of whale-watching disturbance (> 68 min/encounter).	Lusseau, 2004, 2005; Lusseau et al., 2006
E O L CI G	8	Pacific Ocean, Canada	Year-round habitat	WW BT PA C	L LI WW G	Oo	Presence vs absence	BB H E	N/A	Reduced feeding time from 13 to 10% and increased their travel budget by 12.5% when boats were present. Increased energetic expenditure of 3 to 4% due to avoidance and concerning 18% estimated decrease in energy intake due to lost feeding opportunities.	Williams et al., 2002, 2006

Table 10. Summary of studies evaluating very long-term (decadal) effects of whale-watching activities. Codes of study characteristics (based on the key components defined in Bejder & Samuels, 2003), source of human disturbance, whale-watching industry characteristics, and short- and long-term measures refer to factors listed in Table 4. Definitions of area use type (year-round habitat, breeding ground, feeding ground, migratory corridor, multiple use, etc.) followed Weinrich et al. (2010). Target species are listed with abbreviation of their scientific names and include humpback whales (*Megaptera novaeangliae*; Mn) and Indo-Pacific bottlenose dolphins (*Tursiops aduncus*; Ta).

Study characteristics	Data collection length (y)	Location	Area use type	Source of anthropogenic disturbance	Whale-watching industry characteristics	Target species	Exposure measure	Short-term measures	Long-term measures	Detailed measured or modeled results	Reference
E H B CI I	26	Atlantic Ocean, USA	Feeding ground	WW BT	L HI WW P G R	Mn	Exposure time; # boat–animal interactions	N/A	CR CS	No significant correlation between encounter rate nor duration on either measure of reproductive success. No effect of confounding factor (prey density).	Weinrich & Corbelli, 2009
E O B WE G	3+2, 15 y apart	Pacific Ocean, Australia	Year-round habitat	WW	L HI SW P G R	Ta	Encounter duration; sighting rate	H V	SR	Increase of effect (avoidance or attraction) responses of 52.4%; significant decrease in sighting success rate between the two seasons from 59.4 to 46.6%.	Filby et al., 2014
E H B+L CI G	15	Indian Ocean, Australia	Year-round habitat	WW	L LI WW P G R	Ta	Exposure time (h/d)	H V	PD	Time spent with dolphins increased by 74.9% and average abundance dropped by 14.9%/km^2 (1 of every 7 individuals) due to the additional whale-watching operator.	Bejder & Samuels, 2003; Bejder et al., 2006
E O+H B CI I	20	Indian Ocean, Australia	Year-round habitat	WW O	P	Ta	Provisioning status and frequency	N/A	CS	Provisioned females showed half the reproductive success of non-provisioned females, and only 1/3 of calves born to provisioned mothers were successfully weaned.	Senigaglia et al., 2019

Western Australia. This area is inhabited by one of the longest studied populations of bottlenose dolphins for which research activities started before the whale-watching industry began in 1993. They compared relative dolphin abundance between two adjacent sites ("control" and "tourism") with different anthropogenic pressures across three consecutive 4.5 y time periods, corresponding to before commencement of whale watching, with one tour vessel operating (up to 4 trips per day) and with two vessels operating (up to 8 trips per day combined). Bejder et al. demonstrated significant increases in dolphin exposure to tour vessels with concomitant decreases in the average relative abundance in the tourism site. Increases in relative dolphin abundance in the adjacent control site were not detected. The impact of whale-watching disturbance may be negligible on the large and genetically diverse Shark Bay dolphin population, but, according to Bejder et al., the effect of tour vessels on dolphin abundance in a region of low-level tourism calls into question the presumption that dolphin-watching tourism is benign.

Multi-Year to Decadal-Scale Studies of Other Disturbances—A relatively small number of long-term (many year) studies have evaluated disturbances from non-whale-watching anthropogenic activities on cetacean vital rates and linked these to long-term population consequences. We selected and evaluated selected studies that reported sufficient information (New et al., 2013, 2014; Pirotta et al., 2014, 2015) and provide a summary assessment and synthesis of each paper in Table 11, with additional details provided in Appendix 2.

Multi-Year and Predictive Studies on Population-Level Consequences of Multiple and Concurrent Threats via Modeling Approach—New technologies are being developed to allow rapid assessment of vital rates in the wild (Booth et al., 2020; Christiansen et al., 2020; Currie et al., 2021). However, given the challenges of obtaining such measurements over multiple generational time frames, studies have increasingly employed modeling approaches to assess the effects of long-term disturbance and multiple stressors (Pirotta et al., 2019). Ecological and behavioral models allow the consideration of multiple potential sources of cumulative impacts in predicting population-level consequences from short-term behavioral responses. Quantitative tools can be used to forecast population trajectories under different circumstances (i.e., additional management actions implemented, increased human pressure, climate change, etc.), including decision trees and simulation-based analyses such as population viability analyses (PVAs) (Strindberg & O'Brien, 2012; Lacy et al., 2017). Such methods incorporate the combined effects of stochastic processes, such as demographic, environmental, and genetic variability, and deterministic processes (e.g., overexploitation, human disturbance, habitat degradation) to simulate population dynamics and assess potential extinction risk (Himes Boor, 2014). Moreover, a theoretical framework is currently being developed to quantify the Population Consequences of Multiple Stressors and assess the potential conservation threat associated with the exposure of wildlife to multiple stressors. Challenges in the implementation of both PVA and PCoMS remain in the large volume of data required, encompassing multiple years and considerable amounts of funding. Few studies successfully employed such methodologies using empirical data (Lacy et al., 2017; Senigaglia, 2020).

Lacy et al. (2017) explored the cumulative effects of multiple stressors on the endangered population of SRKWs in the North Pacific with the aim of prioritizing conservation actions and guiding the implementation of effective management plans. PVAs were used to model population growth under different scenarios which varied in the levels of anthropogenic stressors (including noise disturbance measured as loss of acoustic communication space), pollution, and prey availability. The relative importance of anthropogenic stressors was also assessed. The comprehensive dataset used encompassed multiple decades of killer whales' demographic parameters and 40 y of data on Chinook salmon (*Oncorhynchus tshawytscha*) survival and reproduction. The results of this study confirmed the fragility of the population to any increase in disturbance, with prey limitation being the most influential determinant of a potential population decrease. However, the authors conclude that mitigation of SRKW population decline via prey management alone would be unfeasible, while reducing noise and chemical pollution would be insufficient to meet the recovery target. Only a combination of decreased disturbance and a 15% increase in prey availability would allow for a forecasted population growth of just 2.3%. This study highlights the importance of considering multiple sources of disturbance when developing effective management plans, and the importance of employing predictive modeling and new technologies to assess human impact on marine mammals.

Table 11. Summary of studies evaluating long-term effects of anthropogenic activities (other than whale watching). Codes of study characteristics (based on the key components defined in Bejder & Samuels, 2003), source of anthropogenic disturbance, industry characteristics, and short- and long-term measures refer to factors listed in Table 4. Definitions of area use type (year-round habitat, breeding ground, feeding ground, migratory corridor, multiple use, etc.) followed Weinrich et al. (2010). Target species are listed with abbreviation of their scientific names and include bottlenose dolphins (*Tursiops truncatus*; Tt) and southern elephant seals (*Mirounga leonina*; Ml).

Study characteristics	Data collection length (y)	Location	Area use type	Source of anthropogenic disturbance	Whale-watching industry characteristics	Target species	Exposure measure	Short-term measures	Long-term measures	Detailed measured or modeled results	Reference
C O WE I	6	North Sea, Scotland	Year-round habitat	WW P M F BT PC PA	L HI WW G (whale watching present but not the focus of the study)	Tt	Exposure time (min/d)	N/A	N/A	High inter-individual spatial variability leading to variable exposure to vessel traffic with some individuals consistently under- or overexposed.	Pirotta et al., 2014
C O BDA I	6	North Sea, Scotland	Year-round habitat	WW P M F BT PA	L HI WW G (whale watching present but not the focus of the study)	Tt	N/A	BB M	H CS	High inter-individual differences in the motivational states that were consistent across years. Exposure and motivational states did not differ as a result of the modelled increase in boat traffic that is unlikely to cause a substantial behavioral disruption of calf survival traffic and dredging activity except during the operational phase.	Pirotta et al., 2015
M CE BDA G	N/A	North Sea, Scotland	Year-round habitat	WW P M F BT PA	L HI WW V (whale watching present but not the focus of the study)	Tt	# vessels	BB M G	H	Simulation did not detect long-term consequences of the increased disturbance (more than sixfold increase in vessel traffic).	New et al., 2013
C O BDA I	6	Pacific Ocean, Australia	Year-round habitat	BT F C	N/A	Ml	N/A	BB BC	H PD	Significant behavioral disruption (> 50% of an average foraging trip at sea) in one year resulted in a small (0.4%) decline in population size the following year. Persistent disruption (e.g., 30 y) caused by the long-term effects of climate change could result in a 0.3% decline in individual fitness and a 10% decline in population size.	New et al., 2014

Discussion

As we continue to study and understand aspects of the effects of noise exposure on marine mammal behavior, it has in some ways become more challenging to provide simple assessments of what kind of exposures will initiate responses and of what severity. Whereas initial assessments and regulatory approaches focused almost entirely on received noise levels (in simple sound pressure units) with proposed step-function thresholds for broad taxa, science is telling us there is much more nuance required. It was unrealistic to expect that diverse populations would all respond to various sounds at the same RL. Tyack & Thomas (2019) demonstrate that using an all-or-nothing threshold, ignoring the variation inherent even within one population responding to one signal, can underestimate effects by a factor of 280 for the dose-response function estimated by Miller et al. (2014). Responses of wildlife to sounds are probabilistic in nature both within and across individuals. Considerable variability in response type and magnitude has been observed for similar noise exposures as a function of species, age/sex class, individual behavioral state, and a host of interacting biological and ecological contextual factors (e.g., Richardson et al., 1990; Southall et al., 2007, 2019b; Ellison et al., 2012; NAS, 2017).

