ICES Journal of Marine Science

ICES Journal of Marine Science (2016), 73(3), 572–581. doi:10.1093/icesjms/fsv118

Contribution to Special Issue: 'Towards a Broader Perspective on Ocean Acidification Research'

Original Article

Experimental design in ocean acidification research: problems and solutions

Christopher E. Cornwall[1,2]* and Catriona L. Hurd[1]

[1]*Institute for Marine and Antarctic Studies, University of Tasmania, Private Bag 129, Hobart, TAS 7001, Australia*
[2]*School of Earth and Environment and ARC Centre of Excellence in Coral Reef Studies, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*

**Corresponding author: tel: +61 8 6488 3644; fax: +61 3 62262973; e-mail: christopher.cornwall@uwa.edu.au*

Cornwall, C. E., and Hurd, C. L. Experimental design in ocean acidification research: problems and solutions. – ICES Journal of Marine Science, 73: 572–581.

Received 26 February 2015; revised 8 June 2015; accepted 10 June 2015; advance access publication 8 July 2015.

Ocean acidification has been identified as a risk to marine ecosystems, and substantial scientific effort has been expended on investigating its effects, mostly in laboratory manipulation experiments. However, performing these manipulations correctly can be logistically difficult, and correctly designing experiments is complex, in part because of the rigorous requirements for manipulating and monitoring seawater carbonate chemistry. To assess the use of appropriate experimental design in ocean acidification research, 465 studies published between 1993 and 2014 were surveyed, focusing on the methods used to replicate experimental units. The proportion of studies that had interdependent or non-randomly interspersed treatment replicates, or did not report sufficient methodological details was 95%. Furthermore, 21% of studies did not provide any details of experimental design, 17% of studies otherwise segregated all the replicates for one treatment in one space, 15% of studies replicated CO_2 treatments in a way that made replicates more interdependent within treatments than between treatments, and 13% of studies did not report if replicates of all treatments were randomly interspersed. As a consequence, the number of experimental units used per treatment in studies was low (mean = 2.0). In a comparable analysis, there was a significant decrease in the number of published studies that employed inappropriate chemical methods of manipulating seawater (i.e. acid–base only additions) from 21 to 3%, following the release of the "Guide to best practices for ocean acidification research and data reporting" in 2010; however, no such increase in the use of appropriate replication and experimental design was observed after 2010. We provide guidelines on how to design ocean acidification laboratory experiments that incorporate the rigorous requirements for monitoring and measuring carbonate chemistry with a level of replication that increases the chances of accurate detection of biological responses to ocean acidification.

Keywords: experimental design, manipulation experiments, ocean acidification, pseudoreplication, replication.

Introduction

Ocean acidification is a potential threat to marine ecosystems through its effect on the physiology and ecology of many marine species (e.g. Kroeker *et al.*, 2013b). As the research effort being invested into examining the effects of ocean acidification on marine systems increases, publications on the topic are increasing exponentially (Gattuso *et al.*, 2012; Riebesell and Gattuso, 2015). Experimental manipulations of CO_2 concentrations in the field are difficult, and the number of field studies are limited to a few locales where high CO_2 naturally occurs, or where large scale, costly, and labour intensive experiments have been employed (Barry *et al.*, 2010; Gattuso *et al.*, 2014). Consequently, the most studies are conducted in the laboratory (see reviews by Wernberg *et al.*, 2012; Kroeker *et al.*, 2013b), where CO_2 concentrations can potentially be controlled and reported accurately, and their effects isolated from those of other environmental variables. A multidisciplinary field of research such as ocean acidification brings together various research expertise including engineering sophisticated systems to manipulate and monitor seawater carbonate chemistry, building systems that can house organisms for long periods, and

expertise in measuring the appropriate physiological/biogeochemical/ecological responses to ocean acidification. Crucially, if the field of ocean acidification is to progress, these manipulation experiments need to be performed in such a way that clear conclusions can be drawn from the results.

One of the major challenges for ocean acidification research is that seawater carbonate chemistry must be manipulated correctly in order for experimental treatments to approximately simulate future high CO_2 oceans. Outwardly, this seems rather simple: seawater pH is manipulated using CO_2 gas or the chemically equivalent method of HCl and dissolved inorganic carbon (DIC; usually in the form of $NaHCO_3$ or Na_2CO_3), in a way that mimics changes in the future seawater, increasing in CO_2 concentrations and declines in pH. That is, total alkalinity (A_T) remains constant and DIC increases (Rost *et al.*, 2008; Hurd *et al.*, 2009; Gattuso *et al.*, 2010).

Other logistical constraints, however, make ocean acidification experiments more difficult than those in other related fields of biological research. Experimental tanks often need to be sealed because CO_2 can degas when exposed to the atmosphere, and the tanks require sufficient rates of seawater flow-through because the metabolic activity of organisms can modify seawater DIC and A_T (Rost *et al.*, 2008). The constant addition of chemicals (i.e. CO_2 or HCl/DIC) into experimental tanks over long periods of time, without directly exposing organisms to un-mixed chemicals and seawater, adds another aspect of logistic difficulty to ocean acidification research that is not present in many other manipulation experiments (Hurd *et al.*, 2009; Riebesell *et al.*, 2010b). Appropriate methods must also be employed to determine whether the seawater that organisms are exposed to represents the desired treatments required for ocean acidification research; at least two components of the seawater carbonate system must also be measured (pH, A_T, DIC, or pCO_2). If pH is used to parametrize the seawater carbonate system, then it must be measured using spectrophotometers or electrodes calibrated using TRIS buffers, not with electrodes calibrated using NBS buffers (Dickson *et al.*, 2007; Dickson, 2010). Also, to eliminate or reduce the chance of experimental artefacts or pre-existing gradients in other factors influencing treatments differently, treatments within manipulation experiments must all contain adequate numbers of randomly interspersed and independent treatment replicates (Hurlbert, 1984; Hurlbert and White, 1993; Hurd *et al.*, 2009; Hurlbert, 2009; Wernberg *et al.*, 2012). To complicate matters, ocean acidification will not be occurring in isolation, as other anthropogenic effects such as global warming and localized altered salinity, nutrients (nitrogen, phosphorous), light regimes, storm occurrence, and land-born pollutants will be occurring in synergy (Feely *et al.*, 2004; Boyd, 2011; Ciais *et al.*, 2013). Factorial designs where CO_2 is manipulated in combination with other factors therefore add further complexity and logistical challenges to experimental designs (Havenhand *et al.*, 2010; Wernberg *et al.*, 2012).