Effective management of the effects of sound on marine mammals requires criteria that recognize the inherent variability of how animals respond to different doses of sound in different situations. Methods have been developed to use results of behavioral response studies to estimate probabilistic functions relating to exposure and response. It is possible to estimate how the probability of an individual animal responding varies as a function of acoustic exposure, using any of several parameters for acoustic dosage and pooling data from any set of populations, settings, and sound types. However, pooling data from less sensitive species with those from highly sensitive ones will result in exposure-response functions that may fail to predict responses of the sensitive species and may fail management goals.

Documented variability in patterns of response to anthropogenic sounds raises questions about which taxa, contexts, and sound types should be pooled and where discrete probabilistic functions are needed to make informed decisions. The best method for analyzing specific issues, such as effects of seismic surveys on bowhead whales in shallow Arctic seas, would be the derivation of a discrete function with sufficient sample size for this particular setting. However, it may not be realistic to develop separate functions for each population, sound source, and setting. This raises the question of when it may be valid to extrapolate from one situation to another, or how to pool data when sample sizes are not large enough.

Southall et al. (2007) initially pooled taxa by hearing groups and had segregated exposures using purely acoustic pulse or non-pulse categories more relevant to auditory impact differences. Remarkable progress has been made in the past decade in developing methods to estimate probabilistic dose-response functions that can be used to predict the probability of response (Miller et al., 2014) at different severity levels (Harris et al., 2015). The Bayesian methods described in Miller et al. (2014) use reasonable *a priori* assumptions to efficiently derive strong statistical power from modest sample sizes. As the number of response studies increases, we have more options of using methods that empirically test which taxa, sounds, and contexts show similar enough dose-response patterns to indicate pooling them and which are so different that they are better treated separately (e.g., Harris et al., 2015). We advocate here for a rational, common-sense framework with which to systematically and objectively assess available science and yield a manageable number of probabilistic response functions with which to make informed decisions.

These kinds of exposure-response methods and this framework require the ability to integrate data from many separate studies with common measures. We strongly advocate for much more robust and systematic reporting of key exposure, contextual, and response metrics in both experimental and observational studies. Multiple and differential noise exposure metrics are clearly needed, including multiple SPL and SEL conditions as well as derived variables including SNR. This relates to all studies of marine mammal behavioral response, whether in the field or in the lab and whether acute or chronic exposure. While some of those identified in Table 1 may be more or less relevant than others, depending on the species and context of exposure, a common observation from Southall et al. (2007) that remains largely true here is that many studies simply provide insufficient details regarding these metrics to be as useful as they could be. To make matters worse, many studies certainly had such information collected but just not reported. These should be much more conspicuously and systematically reported in papers or, where not possible given space limitations, provided in supplementary materials of published papers (e.g., DeRuiter et al., 2013; Southall et al., 2019b).

In terms of assessing discrete (acute) exposure-response events, we made substantial progress and associated observations in several areas. First, in acknowledging the radically different contexts of discrete (acute) exposure-response events in captive and free-ranging conditions, we formalized the distinction between them and disentangled the severity scoring approach for them. The revised severity scale for captive responses further distinguishes between the different contexts of disruption of food-reinforced behavior and those observed in other conditions. These should logically be considered separately, which is of relevance for entities concerned with responses of animals in captive conditions. In terms of free-ranging animals, we propose a thorough reassessment and adaptation of the existing severity scale for discrete (acute) exposure-response events, taking a vital rate perspective with responses of progressive severity in different "tracks." For free-ranging marine mammals, we directly applied this severity scale to a subset of the marine mammal literature using a systematic assessment method based on the relevant details provided in candidate studies.

Herein, we made a very important distinction from how Southall et al. (2007) conducted a similar evaluation of the initial severity scale. Rather than segregating noise exposure into the "pulse" and "non-pulse" acoustic categories derived more for the purposes of evaluating auditory effects such as temporary threshold shifts in hearing, we segregated them into functional noise categories. By segregating noise sources into the effective industrial categories of active sonar, industrial (continuous), pile driving, and seismic airgun surveys, we were seeking to not only evaluate things that share some general contextual similarities at some level but that also are more likely to be regulated and hopefully mitigated using common practices. To be clear, a complete and thorough evaluation of all available studies in any of these industrial noise categories was well beyond the scope of this article. We did this deliberately to emphasize the point that such noise category segregation, including potential additional segregation for the purposes of deriving meta-analysis response functions (e.g., more sensitive and more tolerant species within sound types), would be a logical approach. The derivation of such group and noise type exposure-response probabilistic functions is an area of extensive and active research effort. This is also beyond the scope of this article. However, these could include, for example, Bayesian hierarchical models for generating differential forms of behavioral response curves (with uncertainty), model selection methods for assessing support for species pooling, and reversible-jump Markov chain and Monte Carlo methods to select those models best supported by the data.

A key observation and segregation of the severity assessment methods and results provided herein relates to differences in the temporal and spatial scales of exposure scenarios and associated studies. As discussed, we evaluated a number of important longer-term, population-level studies using the acute response severity scale with disappointing and/or potentially misleading or unfair conclusions. Put simply, these more conventional means of scoring response severity for known, discrete exposures at the individual or defined level are less applicable for studies conducted for less well-defined groups and/or local population levels and for repeated or chronic exposures.

Borrowing heavily from the literature associated with the study of chronic exposure in scenarios lasting years or decades (e.g., whale watching), we derived systematic review and assessment approaches. While these are perhaps more limited in yielding quantitative severity scores, they do provide a useful means of synthesizing results presented on variable (including long) temporal scales placed in a population context. From these assessments, the following general and specific conclusions arise. Longitudinal studies of marine mammals over the entire span of their home range are of particular importance in quantifying the ability of the targeted animal to compensate for disturbance and the associated impact(s). Animals living in depleted habitats or that rely on heterogeneously distributed resources will be less resilient, and the impact of disturbance on vital rates will be higher (Lusseau, 2014). Population-level consequences are mediated by individual responses (Lusseau, 2014), and herein we highlighted the importance of considering individual exposure rates (Christiansen & Lusseau, 2015; Pirotta et al., 2015).

For instance, the combined results of reviews of annual to decadal scales illustrate differences in impacts of whale watching between mysticetes and odontocetes, likely as a result of variabilities in their compensatory opportunities. It appears that mysticetes, although affected by vessel presence both physiologically (increases in stress hormones; see Rolland et al., 2012) and behaviorally (disruption of feeding activity; see Christiansen & Lusseau, 2014), may be more readily able to compensate for short-term effects that may not translate into long-term impact. Mysticetes are capital breeders; thus, they accumulate energy while in feeding grounds and transfer energy to calves while in breeding grounds. We might expect disturbance to have different consequences in these two areas and differences in compensative measures adopted by the animals. Migratory whales may be able to compensate for behavioral disruption in one part of their foraging grounds by feeding in other areas. A number of factors determine compensation ability in mysticetes, including individual cumulative exposure, presence and impact of concurrent disturbance sources, centrality and relative importance of study area within the home range of the animal, prey availability, and predictability (Lusseau, 2014). While these natural history factors may result in differential sensitivity to some kinds of disturbance, the threatened or endangered (in some species critically) status of many baleen whales and associated requisite conservative approaches to management are clearly relevant associated considerations.

Conversely, odontocetes, at least in some contexts, may experience a higher cumulative exposure to whale-watching disturbance due to their localized and coastal home range (Lusseau, 2014; Senigaglia et al., 2016). They also might be less resilient than mysticetes to overexploitation (Wade et al., 2012), requiring a different management approach (Williams et al., 2009a) because of their socioecology. Moreover, heterogeneity in spatial use over the animal home range can arise due to environmental factors such as uneven distribution of prey (Hastie et al., 2004), predation risk responses (Heithaus & Dill, 2006), or behavioral differences among individuals (de Stephanis et al., 2008; Senigaglia et al., 2012). These area usage differences may potentially lead to individuals' over- or under-exposure to whale-watching disturbance (Pirotta et al., 2015).

In summary, simple all-or-nothing thresholds that attempt to relate single noise exposure parameters (e.g., received noise level) and behavioral response across broad taxonomic grouping and sound types can lead to severe errors in predicting effects. Differences between species, individuals, exposure situational context, the temporal and spatial scales over which they occur, and the potential interacting effects of multiple stressors can lead to inherent variability in the probability and severity of behavioral responses. Studies that quantify this variability can estimate probabilistic response functions, which then provide much better accuracy for predicting effects and identifying potentially important contextual covariates. If one can pool data from different studies, methods are available to empirically decide which taxonomic, acoustic, or contextual factors so alter the dose-response function as to require estimating a separate function. Few behavioral response studies report the critical data in a systematic, structured, objective way. Such reporting is essential for these studies to achieve their full potential for improving management of the effects of sound on marine mammals.

Acknowledgments

We would like to acknowledge all members from the Southall et al. (2007) panel who participated in the process leading to this article, including Brandon Southall, Ann Bowles, William Ellison, James Finneran, Roger Gentry, Charles Greene, Jr., Darlene Ketten, James Miller, Paul Nachtigall, and Peter Tyack. Colleen Reichmuth, Doug Nowacek, and Lars Bejder were subsequently added to the panel. Each of these individuals participated intellectually in the current effort, with a subset contributing as coauthors to this article. P. Tyack acknowledges support for his time from the Strategic Environmental Research and Development Program Contract #RC21-3091 and Office of Naval Research Grant #N000142112096. L. Bejder's participation was made possible via University of Hawai'i's Institute of Marine Biology and School of Ocean and Earth Science and Technology Contribution Numbers 1862 and 11378, respectively. We would also like to acknowledge support from the Joint Industry Programme for Sound and Marine Life for panelists' travel and logistical support and for their support of open access for this publication.

Literature Cited

Bee, M. A., & Micheyl, C. (2008). The "Cocktail Party Problem": What is it? How can it be solved? And why should animal behaviorists study it? *Journal of Comparative Psychology*, *122*(3), 235-251. https://doi.org/10.1037/0735-7036.122.3.235

Bejder, L., & Samuels, A. (2003). Evaluating the effects of nature-based tourism on cetaceans. *Marine Mammals: Fisheries, Tourism and Management Issues*, *1*, 229-256.

Bejder, L., Samuels, A., Whitehead, H., Gales, N., Mann, J., Connor, R., Heithaus, M., Watson-Capps, J., Flaherty, C., & Krützen, M. (2006). Decline in relative abundance of bottlenose dolphins exposed to long-term disturbance. *Conservation Biology*, *20*(6), 1791-1798. https://doi.org/10.1111/j.1523-1739.2006.00540.x

Bertulli, C. G., Leeney, R. H., Barreau, T., & Matassa, D. S. (2016). Can whale-watching and whaling co-exist? Tourist perceptions in Iceland. *Journal of the Marine Biological Association of the United Kingdom*, *96*(4), 969-977. https://doi.org/10.1017/S002531541400006X

Blackwell, S. B., Lawson, J. W., & Williams, M. T. (2004). Tolerance by ringed seals (*Phoca hispida*) to impact pipe-driving and construction sounds at an oil production island. *The Journal of the Acoustical Society of America*, *115*(5), 2346-2357. https://doi.org/10.1121/1.1701899

Blickley, J. L., Blackwood, D., & Patricelli, G. L. (2012). Experimental evidence for the effects of chronic anthropogenic noise on abundance of greater sage-grouse at leks. *Conservation Biology*, *26*(3), 461-471. https://doi.org/10.1111/j.1523-1739.2012.01840.x

Booth, C. G., Sinclair, R. R., & Harwood, J. (2020). Methods for monitoring for the population consequences of disturbance in marine mammals: A review. *Frontiers in Marine Science*, *7*, 115. https://doi.org/10.3389/fmars.2020.00115

Bowles, A. E., & Anderson, R. C. (2012). Behavioral responses and habituation of small cetaceans and pinnipeds to novel objects and simulated fishing gear with and without a pinger. *Aquatic Mammals*, *38*(2), 161-188. https://doi.org/10.1578/AM.38.2.2012.161

Brandt, M. J., Diederichs, A., & Nehls, G. (2009). *Harbour porpoise responses to pile driving at the Horns Rev II offshore wind farm in the Danish North Sea*. Final Report to DONG Energy.

Brandt, M. J., Diederichs, A., Betke, K., & Nehls, G. (2011). Responses of harbour porpoises to pile driving at the Horns Rev II offshore wind farm in the Danish North Sea. *Marine Ecology Progress Series*, *421*, 205-216. https://doi.org/10.3354/meps08888

Christiansen, F., & Lusseau, D. (2014). Understanding the ecological effects of whale-watching on cetaceans. In J. Higham, L. Bejder, & R. Williams (Eds.), *Whale-watching: Sustainable tourism and ecological management* (pp. 177-192). Cambridge University Press. https://doi.org/10.1017/CBO9781139018166.016

Christiansen, F., & Lusseau, D. (2015). Linking behavior to vital rates to measure the effects of non-lethal disturbance on wildlife. *Conservation Letters*, *8*(6), 424-431. https://doi.org/10.1111/conl.12166

Christiansen, F., Dawson, S. M., Durban, J. W., Fearnbach, H., Miller, C. A., Bejder, L., Uhart, M., Sironi, M., Corkeron, P., Rayment, W., Leunissen, E., Haria, E., Ward, R., Warick, H. A., Kerr, I., Lynn, M. S., Pettis, H. M., & Moore, M. J. (2020). Population comparison of right whale body condition reveals poor state of the North Atlantic right whale. *Marine Ecology Progress Series*, *640*, 1-16. https://doi.org/10.3354/meps13299

Currie, J. J., van Aswegen, M., Stack, S. H., West, K. L., Vivier, F., & Bejder, L. (2021). Rapid weight loss in free ranging pygmy killer whales (*Feresa attenuata*) and the implications for anthropogenic disturbance of odontocetes. *Scientific Reports*, *11*, 8181. https://doi.org/10.1038/s41598-021-87514-2

DeRuiter, S. L., Southall, B. L., Calambokidis, J., Zimmer, W. M. X., Sadykova, D., Falcone, E. A., Friedlaender, A. S., Joseph, J. E., Moretti, D., Schorr, G. S., Thomas, L., & Tyack, P. L. (2013). First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active sonar. *Biology Letters*, *9*, 20130223. https://doi.org/10.1098/rsbl.2013.0223

de Stephanis, R., García-Tíscar, S., Verborgh, P., Esteban-Pavo, R., Pérez, S., Minvielle-Sebastia, L., & Guinet, C. (2008). Diet of the social groups of long-finned pilot whales (*Globicephala melas*) in the Strait of Gibraltar. *Marine Biology*, *154*(4), 603-612.

Dunlop, R. A., Noad, M. J., McCauley, R. D., Kniest, E., Paton, D., & Cato, D. H. (2015). The behavioural response of humpback whales (*Megaptera novaeangliae*) to a 20 cubic inch air gun. *Aquatic Mammals*, *41*(4), 412-433. https://doi.org/10.1578/AM.41.4.2015.412

Dunlop, R. A., Noad, M. J., McCauley, R. D., Scott-Hayward, L., Kniest, E., Slade, R., Paton, D., & Cato, D. H. (2017). Determining the behavioural dose–response relationship of marine mammals to air gun noise and source proximity. *Journal of Experimental Biology*, *220*(16), 2878-2886. https://doi.org/10.1242/jeb.160192

Ellison, W. T., Southall, B. L., Clark, C. W., & Frankel, A. S. (2012). A new context-based approach to assess marine mammal behavioral responses to anthropogenic sounds. *Conservation Biology*, *26*(1), 21-28. https://doi.org/10.1111/j.1523-1739.2011.01803.x

Filby, N. E., Stockin, K. A., & Scarpaci, C. (2014). Long-term responses of Burrunan dolphins (*Tursiops australis*) to swim-with dolphin tourism in Port Phillip Bay, Victoria, Australia: A population at risk. *Global Ecology and Conservation*, *2*, 62-71. https://doi.org/10.1016/j.gecco.2014.08.006

Gailey, G., Sychenko, O., McDonald, T., Racca, R., Rutenko, A., & Bröker, K. (2016). Behavioural responses of western gray whales to a 4-D seismic survey off northeastern Sakhalin Island, Russia. *Endangered Species Research*, *30*, 53-71. https://doi.org/10.3354/esr00713

Gomez, C., Lawson, J. W., Wright, A. J., Buren, A. D., Tollit, D., & Lesage, V. (2016). A systematic review on the behavioural responses of wild marine mammals to noise: The disparity between science and policy. *Canadian Journal of Zoology*, *94*, 801-819. https://doi.org/10.1139/cjz-2016-0098

Gordon, J. C. D., Leaper, R., Hartley, F. G., & Chappell, O. (1992). *Effects of whale-watching vessels on the surface and underwater acoustic behaviour of sperm whales off Kaikoura, New Zealand* (Science and Research Series). New Zealand Department of Conservation. 64 pp.