The inappropriate assignment of experimental units can be a problem in any form of research, and Hurlbert (1984) defines this inappropriate assignment of experimental units for a given treatment during statistical analyses as "pseudoreplication". Hurlbert defines the appropriate procedures for eliminating the risk of pseudoreplication by randomly interspersing replicates of different treatments with each other, and by removing any interdependence within replicates from the same treatment, i.e. experimental units are randomly interspersed replicates of a treatment. These procedures detailed by Hurlbert (1984) will not be repeated here in detail, but solutions to common problems in ocean acidification manipulation experiments will be mentioned in Discussion. Hurlbert (2013a) also defines experimental units as: "the smallest . . . unit of experimental material to which a single treatment (or treatment combination) is assigned by the experimenter and is dealt with independently . . . ", and defines independent experimental units as being units assigned to the same treatment that will not be subject to conditions that are more similar than conditions that are imposed on units from another treatment, other than the treatment under investigation (Cox, 1958; Kozlov and Hurlbert, 2006; Hurlbert, 2013a). The experimental unit can be constrained by two principles: (i) experimental units within one treatment must not influence each other more than they influence experimental units within another treatment and (ii) factors other than the treatment in question (e.g. seawater source, light, etc.) must, on average, be equal across all treatments (Hurlbert, 2013a). If these principles are adhered to, it will greatly reduce the risk of non-treatment effects differentially influencing one treatment and not others; these should be the basic tenets of experimental design. Regardless of the degree of precision that the treatment is applied and its effects measured, if treatment effects are confused with the effects of other factors not under investigation, then an accurate assessment of the effects of the treatment cannot be made. Hurlbert and White (1993) further define three types of pseudoreplication: (i) simple pseudoreplication, where there is one experimental unit per treatment and multiple individuals in one experimental unit whose responses to the treatment are measured (defined by Hurlbert, 2009 as the "evaluation unit") and treated as though they are independent experimental units; (ii) temporal pseudoreplication, where multiple measurements are made though time on the same experimental unit and treated as independent experimental units of a treatment; (iii) sacrificial pseudoreplication, where there are multiple experimental units per treatment and multiple individuals within each experimental unit, but the individuals are treated as the experimental units during statistical analysis. These three definitions demonstrate how a misinterpretation as to what constitutes an experimental unit vs. an "evaluation unit" could lead to inappropriate design and analysis.

The field of ocean acidification research is rapidly expanding, and so far the scientific community has revealed information crucial to understanding its future impacts at an impressive rate unprecedented in many other fields of research (Riebesell and Gattuso, 2015). However, if the field is to progress, then we need to maximize the information provided by each future study. The purpose of this study is not to exhaustively detail how to correctly replicate experiments of all types; these have already been fully explained previously (Cox, 1958; Hurlbert, 1984; Mead, 1988). Nor is it our goal to undermine previously conducted research, which has significantly advanced understanding of the potential effects of ocean acidification on biological systems (Riebesell and Gattuso, 2015). The objectives of this study is to highlight how well the basic principles for the design and analysis of experiments are followed (e.g. Cox, 1958; Hurlbert, 1984; Mead, 1988; Hurd *et al.*, 2009; Havenhand *et al.*, 2010; Wernberg *et al.*, 2012), and to highlight how the mistreatment of experimental units (not just during statistical analysis) hinders the ability to accurately predict the effects of ocean acidification in certain circumstances. Importantly, it was our goal to provide solutions to many commonly encountered problems in experimental design. The appropriate procedures to replicate experimental units in ocean acidification manipulation experiments are identified, and the prevalence of these appropriate methodological approaches is quantified. As well as presenting potential pitfalls in experimental design, a range of solutions to logistical limitations that are imposed by different CO_2-manipulation system designs

are provided, to increase the inference that can be drawn from future manipulation experiments.

Methods

To measure the prevalence of different designs in ocean acidification manipulation experiments, we searched the database Scopus http://www.scopus.com/ using the term "ocean acidification". We cross checked this search with the database used by Kroeker *et al.* (2013b) and the ocean acidification blog http://news-oceanacidification-icc.org/. Four hundred and sixty-five studies published between 1993 and February 2014 were examined, along with any others that came across our desk thereafter (Supplementary Table SI). Sixty-two of these studies were also analysed by Wernberg *et al.* (2012). Studies were only analysed if they met specific criteria. That is, the studies had to report: (i) the results of research containing the assessment of biological responses to experimental manipulations of seawater CO_2; (ii) state that their results were directly applicable to predicting the impacts of ocean acidification, increases in seawater CO_2 concentrations predicted for the future, or another analogous term. Studies that were investigating the impacts of carbon sequestration in the seabed, and made no mention of future high CO_2 seawater, ocean acidification, etc., were excluded. (iii) They needed to be conducted for longer than 24 h, in other words, we excluded short-term "response-assays". (iv) They had to be published in peer reviewed journals to be included. We excluded field-based correlative surveys due to the logistical limitations imposed on them by lack of availability of multiple "treatment" sites in some locations, and because treatments were not "manipulated" by the researchers usually. This is not to say that we viewed these studies as less worthy than laboratory studies.

After papers were identified that met the above criteria, their experimental design in respect to replication was examined. Specific designs were grouped as per the scheme outlined in Figure 1 in Hurlbert (1984) which we have re-constructed in the context of ocean acidification research (Figure 1), showing potential aquaria or "experimental tank" arrays. Hulbert considers designs whose letters start with an "A" are an appropriate design choice, while those with "B" are an inappropriate choice. They are A-1 completely randomized, A-2 randomized block, A-3 systematic, B-1 simple segregation, B-2 clumped segregation, B-3 isolative segregation, B-4 randomized but with interdependent treatment replicates, and B-5 no replication. In each study, the number of actual experimental units per treatment was also recorded, rather than those stated by the authors. In the context of ocean acidification research, we provide the following examples for each of the B type designs (Figure 1): (B-1) three different pCO_2 treatments are used and experimental tanks are lined up on a bench in order from lowest to highest pCO_2; (B-2) three different shaker tables are used next to each other, and each table contains the replicates of only one treatment; (B-3) three different experimental tanks are used, and each experimental tank contains all the individuals of one treatment only; (B-4) three different header tanks are used to supply seawater for an entire experiment, and one header tank provides seawater for all the replicates of one treatment only; (B-5) one individual per treatment is placed in one tank, and only one tank is used for each treatment of an experiment. In all these examples, pre-existing gradients in other factors or chance events could obscure the effects of the factor under investigation.

In our survey, we assumed that cylinders/containers of CO_2 gas, HCl, or a form of DIC is not likely to be a source of experimental artefact. In other words, we considered that using one source of chemicals to modify carbonate chemistry, and applying it to seawater in all elevated CO_2 treatments would not result in a lack of independence, as we considered that the likelihood of a contaminant such as a type of micro-organism (fungi, bacteria, diatoms, etc.)