Götz, T., & Janik, V. M. (2011). Repeated elicitation of the acoustic startle reflex leads to sensitisation in subsequent avoidance behaviour and induces fear conditioning. *BMC Neuroscience*, *12*(1), 1-13. https://doi.org/10.1186/1471-2202-12-30

Harris, C. M., Sadykova, D., DeRuiter, S. L., Tyack, P. L., Miller, P. J. O., Kvadsheim, P. H., Lam, F-P. A., & Thomas, L. (2015). Dose response severity functions for acoustic disturbance in cetaceans using recurrent event survival analysis. *Ecosphere*, *6*(11), 1-14. https://doi.org/10.1890/ES15-00242.1

Harris, C. M., Thomas, L., Sadykova, D., DeRuiter, S. L., Tyack, P. L., Southall, B. L., Read, A. J., & Miller, P. J. O. (2016). The challenges of analyzing behavioral response study data: An overview of the MOCHA (Multi-study Ocean acoustics Human effects Analysis) Project. In A. N. Popper & A. Hawkins (Eds.), *Advances in experimental medicine and biology: The effects of noise on aquatic life II* (Vol. 875, pp. 399-407). Springer. https://doi.org/10.1007/978-1-4939-2981-8_47

Hastie, G. D., Donovan, C., Götz, T., & Janik, V. M. (2014). Behavioral responses by grey seals (*Halichoerus grypus*) to high frequency sonar. *Marine Pollution Bulletin*, *79*(1-2), 205-210. https://doi.org/10.1016/j.marpolbul.2013.12.013

Hastie, G. D., Wilson, B., Wilson, L. J., Parsons, K. M., & Thompson, P. M. (2004). Functional mechanisms underlying cetacean distribution patterns: Hotspots for bottlenose dolphins are linked to foraging. *Marine Biology*, *144*(2), 397-403. https://doi.org/10.1007/s00227-003-1195-4

Heithaus, M. R., & Dill, L. M. (2006). Does tiger shark predation risk influence foraging habitat use by bottlenose dolphins at multiple spatial scales? *OIKOS*, *114*(2), 257-264. https://doi.org/10.1111/j.2006.0030-1299.14443.x

Himes Boor, G. K. (2014). A framework for developing objective and measurable recovery criteria for threatened and endangered species. *Conservation Biology*, *28*(1), 33-43. https://doi.org/10.1111/cobi.12155

Holt, M. M., Noren, D. P., Veirs, V., Emmons, C. K., & Veirs, S. (2009). Speaking up: Killer whales (*Orcinus orca*) increase their call amplitude in response to vessel noise. *The Journal of the Acoustical Society of America*, *125*(1), EL27-EL32. https://doi.org/10.1121/1.3040028

Isojunno, S., Curé, C., Kvadsheim, P. H., Lam, F-P. A., Tyack, P. L., Wensveen, P. J., & Miller, P. J. O. (2016). Sperm whales reduce foraging effort during exposure to 1-2 kHz sonar and killer whale sounds. *Ecological Applications*, *26*(1), 77-93. https://doi.org/10.1890/15-0040

Kendall, L. S., Sirovic, A., & Roth, E. H. (2013). Effects of construction noise on the Cook Inlet beluga whale (*Delphinapterus leucas*) vocal behavior. *Canadian Acoustics*, *41*(3), 3-13.

Lacy, R. C., Williams, R., Ashe, E., Balcomb III, K. C., Brent, L. J. N., Clark, C. W., Croft, D. P., Giles, D. A., MacDuffee, M., & Paquet, P. C. (2017). Evaluating anthropogenic threats to endangered killer whales to inform effective recovery plans. *Scientific Reports*, *7*(1), 14119. https://doi.org/10.1038/s41598-017-14471-0

Lusseau, D. (2004). The hidden cost of tourism: Detecting long-term effects of tourism using behavioral information. *Ecology and Society*, *9*(1), 2. https://doi.org/10.5751/ES-00614-090102

Lusseau, D. (2005). Residency pattern of bottlenose dolphins *Tursiops* spp. in Milford Sound, New Zealand, is related to boat traffic. *Marine Ecology Progress Series*, *295*, 265-272. https://doi.org/10.3354/meps295265

Lusseau, D. (2014). Ecological constraints and the propensity for population consequences of whale-watching disturbances. In J. Higham, L. Bejder, & R. Williams (Eds.), *Whale-watching: Sustainable tourism and ecological management* (pp. 229-242). Cambridge University Press. https://doi.org/10.1017/CBO9781139018166.019

Lusseau, D., Slooten, L., & Currey, R. J. (2006). Unsustainable dolphin-watching tourism in Fiordland, New Zealand. *Tourism in Marine Environments*, *3*(2), 173-178. https://doi.org/10.3727/154427306779435184

Madsen, P. T. (2005). Marine mammals and noise: Problems with root mean square sound pressure levels for transients. *The Journal of the Acoustical Society of America*, *117*(6), 3952-3957. https://doi.org/10.1121/1.1921508

Madsen, P. T., Møhl, B., Nielsen, B. K., & Wahlberg, M. (2002). Male sperm whale behavior during exposures to distant seismic survey pulses. *Aquatic Mammals*, *28*(3), 231-240.

Malme, C. I., Würsig, B., Bird, J. E., & Tyack, P. L. (1986). *Behavioral responses of gray whales to industrial noise: Feeding observations and predictive modeling* (Report 6265; OCS Study MMS 88-0048; NTIS PB88-249008). BBN.

Martin, S. (2012). *Whale watching in Iceland: An assessment of whale watching activities on Skjalfandi Bay* (Master's thesis). University of the Westfjords, Ísafjörður, Iceland.

McHuron, E. A., Costa, D. P., Schwarz, L., & Mangel, M. (2017). State-dependent behavioural theory for assessing the fitness consequences of anthropogenic disturbance on capital and income breeders. *Methods in Ecology and Evolution*, *8*(5), 552-560. https://doi.org/10.1111/2041-210X.12701

McKenna, M. F., Shannon, G., & Fristrup, K. (2016). Characterizing anthropogenic noise to improve understanding and management of impacts to wildlife. *Endangered Species Research*, *31*, 279-291. https://doi.org/10.3354/esr00760

Miller, P. J. O., Johnson, M. P., Madsen, P. T., Biassoni, N., Quero, M., & Tyack, P. L. (2009). Using at-sea experiments to study the effects of airguns on the foraging behavior of sperm whales in the Gulf of Mexico. *Deep-Sea Research*, *56*, 1168-1181. https://doi.org/10.1016/j.dsr.2009.02.008

Miller, P. J. O., Kvadsheim, P. H., Lam, F-P. A., Wensveen, P. J., Antunes, R., Alves, A. C., Visser, F., Kleivane, L., Tyack, P., & Sivle, L. D. (2012). The severity of behavioral changes observed during experimental exposures of killer (*Orcinus orca*), long-finned pilot (*Globicephala melas*), and sperm (*Physeter macrocephalus*) whales to naval sonar. *Aquatic Mammals*, *38*(4), 362-401. https://doi.org/10.1578/AM.38.4.2012.362

Miller, P. J. O., Antunes, R. N., Wensveen, P. J., Samarra, F. I. P., Alves, A. C., Tyack, P. L., Kvadsheim, P. H., Kleivane, L., Lam, F-P. A., Ainslie, M. A., & Thomas, L. (2014). Dose-response relationships for the onset of avoidance of sonar by free-ranging killer whales. *The Journal of the Acoustical Society of America*, *135*(2), 975-993. https://doi.org/10.1121/1.4861346

Moretti, D., Thomas, L., Marques, T., Harwood, J., Dilley, A., Neales, B., Shaffer, J., McCarthy, E., New, L., Jarvis, S., & Morrissey, R. (2014). A risk function for behavioral disruption of Blainville's beaked whales (*Mesoplodon densirostris*) from mid-frequency active sonar. *PLOS ONE*, *9*, e85064. https://doi.org/10.1371/journal.pone.0085064

National Academies of Sciences, Medicine, and Engineering (NAS). (2017). *Approaches to understanding the cumulative effects of stressors on marine mammals*. The National Academies Press.

National Research Council (NRC). (2000). *Marine mammals and low-frequency sound*. The National Academies Press.

New, L. F., Harwood, J., Thomas, L., Donovan, C., Clark, J. S., Hastie, G., Thompson, P. M., Cheney, B., Scott-Hayward, L., & Lusseau, D. (2013). Modelling the biological significance of behavioural change in coastal bottlenose dolphins in response to disturbance. *Functional Ecology*, *27*(2), 314-322. https://doi.org/10.1111/1365-2435.12052

New, L. F., Clark, J. S., Costa, D. P., Fleishman, E., Hindell, M. A., Klanjšček, T., Lusseau, D., Kraus, S., McMahon, C. R., Robinson, P. W., Schick, R. S., Schwarz, L. K., Simmons, S. E., Thomas, L., Tyack, P., & Harwood, J. (2014). Using short-term measures of behavior to estimate long-term fitness of southern elephant seals.

Marine Ecology Progress Series, *496*, 99-108. https://doi.org/10.3354/meps10547

Nowacek, D. P., Johnson, M. P., & Tyack, P. L. (2004). North Atlantic right whales (*Eubalaena glacialis*) ignore ships but respond to alerting stimuli. *Proceedings of the Royal Society B: Biological Sciences*, *271*(1536), 227-231. https://doi.org/10.1098/rspb.2003.2570

Nowacek, D. P., Thorne, L. H., Johnson, D. W., & Tyack, P. L. (2007). Responses of cetaceans to anthropogenic noise. *Mammal Review*, *37*(2), 81-115. https://doi.org/10.1111/j.1365-2907.2007.00104.x

Occupational Safety and Health Administration (OSHA). (1999). *OSHA technical manual* (Section VII, Chapter 1; OSHA Technical Directive TED 01-00-015 [TED 1-0.15A]). U.S. Department of Labor.

Payne, R., & Webb, D. (1971). Orientation by means of long range acoustic signaling in baleen whales. *Annals of the New York Academy of Sciences*, *188*(1), 110-141. https://doi.org/10.1111/j.1749-6632.1971.tb13093.x

Pirotta, E., Thompson, P. M., Cheney, B., Donovan, C. R., & Lusseau, D. (2014). Estimating spatial, temporal and individual variability in dolphin cumulative exposure to boat traffic using spatially explicit capture–recapture methods. *Animal Conservation*, *18*(1), 20-31. https://doi.org/10.1111/acv.12132

Pirotta, E., Harwood, J., Thompson, P. M., New, L., Cheney, B., Arso, M., Hammond, P. S., Donovan, C., & Lusseau, D. (2015). Predicting the effects of human developments on individual dolphins to understand potential long-term population consequences. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1818), 20152109. https://doi.org/10.1098/rspb.2015.2109

Pirotta, E., Booth, C. G., Cade, D. E., Calambokidis, J., Costa, D. P., Fahlbusch, J. A., Friedlaender, A. S., Goldbogen, J. A., Harwood, J., Hazen, E. L., New, L., & Southall, B. L. (2021). Context-dependent variability in the predicted daily energetic costs of disturbance in blue whales. *Conservation Physiology*, *9*(1). https://doi.org/10.1093/conphys/coaa137

Pirotta, E., Mangel, M., Costa, D. P., Goldbogen, J. A., Harwood, J., Hin, V., Irvine, L. M., Mate, B. R., McHuron, E. A., Palacios, D. M., Schwarts, L. K., & New, L. (2019). Anthropogenic disturbance in a changing environment: Modelling lifetime reproductive success to predict the consequences of multiple stressors on a migratory population. *OIKOS*, *128*, 1340-1357. https://doi.org/10.1111/oik.06146

Pirotta, E., Booth, C. G., Costa, D. P., Fleishman, E., Kraus, S. D., Lusseau, D., Moretti, M., New, L. F., Schick, R. S., Schwarz, L. K., Simmons, S. E., Thomas, L., Tyack, P. L., Weise, M. J., Wells, R. S., & Harwood, J. (2018). Understanding the population consequences of disturbance. *Ecology and Evolution*, *8*(19), 9934-9946. https://doi.org/10.1002/ece3.4458

Richardson, W. J., & Würsig, B. (1997). Influences of man-made noise and other human actions on cetacean behaviour. *Marine & Freshwater Behaviour & Physiology*, *29*(1-4), 183-209. https://doi.org/10.1080/10236249709379006

Richardson, W. J., Würsig, B., & Greene, C. R., Jr. (1990). Reactions of bowhead whales, *Balaena mysticetus*, to drilling and dredging noise in the Canadian Beaufort Sea. *Marine Environmental Research*, *29*, 135-160.

Rolland, R. M., Parks, S. E., Hunt, K. E., Castellote, M., Corkeron, P. J., Nowacek, D. P., Wasser, S. K., & Kraus, S. D. (2012). Evidence that ship noise increases stress in right whales. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1737), 2363-2368. https://doi.org/10.1098/rspb.2011.2429

Schmidt, B., Kirpalani, H., Rosenbaum, P., & Cadman, D. (1988). Strengths and limitations of the Apgar score: A critical appraisal. *Journal of Clinical Epidemiology*, *41*(9), 843-850. https://doi.org/10.1016/0895-4356(88)90100-X

Senigaglia, V. (2020). *Ecological consequences and social drivers of human–wildlife interactions: The case of food-provisioning of bottlenose dolphins in Bunbury, Western Australia* (Ph.D. thesis). Murdoch University, Perth, Australia.

Senigaglia, V., de Stephanis, R., Verborgh, P., & Lusseau, D. (2012). The role of synchronized swimming as affiliative and anti-predatory behavior in long-finned pilot whales. *Behavioural Processes*, *91*(1), 8-14. https://doi.org/10.1016/j.beproc.2012.04.011

Senigaglia, V., Christiansen, F., Sprogis, K. R., Symons, J., & Bejder, L. (2019). Food-provisioning negatively affects calf survival and female reproductive success in bottlenose dolphins. *Scientific Reports*, *9*, 8981. https://doi.org/10.1038/s41598-019-45395-6

Senigaglia, V., Christiansen, F., Bejder, L., Gendron, D., Lundquist, D., Noren, D. P., Schaffar, A., Smith, J. C., Williams, R., Martinez, E., Lusseau, D., & Stockin, K. (2016). Meta-analyses of whale-watching impact studies: Comparisons of cetacean responses to disturbance. *Marine Ecology Progress Series*, *542*, 251-263. https://doi.org/10.3354/meps11497

Simmonds, M. P., Dolman, S. J., Jasny, M., Parsons, E. C. M., Weilgart, L., Wright, A. J., & Leaper, R. (2014). Marine noise pollution-increasing recognition but need for more practical action. *Journal of Ocean Technology*, *9*(1), 71-90.

Southall, B. L. (2017). Noise. In B. Würsig, J. G. M. Thewissen, & K. Kovacs (Eds.), *Encyclopedia of marine mammals* (3rd ed., pp. 699-707). Academic Press.

Southall, B. L. (2021). Evolutions in marine mammal noise exposure criteria. *Acoustics Today*, *17*(2), 52-60. https://acousticstoday.org/wp-content/uploads/2021/06/2021summerEvolutions-in-Marine-Mammal-Noise-Exposure-Criteria-Brandon-L.-Southall.pdf

Southall, B. L., Nowacek, D. P., Miller, P. J. O., & Tyack, P. L. (2016). Experimental field studies to measure behavioral responses of cetaceans to sonar. *Endangered Species Research*, *31*, 293-315. https://doi.org/10.3354/esr00764

Southall, B. L., Finneran, J. J., Reichmuth, C., Nachtigall, P. E., Ketten, D. R., Bowles, A. E., Ellison, W. T., Nowacek, D. P., & Tyack, P. L. (2019a). Marine mammal noise exposure criteria: Updated scientific recommendations for residual hearing effects. *Aquatic Mammals*, *45*(2), 125-232. https://doi.org/10.1578/AM.45.2.2019.125

Southall, B. L., Bowles, A. E., Ellison, W. T., Finneran, J. J., Gentry, R. L., Greene, C. R., Kastak, D., Ketten, D. R., Miller, J. H., Nachtigall, P. E., Richardson, W. J., Thomas, J. A., & Tyack, P. L. (2007). Marine mammal noise exposure criteria. *Aquatic Mammals*, *33*(4), 411-522. https://doi.org/10.1578/AM.33.4.2007.411

Southall, B. L., DeRuiter, S. L., Friedlaender, A., Stimpert, A. K., Goldbogen, J. A., Hazen, E., Casey, C., Fregosi, S., Cade, D. E., Allen, A. N., Harris, C. M., Schorr, G., Moretti, D., Guan, S., & Calambokidis, J. (2019b). Behavioral responses of individual blue whales (*Balaenoptera musculus*) to mid-frequency military sonar. *Journal of Experimental Biology*, *222*(5), 1-15. https://doi.org/10.1242/jeb.190637

Strindberg, S., & O'Brien, T. (2012). *A decision tree for monitoring wildlife to assess the effectiveness of conservation interventions*. Wildlife Conservation Society.

Thompson, P. M., Lusseau, D., Barton, T., Simmons, D., Rusin, J., & Bailey, H. (2010). Assessing the responses of coastal cetaceans to the construction of offshore wind turbines. *Marine Pollution Bulletin*, *60*(8), 1200-1208. https://doi.org/10.1016/j.marpolbul.2010.03.030

Tyack, P. L., & Thomas, L. (2019). Using dose–response functions to improve calculations of the impact of anthropogenic noise. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *29*, 242-253. https://doi.org/10.1002/aqc.3149

Tyack, P. L., Zimmer, W. M. X., Moretti, D., Southall, B. L., Claridge, D. E., Durban, J. W., Clark, C. W., D'Amico, A., DiMarzio, N., Jarvis, S., McCarthy, E., Morrissey, R., Ward, J., & Boyd, I. L. (2011). Beaked whales respond to simulated and actual navy sonar. *PLOS ONE*, *6*(3), e17009. https://doi.org/10.1371/journal.pone.0017009

Tyne, J., Loneragan, N., & Bejder, L. (2014). The use of area–time closures as a tool to manage cetacean-watch tourism. In J. Higham, L. Bejder, & R. Williams (Eds.), *Whale-watching: Sustainable tourism and ecological management* (pp. 242-260). Cambridge University Press. https://doi.org/10.1017/CBO9781139018166.020

U.S. Department of the Navy. (2017). *Criteria and thresholds for U.S. Navy acoustic and explosive effects analysis (Phase III)* (Technical Report). SSC Pacific, San Diego. 194 pp.