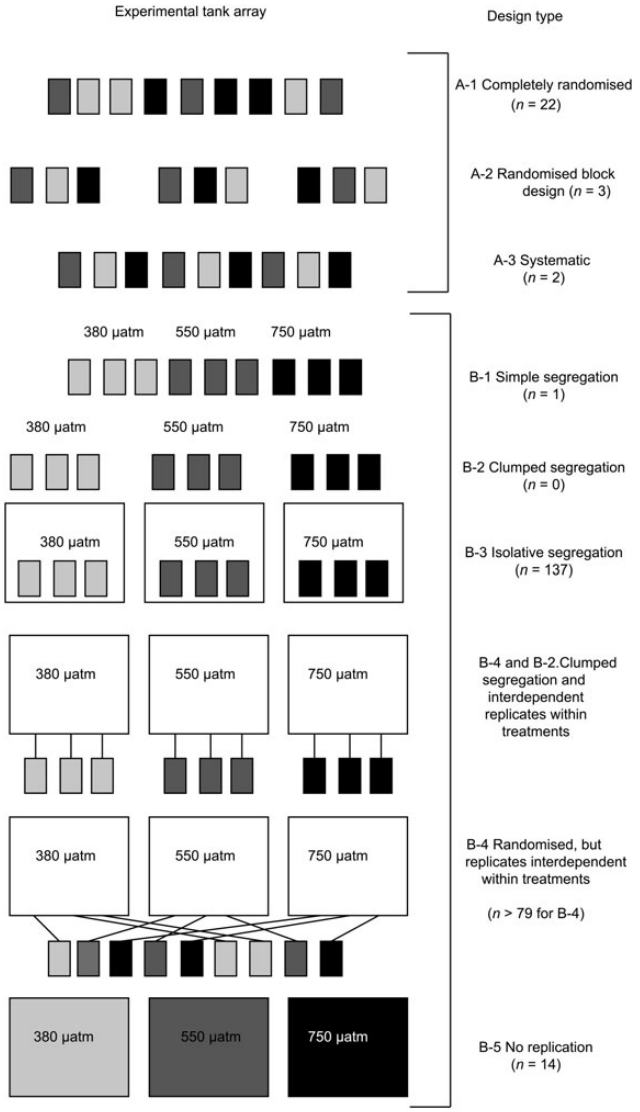

Figure 1. Different designs of tank arrays, modified and re-drawn from Hurlbert (1984) in the context of ocean acidification research. Different coloured tanks correspond to different CO_2 treatments. Design types preceded with an A are acceptable, while those preceded by a B are unacceptable ways to replicate experimental units of a treatment according to Hulbert. (A-1) Completely randomized. (A-2) Randomized block design. (A-3) Systematic. (B-1) Simple segregation. (B-2) Clumped segregation. (B-3) Isolative segregation. (B-4) Randomized but all replicates of one treatment interdependent with themselves more than other treatments (e.g. one header tank of seawater per treatment). (B-5) No replication. An example of replicates within treatments that are interdependent are treatment replicates that all share a common header tank that is not shared with replicates of other treatments large white boxes denote tanks of seawater that do not contain the organisms that are housed. Smaller boxes denote tanks that the organisms are housed in, with different colours representing different pCO_2 levels. n = number of studies using this design type in ocean acidification research. See results and methods for more details.

contaminating that container and changing the treatment was extremely low compared with the likelihood of micro-organisms living in tanks of seawater. This could be an issue, but was outside of the bounds of our analysis, as it is difficult to expect studies to report the number of containers of chemicals used.

The "Guide to Best Practices for Ocean Acidification Research and Data Reporting" (Dickson, 2010; Gattuso *et al.*, 2010; Riebesell *et al.*, 2010a) provided rationale and methods detailing how to manipulate seawater carbonate chemistry in a way that most accurately mimics the changes predicted for a future ocean, and guidelines on the appropriate methods to characterize seawater carbonate chemistry. To determine the impact of this publication on the frequency of studies that used appropriate methods to manipulate and monitor seawater carbonate chemistry, we examined the same 465 papers to determine if they altered seawater chemistry in a way that simulates future ocean acidification (using methods that should increase DIC and keep total alkalinity (A_T) constant) or not (methods resulting in constant DIC and decreased A_T). In other words, we examined whether HCl was used by itself to reduce seawater pH without the addition of DIC. Two components of the seawater carbonate system must be measured (pH, A_T, DIC, or pCO_2) along with temperature and salinity to appropriately measure seawater carbonate chemistry, and if pH is measured it should be measured on the total scale: involving using TRIS buffers to calibrate electrodes, or using spectrophotometric measurements (see Dickson *et al.*, 2007; Dickson, 2010 for complete guidelines). We also recorded the number of papers that measured two or more of seawater pH_T, A_T, DIC, or pCO_2.

Statistical analysis

We analysed whether or not there was a difference pre- and post-2010 (the year of the publication of The Guide to Best Practices in Ocean Acidification and Data Reporting) between the proportion of studies that possibly used inappropriate methods: (i) to manipulate seawater carbonate chemistry; (ii) to measure two or more of pH_T, A_T, DIC, or pCO_2; and (iii) to design/analyse experimental units. This was done using a z-test to examine differences between two proportions.

Definitions of tank types

In ocean acidification experiments, different containers (referred to here as tanks) have different purposes, and for clarity we assign names to each tank type. All manipulation experiments must have at least one of these tanks types, but not all laboratory manipulation experiments will use all types, and designs could incorporate tanks that have dual purposes. They are as follows: (i) storage tank: location where seawater is stored before being altered to create CO_2 treatments. (ii) Mixing tank: location where chemicals and seawater are mixed together to create CO_2 treatments before contact with the study species. (iii) Header tank: location where seawater is stored after mixing with chemicals, but before exposure to organisms. (iv) Experimental tank: the putative experimental unit, location where organisms are housed in the treatment seawater. Note that these "tanks" may not all be traditional aquaria, but can be any type of container that stores seawater for any length of time, and that in some designs storage "tanks" could technically be the ocean or pipes leading to tanks.

Results

The use of interdependent or non-randomly interspersed treatment replicates in experimental designs (i.e. B designs) was more common (173 studies) than the use of independent, randomly interspersed replicates of treatments (i.e. A designs, 28 studies) among ocean acidification manipulation studies (Figure 2). Note that 20 B-designed studies used random or nested factors in their analysis to attempt to deal with this type of data, and those without other design problems were therefore classified as A designs. A further 17 B studies initially also used this approach and dropped the random effect when it was not significant. The most common design type was B-3 (isolative segregation—Figure 1). Most studies that used designs that were classified as B type designs did so by using one of the following methods: (i) Multiple individuals were housed within one experimental tank but each individual was treated as an experimental unit (i.e. B-3 if only one tank per treatment existed); in this case, the tank itself is the experimental unit, giving $n = 1$ in this example. Often the subsampling of individuals from one replicate container was not mentioned, but then the degrees of freedom in the statistical analysis were higher than the number of experimental units per treatment (i.e. experimental tanks), which occurred in nine studies where the design and sampling procedures could not otherwise attribute it to another type of design. (ii) Utilizing one header tank per treatment, then having the header tank periodically or continuously deliver seawater to multiple experimental tanks, with no acknowledgment by the authors that the experimental tanks are interdependent (68 studies; design B-4). Note: it was deemed appropriate if all experimental units of all treatments are supplied with seawater from one header tank at the beginning of the experiment only. (iii) All replicates of one treatment were grown in one room or water bath, while the other treatments were housed in different room or water bath (B-3). There was no difference between the proportion of studies using appropriate designs before or after 2010 (94.3% in 2010 and before, 93.7% after 2010; $z = -0.25$, $P = 0.79$).