Wade, P. R., Reeves, R. R., & Mesnick, S. L. (2012). Social and behavioural factors in cetacean responses to overexploitation: Are odontocetes less "resilient" than mysticetes? *Journal of Marine Biology*, *2012*, 567276. https://doi.org/10.1155/2012/567276

Weinrich, M., & Corbelli, C. (2009). Does whale watching in southern New England impact humpback whale (*Megaptera novaeangliae*) calf production or calf survival? *Biological Conservation*, *142*(12), 2931-2940. https://doi.org/10.1016/j.biocon.2009.07.018

Weinrich, M., Pekarcik, C., & Tackaberry, J. (2010). The effectiveness of dedicated observers in reducing risks of marine mammal collisions with ferries: A test of the technique. *Marine Mammal Science*, *26*(2), 460-470. https://doi.org/10.1111/j.1748-7692.2009.00343.x

Wensveen, P. J., Isojunno, S., Hansen, R. R., von Benda-Beckmann, A. M., Kleivane, L., van IJsselmuide, S., Lam, F-P. A., Kvadsheim, P. H., DeRuiter, S. L., Curé, C., Narazaki, T., Tyack, P. L., & Miller, P. J. O. (2019). Northern bottlenose whales in a pristine environment respond strongly to close and distant navy sonar signals. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1899), 20182592. https://doi.org/10.1098/rspb.2018.2592

Williams, R., Lusseau, D., & Hammond, P. S. (2006). Estimating relative energetic costs of human disturbance to killer whales (*Orcinus orca*). *Biological Conservation*, *133*(3), 301-311. https://doi.org/10.1016/j.biocon.2006.06.010

Williams, R., Lusseau, D., & Hammond, P. S. (2009a). The role of social aggregations and protected areas in killer whale conservation: The mixed blessing of critical habitat. *Biological Conservation*, *142*(4), 709-719. https://doi.org/10.1016/j.biocon.2008.12.004

Williams, R., Trites, A. W., & Bain, D. E. (2002). Behavioural responses of killer whales (*Orcinus orca*) to whale-watching boats: Opportunistic observations and experimental approaches. *Journal of Zoology*, *256*(2), 255-270. https://doi.org/10.1017/S0952836902000298

Williams, R., Bain, D. E., Smith, J. C., & Lusseau, D. (2009b). Effects of vessels on behaviour patterns of individual southern resident killer whales *Orcinus orca*. *Endangered Species Research*, *6*(3), 199-209. https://doi.org/10.3354/esr00150

Williams, R., Krkošek, M., Ashe, E., Branch, T. A., Clark, S., Hammond, P. S., Hoyt, E., Noren, D. P., Rosen, D., & Winship, A. (2011). Competing conservation objectives for predators and prey: Estimating killer whale prey requirements for Chinook salmon. *PLOS ONE*, *6*(11), e26738. https://doi.org/10.1371/journal.pone.0026738

Appendix 1. Description of Subject-Specific, Contextual, and Noise Exposure Metrics

The following gives more comprehensive explanations of recommended metrics for disturbance studies listed in Table 1A, B, and C.

Table 1A. Subject-Specific (Individual or Group) Variables

Species – Species should be given, including metadata regarding the identification method and reference resources. For example, we have used Society for Committee on Taxonomy (2021) and Southall et al. (2019), Appendices 1-6, in our assessments of the literature. Any possible ambiguities in identification should be made explicit in the metadata for the study.

Functional Hearing Group – The reference resource on functional hearing groups should be specified. We have used Southall et al. (2019) here, but it is possible that other references will become relevant as additional or alternate hearing groups are refined or developed in future criteria.

Subject Individual Identifier (Where Applicable) – The subjects of observations may be defined as groups or individuals. As discussed in the article, separate processes may be required for long-term studies in which case the group may be a local population. If the subject is an individual, the code for that individual should be specified. In metadata for the observation, the time-scale of the identifier should be specified clearly (e.g., within track, during trial, during trial series, over the period that an instrument is attached, or long-term using photographic or other permanent identification). If there are reasons to suspect that different individuals were observed in successive time blocks (e.g., because subjects were migrants), this should also be specified.

Subject Weighting – If subjects are exposed to multiple stimuli or to the same stimulus multiple times, individual responses should be weighted by $1/n$ to ensure that analyses are not biased due to imbalances in sample size where n is the number of stimuli or stimulus presentations. For instance, if individual x is exposed three times, there will be three severity scores (s_x1, s_x2, s_x3) for that individual, with either different or identical severity scores. Each would be given a subject weight of $1/n = 0.33$ such that the overall contribution of that individual to an across-individual assessment of response probability would be $N = 1$, with a severity score of S_x as follows:

$$S_x = \left(s_x 1 * \frac{1}{n}\right) + \left(s_x 2 * \frac{1}{n}\right) + \left(s_x 3 * \frac{1}{n}\right)$$

Censored Data? – It is possible that subject responses are observed even at the lowest exposure level in a series of planned experiments, indicating that the threshold for response is lower than the minimum of the test range (left-censored). Alternatively, animals may not respond at even the highest level of exposure (right-censored). The range of exposure levels should be specified and censoring indicated.

Age Class (If Known) – The age class of a subject should be specified, including the resource or method used to determine age class.

Sex (If Known) – The sex of the subject should be specified, including the resource or method used to determine sex.

Calf Present? (If Female) – The presence of calves should be encoded. This metric is known to have an effect on the type and probability of responses, but it can be somewhat circular if the sex of the animal accompanying a calf has been determined using its presence. The resource or method for sexing the subject should be specified. For some species, particularly highly social small whales, examples of possible sources of confusion are alloparental caregivers (Augusto et al., 2017) and subadult siblings of both sexes.

Group Size – The size of the group should be specified, and the strategy for calculating the value should be included in metadata. For example, group size might be the average or maximum count from a number of surfacings.

Physical State – When observing free-ranging animals encountered for brief periods, information about physical state may or may not be available. Some proxy metrics, such as sex or the presence of a calf, are included herein. However, other physical states, such as body condition (e.g., illness or starvation), pregnancy or other reproductive state, and hearing impairment, have the potential to affect probability of responses but must be treated as unknown sources of variation.

Because such metrics will be available so rarely, physical condition has not been included as a separate metric here. If study subjects are known individuals in captivity or observed as part of a long-term program, this information should be specified to the extent that it is available.

Behavioral State – Behavioral states are activities, such as feeding and traveling, that an animal engages in over time. States represent functional categories that can be associated with biologically important factors such as energy budgets, survival, or fitness. States are often comprised of a series of shorter-term behavioral events such as fluking up, diving, pursuit, capturing prey, eating, etc. Because these states are not always clearly separable based on a limited source of behavioral data (e.g., surface-based observations, an acoustic tag), an ethogram defining the strategy for separating observed behaviors into state categories must be specified regardless of the degree of detail available about individual behavioral events. Without this information, meta-analyses across studies may be difficult to conduct, particularly studies with different methodologies (e.g., different tracking methods) or performed by different research teams. In this article, we have emphasized states that are measurable at sea and can be related to survival and fitness, particularly foraging and reproductive behaviors. A small number of states have proved quantifiable and potentially variable in controlled exposure experiments (CEEs) across multiple studies. These include deep and shallow feeding, travel with an estimate of speed, social interactions, and calling.

Table 1B. Exposure Context Variables

As described in the text of the publication, context can have a large effect on the intensity of marine mammal responses to noise. Contextual features related to the noise environment are described in Table 1B.

Exposure Novelty (i.e., Is the Source Common/ Rare in the Study Area) – For the most part, the history of exposure of an individual animal or group with a given stimulus is unknown for free-ranging animals. If previous exposure is known (e.g., instrumented animal or captive conditions), it should be specified. If unknown, a measure of the rate of encounter with the stimulus locally can be used as a proxy.

Exposure Similar to Predator Sounds? – Human-made noise (e.g., sonars) can share features with sounds made by predators or other hazards. In addition, test stimuli can be deliberately designed using such sounds, particularly sounds of or similar to those produced by predators. What constitutes similarity still involves many unknowns, but some features, such as frequency sweeps within the frequency range of small whale calls, have aroused responses that cannot be explained adequately by sound amplitude such as avoidance or calling in response to other vocalizations. The best way to make sure this information is available for future meta-analysis is to provide soundfiles and spectrograms with metadata.

Other Species Present in the Area? – Responses to disturbance may be shaped by the presence of other species. Predators represent a special case (addressed below), but other kinds of interspecies interactions may be important as well. For example, they may be competitors or allies in maintaining vigilance, or they may make social interactions more complex.

Predator Species Present in the Area? – Predator presence has the potential to influence disturbance responses in both the terrestrial and marine environments (Frid & Dill, 2002; Rankin et al., 2013; Harris et al., 2018), but the relationship can be complex (Heithaus & Frid, 2003). Predator presence should be documented to the degree possible and included in disturbance datasets, along with metadata on the measures used to survey predator presence.

Other Anthropogenic Presence/Noise in the Area? (Type and Proximity) – In addition to providing information about encounter rate as a proxy for experience with a disturbance, anthropogenic disturbances other than a target/test source can have direct and immediate effects on responses. Properties related to the sound field are described below (e.g., masking noise). However, the nature of the anthropogenic source can also be important (e.g., approaching vessel traffic). This information should be included in study datasets, along with metadata on the measures used to survey for predators.

Table 1C. Noise Exposure Metrics

What follows is a brief explanation of the noise exposure metrics recommended as a minimum for any dataset associated with a disturbance study (listed in Table 1C). The acoustic metrics are defined in more detail in Southall et al. (2007), Appendix A.

Continuous or Intermittent Exposure – Continuous noise differs from intermittent noise in important ways. Continuous noise is the most efficient masking source likely to be encountered by animals because there are no windows in the noise that would allow a target sound to be heard (e.g., Lee et al., 2017). In the case of continuous noise, simple presence/absence is enough specification. For intermittent exposure, however, rise time (see below) and degree of intermittency may be important factors as well (the most basic measure is duty cycle—the proportion of time that it is present).

Interval Between Exposures (s) – The interval between exposures can be an important factor and should be specified. If a series of trials must be conducted in a short period, animals may or may not recover from one exposure before the next begins. Alternatively, if the interval is fairly long, animals may be in different behavioral states when reacting. This metric differs from intermittency in that it applies to successive exposures rather than being a property of individual exposures, although short trial intervals may make exposures grade into a single event from the perspective of the receiving animal.

Individual Duration (s) – The duration of individual stimuli should be specified, including the method used to measure duration.

Individual Rise Time (s) – The rise time of individual stimuli should be specified. Time in seconds is the most usual metric, although dB/s is useful in cases where the maximum level is high and the rise time short because it also captures the difference in level.

Total Exposure Duration (s) – The total duration of a trial exposure begins with the start of the first sound segment of a stimulus and ends with the completion of the last. This should be specified explicitly. Note that this definition captures the exposure from the perspective of the receiver. From the perspective of the experimenter, exposure may be defined as lasting for a period of observation after the sound stops.

Order If Multiple Exposures (Identify Sequence/Order) – The order of a particular stimulus and observation should be specified where animals are exposed multiple times. Metadata should include the scheme for choosing the order of exposure (e.g., increasing or randomized with or without replacement).

Harmonics Present? (None, Few, Many) – Harmonics can alter the detectability of sounds (Cunningham et al., 2016). Therefore, the presence of harmonics, along with an indication of the bandwidth covered, is an important factor.

Root Mean Square Sound Pressure Level – Root mean square sound pressure level (RMS SPL) should be specified at the point where behavior changes or as a minimum and maximum if no change in behavior is observed. It should be specified as a broadband level, with the bandwidth of the recording system specified as well, and the maximum 1/3-octave band level (see "Sensation Level" below for a comment about power spectral density). It is also an important measure of level where sounds with wide and especially widely differing bandwidths are being compared (see discussion in Ellison et al., 2012).

Peak-to-Peak RL – The peak-to-peak received sound pressure level is the difference between the absolute value of the maximum negative and maximum positive instantaneous peak pressure received at the subject animal's level. It should be specified in dB (units of SPL). It is measured in the time domain at the point in time where behavior changes or at the point of maximum exposure if no change in behavior is observed. Attention should be given to the properties of the recording system to ensure that sampling is fast enough (i.e., that it has sufficient bandwidth) to effectively measure the peak.

Sound Exposure Level – Sound exposure level (SEL) is related to the energy in an exposure but can be calculated readily from pressure measurements, so the metric is treated as a measure of sound exposure energy (E), calculated over a given time interval (*T*), with the media-specific reference pressure (*p*), by integrating squared pressures (P^2) in sampling increments (*t*) from 0 to *T*. This enables sounds of differing duration to be related in terms of total energy for purposes of assessing exposure risk. We advocate using a simplified version of the formula presented in Southall et al. (2007) as given below, but see Martin et al. (2019) for a more in-depth discussion. Expressed as a level, the formula is

$$SEL = 10\log_{10}\left(\frac{1}{T_0 p_0^2}\int_0^T P^2 t(dt)\right)$$

SEL is simply the decibel level of the cumulative sum of square pressures (E(*t*)) for a 1-s equivalent duration referenced to the media-specific reference pressure (1 or 20 µPa for water and air, respectively). Consequently, the appropriate units for underwater SEL are dB re 1 µPa^2-s, and the appropriate units for aerial SEL are dB re (20 µPa)^2-s. Note that this means that SEL is referenced to 1 s in both media. SEL should be reported for the point at which behavior changes or at the maximum exposure level if there is no change in behavior. It should be provided for the full bandwidth and maximum 1/3-octave band level, and the bandwidths of the recording system should be included in the metadata.

SELcum – For single events, it is easy to understand the above definition as a measure of the energy in the signal. However, for transients repeated at intervals, there are two general approaches to the calculation of cumulative exposure level. First, it can be calculated in exactly the same way as SEL for individual events, including both the transients and the intervals between. However, if the signal-to-noise ratio of the sounds (e.g., pile driving) is high, there will be a large difference between SEL calculated over the entire period of an exposure and SEL calculated using the second approach, adding together the events only (i.e., without the intervals between). The second approach is typically used to calculate SELcum.

Explicitly, SELcum is calculated for sounds that are intermittent or repeated over time by integrating pressure squared for repeated instances of the sound only when it is present using the equation above (see, also, Southall et al., 2007). Note that this summation procedure essentially generates a single exposure "equivalent" value that does not consider what happens between repeated exposures. The method for isolating the events in the calculation of SELcum should be specified such as the duration encompassing 95% of the energy in the event.

Viewed from the perspective of behavioral disturbance, SEL was devised to provide an "apples-to-apples" comparison of transient sounds of varying duration (e.g., aircraft overflights or passes by snowmobiles). It is sometimes applied to more continuous noise in standardized time intervals (i.e., an hour or an 8-h workday), but other measures are generally used for human community noise applications because the energy in a sound stops being a good predictor of disturbance when integrated over long periods. However, no such standards are available for marine mammal applications for long-term, sustained exposures when animals may be coming in and out of an area and exposures may be intermittent (as in studies described in Appendix 2).

As defined, SEL increases arbitrarily over long periods. The problem is greater in the case of SELcum because quieter periods between exposures to transients such as discharge of an airgun or pile driving are not included. Therefore, for SELcum to be useful, the total exposure time must be constrained somehow—for example, by estimating how long an animal will be near an intermittent sound source. SELcum is a very appropriate measure for experimental transients or playback of fairly short-term stimuli in CEEs, but care is needed when making comparisons across studies or between sound types. This is one of the reasons we have included time-domain features such as intermittency in the list of metrics that must be specified.

SELcum should be provided for the point where behavior changes or for the maximum exposure if there is no behavior change. The intervals over which it is calculated should be defined explicitly. It should be calculated for the full measurement bandwidth and in the maximum 1/3-octave band. The bandwidth of the recording system should be included in the metadata.

Signal-to-Noise Ratio – Signal-to-noise ratio (SNR) is the difference between the level of a target sound and the level of the background noise in the time-frame of the signal. It should be estimated at the time that behavior changes or the point at which the signal is maximal if there is no response. The maximum 1/3-octave band should be used to estimate SNR.

Sensation Level – Sensation level (SnL) is defined as the band-specific difference between the RL of a signal and the species-specific auditory detection threshold, so long as the latter exceeds ambient noise within the band (as in Ellison et al., 2012). We advocate for a 1/3-octave band in making SnL calculations accordingly and for reporting the maximum value for any 1/3-octave band as the effective SnL.

Literature Cited

Augusto, J. F., Frasier, T. R., & Whitehead, H. (2017). Characterizing alloparental care in the pilot whale (*Globicephala melas*) population that summers off Cape Breton, Nova Scotia, Canada. *Marine Mammal Science*, *33*(2), 440-456. https://doi.org/10.1111/mms.12377

Committee on Taxonomy. (2021). *List of marine mammal species and subspecies*. Society for Marine Mammalogy. www.marinemammalscience.org

Cunningham, K. A., Southall, B. L., & Reichmuth, C. (2016). Detection of complex sounds in quiet conditions by seals and sea lions. In A. N. Popper & A. Hawkins (Eds.), *The effects of noise on aquatic life II*: *Advances in experimental medicine and biology* (Vol. 875, pp. 181-187). https://doi.org/10.1007/978-1-4939-2981-8_21

Ellison, W. T., Southall, B. L., Clark, C. W., & Frankel, A. S. (2012). A new context-based approach to assess marine mammal behavioral responses to anthropogenic sounds. *Conservation Biology*, *26*(1), 21-28. https://doi.org/10.1111/j.1523-1739.2011.01803.x

Frid, A., & Dill, L. M. (2002). Human-caused disturbance stimuli as a form of predation risk. *Conservation Ecology*, *6*(1), 11. www.consecol.org/vol6/iss1/art11

Harris, C. M., Thomas, L., Falcone, E. A., Hildebrand, J., Houser, D., Kvadsheim, P. H., Lam, F-P. A., Miller, P. J. O., Moretti, D. J., Read, A. J., Slabbekoorn, H., Southall, B. L., Tyack, P. L., Wartzok, D., & Janik, V. M. (2018). Marine mammals and sonar: Dose-response studies, the risk-disturbance hypothesis and the role of exposure context. *Journal of Applied Ecology*, *55*, 396-404. https://doi.org/10.1111/1365-2664.12955

Heithaus, M. R., & Frid, A. (2003). Optimal diving under the risk of predation. *Journal of Theoretical Biology*, *223*, 79-92. https://doi.org/10.1016/S0022-5193(03)00073-0

Lee, N., Ward, J. L., Vélez, A., Micheyl, C., & Bee, M. A. (2017). Frogs exploit statistical regularities in noisy acoustic scenes to solve cocktail-party-like problems. *Current Biology*, *27*, 743-750. https://doi.org/10.1016/j.cub.2017.01.031

Martin, S. B., Morris, C., Bröker, K., & O'Neill, C. (2019). Sound exposure level as a metric for analyzing and managing underwater soundscapes. *The Journal of the Acoustical Society of America*, *146*, 135-149. https://doi.org/10.1121/1.5113578

Rankin, S., Archer, E., & Barlow, J. (2013). Vocal activity of tropical dolphins is inhibited by the presence of killer whales, *Orcinus orca*. *Marine Mammal Science*, *29*(4), 679-690. https://doi.org/10.1111/j.1748-7692.2012.00613.x

Southall, B. L., Finneran, J. J., Reichmuth, C., Nachtigall, P. E., Ketten, D. R., Bowles, A. E., Ellison, W. T., Nowacek, D. P., & Tyack, P. L. (2019). Marine mammal noise exposure criteria: Updated scientific recommendations for residual hearing effects. *Aquatic Mammals*, *45*(2), 125-232. https://doi.org/10.1578/AM.45.2.2019.125

Southall, B. L., Bowles, A. E., Ellison, W. T., Finneran, J. J., Gentry, R. L., Greene, C. R., Kastak, D., Ketten, D. R., Miller, J. H., Nachtigall, P. E., Richardson, W. J., Thomas, J. A., & Tyack, P. L. (2007). Marine mammal noise exposure criteria. *Aquatic Mammals*, *33*(4), 411-522. https://doi.org/10.1578/AM.33.4.2007.411

Appendix 2. Supplementary Details for Review and Assessment of Population-Level Studies

The following provides additional detail and discussion for population-level studies not discussed in detail in the main text. This includes multi-year whale-watching studies (from Table 9), decadal-scale whale-watching studies (from Table 10), and multi-year to decadal studies of other disturbances (Table 11).