Around half (57%) of the studies did not present sufficient details on the design type to determine the precise design, and 34% of all studies total did not provide sufficient details to determine whether it was an "A" type or "B" type design. Using the limited information that was provided in these studies, it was determined that 28% of studies were either B-3 or B-4 designs (isolative segregation or interdependent replicates), while 24% were A-1 to B-2 designs. In the remainder of cases, insufficient information was provided to narrow down the design to any definitive type. For those studies, often there was no mention of replicate numbers,

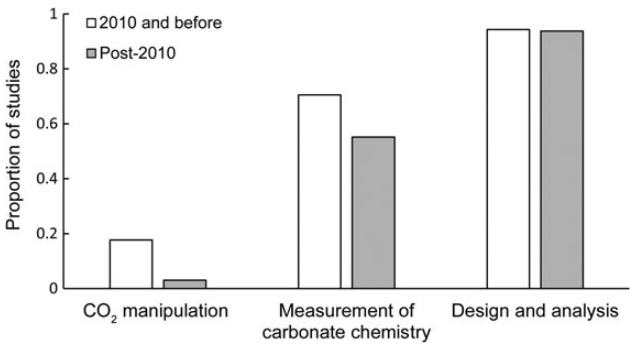

Figure 2. Proportion of studies that possibly use inappropriate methods to manipulate seawater CO_2 concentrations, to measure the carbonate chemistry of seawater, or to design their experiments/miss-assign the experimental unit during statistical analysis. See details in Introduction for descriptions/references to appropriate vs. inappropriate methods.

different tank type numbers, nor how the carbonate chemistry was modified—beyond mentioning the chemicals used—nor how seawater was introduced to the experimental organisms. In many instances the degrees of freedom were not provided, and in some cases it is impossible to derive the statistical differences that were reported by the authors when the true number of experimental units per treatment was so low (i.e. when $n \leq 2$). In other words, individuals were treated as treatment replicates rather than the experimental units being used as treatment replicates.

Among the 30% of studies (i.e. 180) where we could determine if treatment replicates were interspersed randomly or not, 36 studies employed a randomized design, whereas 130 were non-random and 13 employed no replication of treatments. We could not determine whether replicates of all treatments were randomly interspersed in 280 studies; this represented the sole problem in 13% of all studies examined.

Across all studies the number of true experimental units that were used per treatment was extremely low (mean = 2.00, s.e. = 0.12). A disparity between the numbers of actual and stated experimental units per treatment was often caused by studies in which individual organisms within the same tank were treated as independent replicates of that treatment, or where one room was used per treatment.

The proportion of studies that used recommended methods to simulate the carbonate chemistry of a future ocean (91%) was higher than the number of studies that used inappropriate methods (9%). The use of recommended chemical manipulations increased significantly over time, with a sharp increase after 2010 (97% correct after 2010, 78% in 2010 and earlier; Figure 2; $z = 6.5$, $P < 0.01$). The measurement of at least two of pH_T, A_T, DIC, or pCO_2 occurred in 39% of all studies, and there was an increase in this proportion after 2010 (29% pre-2010, 45% post-2010: Figure 2; $z = 3.27$, $P < 0.01$).

Discussion

This analysis identified that the most laboratory manipulation experiments in ocean acidification research used either an inappropriate experimental design and/or data analysis, or did not report these details effectively. Many studies did not report important methods, such as how treatments were created and the number of replicates of each treatment. The tendency for the use of inappropriate experimental design also undermines our confidence in accurately predicting the effects of ocean acidification on the biological responses of marine organisms. There are two contrasting philosophies of interpreting the validity of results with inappropriate design and analysis problems. These are that (i) any research adds to the body of literature, increasing our ability to predict the future effects of ocean acidification. (ii) No value can be taken from such studies, as potential artefacts could alter results, and there is no way of ruling such artefacts out often. As per Hulbert's statement in the seminal paper (1984), "despite errors of design or statics" these papers "nevertheless contain useful information". However, in cases where there is no way to distinguish the effects of potential experimental artefacts or gradients in other factors from the effects of the prescribed treatments (i.e. treatments are completely confounded with time, space, or another factor: i.e. one experimental unit per treatment) then caution should be applied in making conclusions from results. For studies where treatments are not completely confounded (i.e. there is more than one true experimental unit per treatment), but inappropriate analysis has occurred, any statistical significance of the results that are displayed (or estimates of effects sizes or something similar) could still overestimate the true effect size or level of significance. For some studies, re-analysis of the data using the correct statistical framework would facilitate a more accurate assessment of the exact magnitude of biological responses to ocean acidification. However, in some cases, there is no way to distinguish the effects of potential experimental artefacts, gradients in other factors, nor variation in time and space, from the effects of the prescribed treatments. Re-analysis of some of the current literature, potentially using the existing database on ocean acidification research (Nisumaa *et al.*, 2010), would greatly aid in assessing the real biological responses of ocean acidification on marine life and the extent to which ocean acidification research could be biased by non-significant trends.

Confusion regarding what constitutes an experimental unit is evident in ocean acidification research. This is demonstrated by a large proportion of studies that either treated the responses of individuals (i.e. the evaluation unit defined by Hurlbert and White, 1993; Hurlbert, 2009 in terms of statistical analysis) to treatments as experimental units, when multiple individuals were in each tank, or used tank designs where all experimental tanks of one treatment are more interconnected to each other than experimental tanks of other treatments (181 studies total). We consider that this information has not been adequately presented in the context of ocean acidification research previously, which lies at the core of the problem. Therefore, a series of potential solutions is presented to the problems highlighted above for ocean acidification research in the following section.

Advice for performing manipulation experiments: before the ocean acidification manipulation system is designed

In experimental systems that are designed to test the effects of ocean acidification on biological responses of any organism (from bacteria to fish), a number of independent experimental units are needed. It is best practice to have as many independent experimental units as possible to safeguard against confusing non-treatment effects with treatment effects. Not every experimental tank used needs its own storage, mixing, and header tank, as on average all experimental units in one treatment must be equally exposed to the same conditions as all experimental units in another treatment, barring the actual treatment under examination (Hurlbert, 2013a). However, not every experimental tank needs to be completely independent from each other, i.e. every tank does not need an independent source of seawater running from the sea through to individual storage tanks, then to individual mixing tanks, then to individual header tanks, then to the experimental tank. That is why it is equally valid to (i) mix the seawater of each individual experimental unit in one mixing tank per unit (Figure 3a and b), (ii) manipulate the seawater of all experimental units of multiple treatments in one mixing tank (Figure 3c and d), or (iii) mix the seawater for every treatment in multiple mixing tanks (Figure 3e) (i.e. in a randomized block design) as long as this can be taken into account during the statistical analysis (Hurlbert, 1984). The last example (Figure 3e) is a block design whereby an equal number of experimental units from each treatment (but not all experimental units within the experiment) would have their conditions manipulated in the same mixing tank, ensuring no interdependence replicates of the same treatment between blocks (Figures 1 and 3e). The same principals apply to seawater stored for each treatment, where it is equally valid to use individual storage tanks for each experimental unit (Figure 3b), to store seawater for every experimental unit in one storage tank (Figure 3c and d), or equal numbers of experimental units of every treatment in a block design (Figure 3e). However, it

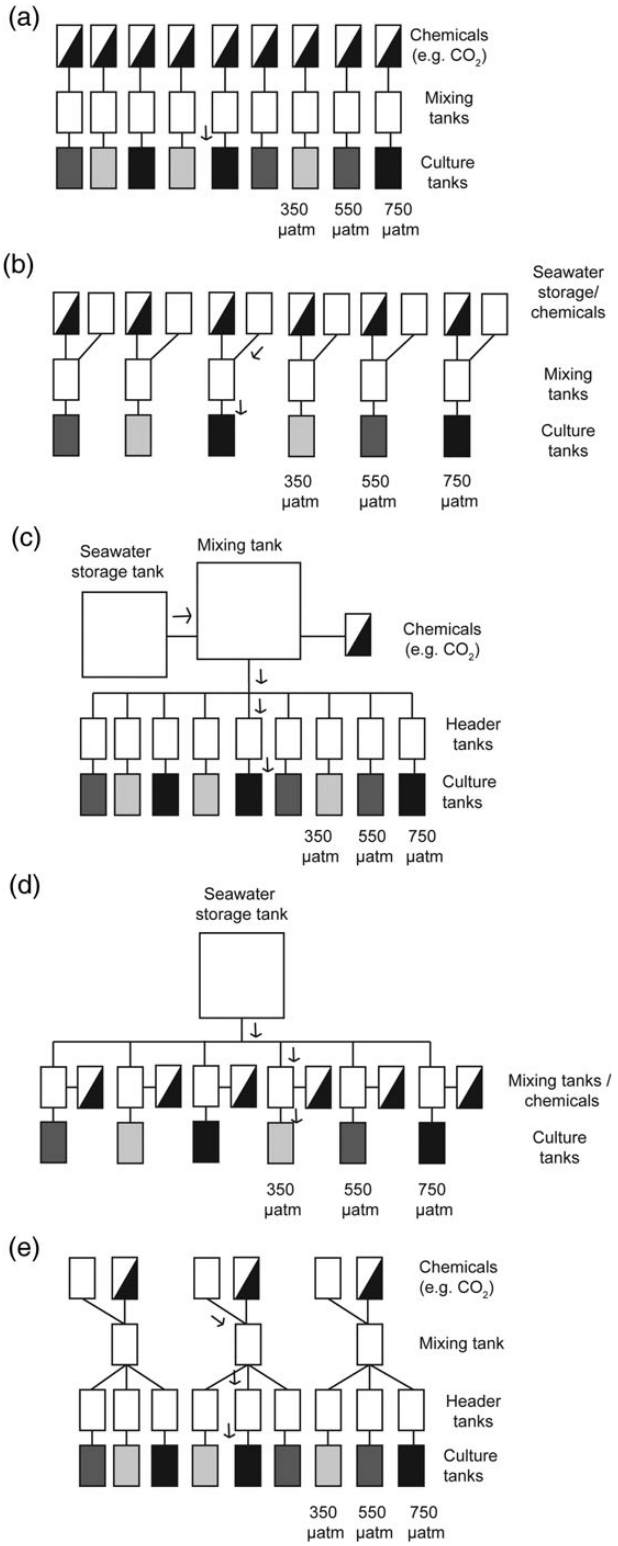

Figure 3. Five tank array types that could be used in ocean acidification research to create treatments with independent experimental units. White tanks indicate space where seawater could be stored other than in experimental tanks. Different coloured experimental tanks correspond to different CO_2 treatments. Half black and white boxes indicate sources of chemicals. See the text for a discussion.

is not valid to mix the seawater of all of one treatment's experimental units in one mixing tank and all of another treatment's experimental units in another mixing tank, and nor is it valid to segregate seawater of all of one treatment from another in one tank during the mixing or storage procedure (i.e. Figure 1, B-4).

The simplest way to manipulate seawater carbonate chemistry to create experimental units that are as independent from each other as they are from other treatment replicated involves at least two different systems of tanks. Several possible methods exist (Figure 3). For each example, a published CO_2-manipulation/housing system that can easily be replicated is referred to. Each of these designs either already fulfil these criteria or are so close to that they could be easily modified to do so by adding one tank or randomly interspersing tanks. The simplest method is to have one mixing tank for each experimental unit/experimental tank (i.e. multiple mixing tanks), so that chemicals are added at the beginning of the experiment from a mixing tank (Figure 3a). Alternatively, this mixing tank can be removed and chemicals added directly to the seawater before cultured organisms are added, or pre-mixed CO_2 at the appropriate level could be added in a way that does not interfere with the study organism. This design is appropriate for experiments that are short term in nature, wherein the organisms do not biologically modify seawater carbonate chemistry significantly over the duration of the experiment (e.g. as in Byrne *et al.*, 2009), or where their biological modification does not affect the study's hypotheses (Schulz *et al.*, 2013).

The second simplest method is to add one storage or header tank per experimental tank to the design (Figure 3b). This is appropriate for longer experiments that could incorporate flow-through seawater and controlled CO_2 conditions (e.g. Munday *et al.*, 2009; Russell *et al.*, 2009). Three more complex methods involve more complicated designs. A third method is to manipulate a source of seawater in one mixing tank for all the experimental units of all treatments (i.e. one mixing tank total) then, after manipulation, the seawater is transferred into separate experimental tanks for each experimental unit, or it is transferred into one header tank then subsequently down into the culture tank in a three-tiered system (Figure 3c). In this example, mixing takes place for one experimental unit at a time, not all of one treatment at a time, and it is important that the mixing order of treatments is randomly interspersed. This is appropriate for an experimental system that delivers seawater semi-continuously using an automated computer-controlled set-up to deliver seawater to the header tanks after measurement and modification (McGraw *et al.*, 2010; Cornwall *et al.*, 2013), such as when a spectrophotometer is used to measure pH and control pH, limiting the number of measurement devices. Another example is if seawater is manually mixed in one tank for all treatments (i.e. one tank total) then transferred into header or experimental tanks. Another modification of this practice is if seawater entering the CO_2-manipulation system is not chemically appropriate for use as controls (i.e. pH is too low or high) and chemicals need to be added before then going to another storage or mixing tank. The fourth design is similar to Figure 3c, where only one storage tank is used, but each culture tank has its own mixing tank (Figure 3d). This design could be used where CO_2 gas is programmed to flow-through the mixing tank at a set rate, or continuously using pre-mixed gases (Findlay *et al.*, 2008; Fangue *et al.*, 2010). The last and most complex design is a block design where one mixing tank manipulates the seawater for one experimental unit for all treatments, although data collected

using this approach needs to incorporate a blocking factor in the analysis (Figure 3e). This fifth design is appropriate if faster flow rates of seawater are used than the design of Figure 3c could permit, for example if multiple spectrophotometers are used in conjunction with a computer-controlled system to measure and control incoming seawater pH. In all these examples, the experimental tank may contain multiple individuals, and analyses can be done using either a simple ANOVA applied to tank means or with a nested ANOVA applied to values for the responses of individual organisms. The decision to include multiple or single individuals in one tank depends on the hypotheses of the study, logistic constraints, and ecological realism.