Population-Level Studies: Multi-Year Whale-Watching Studies

Lusseau (2004, 2005) studied whale-watching impacts on two small bottlenose dolphin populations in Fjordland, New Zealand, specifically in Doubtful Sound and Milford Sound. Impacts were measured by calculating the average amount of time dolphins spent with boats on a given day and the cumulative behavioral budget of dolphins in situations with variable vessel presence. Behavioral budgets were affected in both populations by boat presence, with resting behavior being the most affected behavioral state. Whale watching in Fjordland began in the late 1980s and, due to the isolation of the area, tourism activities account for most of the boat traffic. Milford Sound is more heavily targeted by the tourism industry than Doubtful Sound. However, the cumulative time spent by the dolphins interacting with boats was similar in the two fjords and, despite the lower pressure, the effect was stronger in Doubtful Sound with dolphins' socializing behavior strongly impacted. In Milford Sound, dolphins were more frequently sighted in winter when boat traffic was lower, as well as in the outer part of the fjord where boats spent less time.

Population-Level Studies: Decadal-Scale Whale-Watching Studies

Swim-with tourism carries similar concerns to whale watching for the targeted individual/population (Samuels et al., 2000). By analyzing two datasets collected 15 y apart, Filby et al. (2014) measured long-term effects of the swim-with industry in Port Phillip Bay, Australia. The area is in close proximity to a major urban center covering numerous anthropogenic activities, including exposure to a non-compliant commercial dolphin-swim industry that started in 1986 and included three licensed operators with four vessels each running two trips per day at the time of the study. The study categorized dolphin responses as "effect" (avoidance and approach) or "no effect" (neutral response), demonstrating substantial increases of avoidance responses with resting and small groups avoiding boats more frequently and increasing swim time between seasons. Sighting rates of individually identified animals also decreased within and across seasons. This could be due to a decrease in population size or habitat displacement, either of which might have been affected by disturbance along with many other uncontrolled factors. If disturbance did cause some animals to shift habitat, this does not indicate the extent of disturbance as some individuals might stay subjected to disturbance pressure because of the importance of the site as a calving and breeding ground.

Food provisioning in the context of tourism interactions with marine mammals can elicit behavioural conditioning to human food sources and maladaptive behavior such as begging, which, in turn, has been linked to adverse consequences on survival and reproductive success (Christiansen et al., 2016; Senigaglia et al., 2019). In a recent study, Senigaglia and colleagues (2019) used two decades of citizen science data and over 10 y of systematic survey and demographic information to assess the effect of food provisioning on a population of Indo-Pacific bottlenose dolphins in Bunbury, Australia. Using both Bayes factor analyses and GLMs, the authors found a strong correlation between provisioning and females' reproductive success, defined as the number of weaned calves over the entire reproductive life of an individual. Despite also examining the effects of climate change, begging behavior, and habitat characteristics, results showed that non-provisioned females have double the mean reproductive success than provisioned individuals. Moreover, Senigaglia et al. reported that only one third of the calves born to provisioned mothers were successfully weaned compared to a 77% weaning rate of non-provisioned females. Despite the small sample size, this study highlights the repercussions of a tourism activity involving food handouts on free-ranging dolphins—in particular, when the population is also subjected to multiple and concurrent other stressors and is already declining (Senigaglia, 2020).

Population-Level Studies: Multi-Year to Decadal-Scale Studies of Other Disturbances

Pirotta et al. (2014) measured fine-scale impacts by investigating spatially explicit individual exposure of bottlenose dolphins in Moray Firth, Scotland. Despite the area's protection as a Special Area of Conservation, dolphins are subject to multiple disturbances, including high vessel traffic, fisheries, oil exploration, and recreational activities. Pirotta and colleagues developed individual models combining results from previous studies on habitat utilization, boat traffic, boat disturbance, and construction activities on dolphin behavior. This study illustrated cumulative impacts from different anthropogenic activities with widely variable exposure to vessel traffic. Pirotta et al. (2015) expanded on these results, predicting population-level consequences of aggregate disturbance. Changes in motivational states (energy acquisition vs expenditure) across time were linked to health and calf survival. The model was subsequently applied to three scenarios to predict population consequences: (1) complete satisfaction of motivational state, (2) satisfaction on average, and (3) dissatisfaction with their motivational state (resulting in possible decline caused by individuals being unable to meet their needs). Animal exposure and motivational states did not differ as a result of modeled increases in boat traffic and dredging activity, except during the operational phase when relatively small increases in boat interactions experienced by each individual caused a shift of motivational states toward dissatisfaction. The model did not detect an association between predicted exposure of female dolphins, motivational states, and calf survival.

New et al. (2013) modeled potential interacting effects of the construction of offshore wind farms with different scenarios of vessel traffic on a coastal population of bottlenose dolphins in Moray Firth, Scotland. By simulating the social, spatial, behavioral, and motivational interactions of bottlenose dolphins, they identified a deterministic link between health and motivation and derived impacts from changes in behavioral state due to underlying changes in motivational state influenced by disturbance. The simulation did not detect long-term consequences of increased disturbance despite a more than six-fold increase in vessel traffic. However, the simulated disturbance did not include vessels directly targeting interactions with dolphins nor did it include information on noise levels. Thus, the lack of detectable impact might be due to the type of disturbance and the availability of adjacent "undisturbed" areas within which mobile individuals could avoid human interactions.

Literature Cited

Christiansen, F., McHugh, K. A., Bejder, L., Siegal, E. M., Lusseau, D., McCabe, E. B., Lovewell, G., & Wells, R. S. (2016). Food provisioning increases the risk of injury in a long-lived marine top predator. *Royal Society Open Science*, *3*, 160560. https://doi.org/10.1098/rsos.160560

Filby, N. E., Stockin, K. A., & Scarpaci, C. (2014). Long-term responses of Burrunan dolphins (*Tursiops australis*) to swim-with dolphin tourism in Port Phillip Bay, Victoria, Australia: A population at risk. *Global Ecology and Conservation*, *2*, 62-71. https://doi.org/10.1016/j.gecco.2014.08.006

Lusseau, D. (2004). The hidden cost of tourism: Detecting long-term effects of tourism using behavioral information. *Ecology and Society*, *9*(1), 2. https://doi.org/10.5751/ES-00614-090102

Lusseau, D. (2005). Residency pattern of bottlenose dolphins *Tursiops* spp. in Milford Sound, New Zealand, is related to boat traffic. *Marine Ecology Progress Series*, *295*, 265-272. https://doi.org/10.3354/meps295265

New, L. F., Harwood, J., Thomas, L., Donovan, C., Clark, J. S., Hastie, G., Thompson, P. M., Cheney, B., Scott-Hayward, L., & Lusseau, D. (2013). Modelling the biological significance of behavioural change in coastal bottlenose dolphins in response to disturbance. *Functional Ecology*, *27*(2), 314-322. https://doi.org/10.1111/1365-2435.12052

Pirotta, E., Thompson, P. M., Cheney, B., Donovan, C. R., & Lusseau, D. (2014). Estimating spatial, temporal and individual variability in dolphin cumulative exposure to boat traffic using spatially explicit capture–recapture methods. *Animal Conservation*, *18*(1), 20-31. https://doi.org/10.1111/acv.12132

Pirotta, E., Harwood, J., Thompson, P. M., New, L., Cheney, B., Arso, M., Hammond, P. S., Donovan, C., & Lusseau, D. (2015). Predicting the effects of human developments on individual dolphins to understand potential long-term population consequences. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1818), 20152109. https://doi.org/10.1098/rspb.2015.2109

Samuels, A., Bejder, L., & Heinrich, S. (2000). *A review of the literature pertaining to swimming with wild dolphins* (Contract Number T74463123). Marine Mammal Commission. 57 pp.

Senigaglia, V. (2020). *Ecological consequences and social drivers of human-wildlife interactions: The case of food-provisioning of bottlenose dolphins in Bunbury, Western Australia* (Ph.D. thesis). Murdoch University, Perth, Australia.

Senigaglia, V., Christiansen, F., Sprogis, K. R., Symons, J., & Bejder, L. (2019). Food-provisioning negatively affects calf survival and female reproductive success in bottlenose dolphins. *Scientific Reports*, *9*, 8981. https://doi.org/10.1038/s41598-019-45395-6