In addition to independent replication, there must also be a large enough number of experimental units for sufficient statistical power to accurately detect differences of interest between treatments, and these replicate experimental units must be randomly interspersed when appropriate. The magnitude of the biological response to OA treatments will be species-specific, and within a species the measured response metrics (e.g. growth, photosynthesis, respiration, or calcification rates) will vary among individuals, independent of any measurement error. It is beyond the scope of this review to recommend the use of statistics that employ hypothesis testing vs. other types of statistical approaches (for an excellent discussion, see Ellison *et al.*, 2014) nor to recommend the exact number of experimental units needed for every study; but regardless, the number of experimental units used per treatment should be >2.5, which was the mean in 2014.

Randomization is relatively simple to achieve, given that treatments must all equally on average experience the same conditions other than the treatment ascribed to them, and there are three choices according to Hurlbert (1984): completely randomized (A-1), randomized block (A-2), and systematic (A-3; which is not technically random but still pragmatically achieves an acceptable design). Regression style designs examining biological responses to large numbers of CO_2 treatments could reduce the need for a larger number of experimental units per treatment relative to traditional factorial analysis. Regression style analyses still, however, require a enough experimental units per treatment so that the variation due to differences among individual experimental units is lower than that required to observe any effects of CO_2 treatments. Regression designs also require randomization of treatments. To have spatial or temporal segregation will confound the effects of treatment and location/time. To use an extreme example for the sake of clarity, consider an experiment in which the experimental tanks are placed on a bench, with CO_2 treatments placed from left to right: 200, 400, 600, 800, 1000 ppm pCO_2. In this example, it is impossible to use statistics to differentiate the effects of bench location from the effects of CO_2 treatment. To correctly replicate in this experiment, the experimental tanks would be placed in a random order on the bench.

After the CO_2-manipulation system is designed as B-1 to B-4

In an experimental system where tanks from one treatment are either spatially segregated or are more interdependent with each other than with replicates of other treatments, the easiest way to control for artefacts is to repeat the experiment many times. Thereafter, the effect of tank identity can be incorporated as a random factor. This approach is difficult if there is only one experimental or header tank per treatment, as the effects of time will be difficult to separate from the effects of tank unless a large number of experimental runs are conducted. In some instances, these problems can be overcome during statistical analysis if the treatments are not completely confounded, i.e. if there are multiple header tanks per treatment and not just one, as was the case in 20 studies examined. Also, multiple experimental tanks per header tank could be employed if there are three or more independent header tanks per treatment, but only as long as header tank is treated as the experimental unit. Likewise, multiple individuals could be housed in the same experimental tank using the same principles, where tank is still treated as the experimental unit. This could be a desirable design where variation in individual responses to a treatment is known to be high, but logistical restraints prevent the use of more tanks.

After the experiment is conducted

An experimental culture system can be designed correctly but still result in an inappropriate manipulation experiment when: (i) measurements are interdependent through time (i.e. multiple measurements on one experimental unit over time are treated as independent measurements) and (ii) multiple measurements are made on the same experimental unit and are treated as independent measurements at the same time point. Similarly, treating the responses of multiple individuals within one experimental unit as multiple experimental units is inappropriate (classed as sacrificial pseudoreplication by Hurlbert, 1984; Hurlbert and White, 1993).

Conversely, an inappropriate design can be employed, but variation from non-treatment effects can be controlled for somewhat by treating these as random factors in the statistical analysis, not treating all experimental units at different time points as independent experimental units (classed as temporal pseudoreplication by Hurlbert, 1984; Hurlbert and White, 1993). Running an analysis that determines whether there is an effect of time or space (such as tank identity) on responses then removing this factor from the analysis if there is no effect is also inappropriate according to Hurlbert (1984), i.e. sacrificial pseudoreplication (Hurlbert, 1984; Hurlbert and White, 1993). For time or tank identity, we recommend that they could be incorporated in the analysis as a random factor, not a fixed factor, to incorporate some source of variance from tank identity or time, rather than to increase the degrees of freedom. Another approach is to determine what the mean change in a measured parameter is over time, and treat that mean as the experimental unit, or to take the mean of the evaluation unit's response over all repeated measures. Alternatively, repeated measures within one experimental unit can be removed by randomly removing data points until only one measurement per experimental unit remains.

During the review process

Fifty-seven per cent of studies examined did not report in the methods sufficient detail to assess the experimental design used. Details, such as the layout and numbers of tanks, are very important in allowing scientists to compare results of different studies and Table 1 provides a checklist of essential details that need to be reported in published ocean acidification studies with respect to replication and randomization. It is useful for publications to include a diagram of the experimental set-up, at least the first time an experimental facility is described (e.g. see Figure 1 in Asnaghi *et al.*, 2013).

Some studies did not report degrees of freedom adequately. If the degrees of freedom reported are larger than the total number of experimental units, then samples have been pooled in an inappropriate way. This could lead to reduced *P*-values (Hurlbert, 2013b). This could result in inaccurate findings, and because published experiments are rarely repeated, such spurious or inaccurate results and the conclusions drawn might direct future research

Table 1. A checklist of replication details that need to be included in published research investigating the effects of ocean acidification in a laboratory setting.

Can the following details be clearly determined from the manuscript?
The number of independent experimental units used
The randomization of treatments in space and time (when appropriate)
The configuration of mixing, header, and experimental tanks (when appropriate)
The model construction of the statistical analysis
The degrees of freedom in the statistical analysis
Location where the seawater was manipulated
Location where the measurements of seawater chemistry were taken
Chemicals used to manipulate seawater CO_2 concentrations
The measurement techniques used to characterize seawater carbonate chemistry ($\geq$2 of pH_T, A_T, DIC, and pCO_2; plus temperature and salinity)

erroneously. An accumulation of spurious results may counteract the ability to accurately predict the effects of ocean acidification or to detect any true trends in its potential impacts.

When assessing whether a study has used appropriate methods, it would be pragmatic to determine what was the likelihood that non-treatment effects could confound treatment effects, rather than grouping studies into "pseudoreplicated" vs. "non-pseudoreplicated" categories. Also, studies employing inappropriate statistical and design frameworks may still have important messages to convey. If inappropriate methods were inadvertently used, such that replicates of a treatment are not independent, then data can be reanalysed as described above. However, if this is not possible (i.e. if the resulting degrees of freedom are too small) then no discussion of differences should be made, and only trends can be described with strong caveats mentioned. It is equally important that studies are not miss-categorized by reviewers as using inappropriate methods; by applying the principles discussed here, and in the papers referenced herein, this could be easier for reviewers to achieve.

Conclusions

This study highlights the need for increased attention to the design, performance, analysis, and reporting of experimental procedures used during ocean acidification manipulation experiments. It is encouraging that most studies follow the "Guide to Best Practices for Ocean Acidification Research and Data Reporting (Gattuso *et al.*, 2010)" for manipulating seawater carbonate chemistry, and that this is also improving for measurements of seawater carbonate chemistry; we recommend the same rigour in the use of experimental units in experimental design is followed to avoid artefacts related to inappropriate replication and randomization. Ocean acidification is a relatively new research field, and if stricter guidelines are followed in future research, then greater inferences can be taken from studies, enhancing our ability to accurately predict how changes in ocean carbonate chemistry will affect species and ecosystems in the future. This paper deals specifically with designing manipulation experiments that minimize the effects of chance events or pre-existing gradients in other factors. However, other methodological considerations also need to be taken into account in ocean acidification research to improve the information that can be taken from such studies. Future research would also be improved if the duration of experiments increased, or if they incorporated gradual increases in CO_2 rather than rapid exposure (Kamenos *et al.*, 2013; Munday *et al.*, 2013; Munguia and Alenius, 2013; Gaylord *et al.*, 2015). Future studies should also focus on improving the environmental realism employed in laboratory experiments, for example by attempting to simulate environmentally realistic water motion, and diel cycles in light and pH (when appropriate). Ideally, this added realism would not be sacrificed in an attempt to increase the numbers of experimental units, which further highlights the complexities faced by future ocean acidification research.

A large number of studies have found that ocean acidification will likely negatively impact the calcification or growth of calcifying invertebrates, coccolithophores, calcifying macroalgae, and corals (Riebesell *et al.*, 2000; Gazeau *et al.*, 2007; Anthony *et al.*, 2008; Wood *et al.*, 2008; Byrne *et al.*, 2010), and influence the behavioural traits of invertebrates and fish (Munday *et al.*, 2009; Appelhans *et al.*, 2012; Nilsson *et al.*, 2012). Subsequent shifts in ecosystem structure and function are likely to occur due to the direct biological effects on many ecologically important species (Hall-Spencer *et al.*, 2008; Fabricius *et al.*, 2011; Kroeker *et al.*, 2013a). Yet, meta-analyses and reviews of published studies reveal that the effects of ocean acidification are variable, sometimes contrary to the expected outcome, and that they are species and context dependent, interacting with changes in other environmental stressors in both synergistic and additive ways (Fabry *et al.*, 2008; Pörtner, 2008; Hofmann *et al.*, 2010; Kroeker *et al.*, 2013b). Predicting accurate effects of ocean acidification on marine organisms will need new meta-analyses that can accurately re-analyse published results. Examining the extent to which the inappropriate assignment of experimental units has influenced our current knowledge base, and other potential sources of biases (such as the use of HCl without DIC additions), will involve meta-analyses that incorporate the true variance in responses of species to ocean acidification from all previously published studies. Completing that task then maintaining standards in the publication of ocean acidification research will enable stronger and more accurate predictions regarding the state of our future marine ecosystems. Without the maintenance of standards in ocean acidification research, there is a risk of collecting an ever increasing number of uninterpretable results (Moran, 2014) and misinterpreting likely future outcomes. However, caution must be applied when classifying research as having used inappropriate methods when details are incomplete (as in 34% of studies here), as they could be equally as sound as those with complete methods (Munday *et al.*, 2014). We strongly recommend that future research report sufficient details so that their methods are clear and repeatable. Our ability to more accurately predict the magnitude and context of the effects of ocean acidification rely on continuing to improve best practices in both experimental design, and the measurement and manipulation of seawater carbonate chemistry.

Supplementary material

Supplementary Material is available at *ICESJMS* online.

Acknowledgments

We acknowledge J.-P. Gattuso, S. Hurlbert, D. Moran, B. Russell, M. Wahl, and one anonymous reviewer for feedback on this manuscript, and C.D. Hepburn and C.M. McGraw for discussions over the years on associated topics.

References

Anthony, K. R. N., Kline, D. I., Diaz-Pulido, G., Dove, S., and Hoegh-Guldberg, O. 2008. Ocean acidification causes bleaching and productivity loss in coral reef builders. Proceedings of the National Academy of Sciences, 105: 17442–17446.

Appelhans, Y. S., Thomsen, J., Pansch, C., Melzner, F., and Wahl, M. 2012. Sour times: seawater acidification effects on growth, feeding behaviour and acid-base status of *Asterias rubens* and *Carcinus maenas*. Marine Ecology Progress Series, 459: 85–97.

Asnaghi, V., Chiantore, M., Mangialajo, L., Gazeau, F., Francour, P., Alliouane, S., and Gattuso, J. P. 2013. Cascading effects of ocean acidification in a rocky subtidal community. PLoS ONE, 8: e61978.

Barry, J. P., Hall-Spencer, J. M., and Tyrrell, T. 2010. In situ perturbation experiments: natural venting sites, spatial/temporal gradients in ocean pH, manipulative in situ $p(CO_2)$ perturbations. *In* Guide to Best Practices for Ocean Acidification Research and Data Reporting. Ed. by U. Riebesell, V. J. Fabry, L. Hansson, and J. P. Gattuso. Publications Office of the European Union, Luxembourg.

Boyd, P. W. 2011. Beyond ocean acidification. Nature Geoscience, 4: 273–274.

Byrne, M., Ho, M., Selvakumaraswamy, P., Nguyen, H. D., Dworjanyn, S. A., and Davis, A. R. 2009. Temperature, but not pH, compromises sea urchin fertilization and early development under near-future climate change scenarios. Proceedings of the Royal Society B: Biological Sciences, 276: 1883–1888.

Byrne, M., Ho, M., Wong, E., Soars, N. A., Selvakumaraswamy, P., Shepard-Brennand, H., Dworjanyn, S. A., *et al.* 2010. Unshelled abalone and corrupted urchins: development of marine calcifiers in a changing ocean. Proceedings of the Royal Society B: Biological Sciences, 276: 1883–1888.

Ciais, P., Sabine, C. L., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., *et al.* 2013. Carbon and other biogeochemical cycles. *In* Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Ed. by T. F. Stocker, D. Qin, G-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley. Cambridge University Press, Cambridge, UK and New York, NY, USA.

Cornwall, C. E., Hepburn, C. D., McGraw, C. M., Currie, K. I., Pilditch, C. A., Hunter, K. A., Boyd, P. W., *et al.* 2013. Diurnal fluctuations in seawater pH influence the response of a calcifying macroalga to ocean acidification. Proceedings of the Royal Society B: Biological Sciences, 280: doi:10.1098/rspb.2013.2201.

Cox, D. R. 1958. Planning of Experiments. Academic Press, New York, NY.

Dickson, A. G. 2010. The carbon dioxide system in seawater: equilibrium chemistry and measurements. *In* Guide to Best Practises for Ocean Acidification Research and Data Reporting. Ed. by U. Riebesell, V. J. Fabry, L. Hansson, and J. P. Gattuso. Publications Office of the European Union, Luxembourg.

Dickson, A. G., Sabine, C. L., and Christian, J. R. 2007. Guide to best practices for Ocean CO_2 measurements, North Pacific Marine Science Organization.

Ellison, A. M., Gotelli, N. J., Inouye, B. D., and Strong, D. R. 2014. *P* values, hypothesis testing, and model selection: it's déjà vu all over again. Ecology, 95: 609–610.

Fabricius, K., Langdon, C., Uthicke, S., Humphrey, C., Noonan, S., De'ath, G., Okazaki, R., *et al.* 2011. Losers and winners in coral reefs acclimatized to elevated carbon dioxide concentrations. Nature Climate Change, 1: 165–169.

Fabry, V. J., Seibel, B. A., Feely, R. A., and Orr, J. C. 2008. Impacts of ocean acidification on marine fauna and ecosystem processes. ICES Journal of Marine Science, 65: 414–432.

Fangue, N. A., O'Donnell, M. J., Sewell, M. A., Matson, P. G., MacPherson, A. C., and Hofmann, G. E. 2010. A laboratory-based, experimental system for the study of ocean acidification effects on marine invertebrate larvae. Limnology and Oceanography: Methods, 8: 441–452.

Feely, R. A., Sabine, C. L., Lee, K., Berelson, W., Kleypas, J. A., Fabry, V. J., and Millero, F. J. 2004. Impact of anthropogenic CO_2 on the $CaCO_3$ system in the oceans. Science, 305: 362–366.

Findlay, H. S., Kendall, K. A., Spicer, J. I., Turley, C., and Widdicombe, S. 2008. Novel microcosm system for investigating the effects of elevated carbon dioxide and temperature on intertidal organisms. Aquatic Biology, 3: 51–62.

Gattuso, J. P., Gao, K., Lee, K., Rost, B., and Schulz, K. G. 2010. Approaches and tools to manipulate the carbonate chemistry. *In* Guide to best practices for ocean acidification research and data reporting, pp. 41–51. Ed. by U. Riebesell, V. J. Fabry, L. Hansson, and J. P. Gattuso. Publications Office of the European Union, Luxembourg.

Gattuso, J -P., Kirkwood, W., Barry, J. P., Cox, E., Gazeau, F., Hansson, L., Hendriks, I., *et al.* 2014. Free-ocean CO_2 enrichment (FOCE) systems: present status and future developments. Biogeosciences, 11: 4057–4075.

Gattuso, J. P., Mach, K. J., and Morgan, G. 2012. Ocean acidification and its impacts: an expert survey. Climate Change, 117: 725–738.

Gaylord, B., Kroeker, K. J., Sunday, J. M., Anderson, K. M., Barry, J. P., Brown, N. E., Connell, S. D., *et al.* 2015. Ocean acidification through the lens of ecological theory. Ecology, 96: 3–15.

Gazeau, F., Quiblier, C., Jansen, J. M., Gattuso, J. P., Middelburg, J. J., and Heip, C. H. R. 2007. Impact of elevated CO_2 on shellfish calcification. Geophysical Research Letters, 34: doi: 10.1029/2006GL028554.

Hall-Spencer, J. M., Rodolfo-Metalpa, R., Martin, S., Ransome, E., Fine, M., Turner, S. M., Rowley, S. J., *et al.* 2008. Volcanic carbon dioxide vents show ecosystem effects of ocean acidification. Nature, 454: 96–99.

Havenhand, J., Dupont, S., and Quinn, G. P. 2010. Designing ocean acidification experiments to maximise inference. *In* Guide to best practices for ocean acidification research and data reporting, p. 260. Ed. by U. Riebesell, V. J. Fabry, L. Hansson, and J-P. Gattuso. Publications Office of the European Union, Luxembourg.

Hofmann, G. E., Barry, J. P., Edmunds, P. J., Gates, R. D., Hutchins, D. A., Klinger, T., and Sewell, M. A. 2010. The effect of ocean acidification on calcifying organisms in marine ecosystems: an organism-to-ecosystem perspective. Annual Reviews in Ecology, Evolution, and Systematics, 41: 127–147.

Hurd, C. L., Hepburn, C. D., Currie, K. I., Raven, J. A., and Hunter, K. A. 2009. Testing methods of ocean acidification on algal metabolism: consideration for experimental designs. Journal of Phycology, 45: 1236–1251.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs, 54: 187–211.

Hurlbert, S. H. 2009. The ancient black art and transdisciplinary extent of pseudoreplication. Journal of Comparative Psychology, 123: 434–443.

Hurlbert, S. H. 2013a. Affirmation of the classical terminology for experimental design via a critque of Casella's *Statistical Design*. Agronomony Journal, 105: 412–418.

Hurlbert, S. H. 2013b. Pseudofactorialism, response structures and collective responsibility. Austral Ecology, 38: 646–663.

Hurlbert, S. H., and White, M. D. 1993. Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. Bulletin of Marine Science, 53: 128–153.

Kamenos, N. A., Burdett, H. L., Aloisio, E., Findlay, H. S., Martin, S., Longbone, C., Dunn, J., *et al.* 2013. Coralline algal structure is more sensitive to rate, rather than the magnitude, of ocean acidification. Global Change Biology, 19: 3621–3628.

Kozlov, M., and Hurlbert, S. H. 2006. Pseudoreplication, chatter, and the international nature of science. Journal of Fundamental Biology, 67: 128–135.

Kroeker, K. J., Gambi, M. C., and Micheli, F. 2013a. Community dynamics and ecosystem simplification in a high-CO_2 ocean. Proceedings of the National Academy of Sciences, 110: 12721–12726.

Kroeker, K. J., Kordas, R. L., Crim, R. N., Hendriks, I. E., Ramajo, L., Singh, G. G., Duarte, C. M., *et al.* 2013b. Impacts of ocean

acidification on marine organisms: quantifying sensitivities and interaction with warming. Global Change Biology, 19: 1884–1896.

McGraw, C. M., Cornwall, C. E., Reid, M. R., Currie, K. I., Hepburn, C. D., Boyd, P., Hurd, C. L., *et al.* 2010. An automated pH-controlled culture system for laboratory-based ocean acidification experiments. Limnology and Oceanography: Methods, 8: 686–694.

Mead, R. 1988. The Design of Experiments, Cambridge University Press, Cambridge. 620 pp.

Moran, D. 2014. The importance of accurate CO_2 dosing and measurement in ocean acidification studies. Journal of Experimental Biology, 217: 1827–1829.

Munday, P. L., Dixson, D. L., Donelson, J. M., Jones, G. P., Pratchett, M. S., Devitsina, G. V., and Døving, K. B. 2009. Ocean acidification impairs olfactory discrimination and homing ability of a marine fish. Proceedings of the National Academy of Sciences, 106: 1848–1852.

Munday, P. L., Warner, R. R., Monro, K., Pandolfi, J. M., and Marshall, D. J. 2013. Predicting evolutionary responses to climate change in the sea. Ecology Letters, 16: 1488–1500.

Munday, P. L., Watson, S -A., Chung, W -S., Marshall, N. J., and Nilsson, G. E. 2014. Response to 'The importance of accurate CO_2 dosing and measurement in ocean acidification studies'. Journal of Experimental Biology, 217: 1827–1829.

Munguia, P., and Alenius, B. 2013. The role of preconditoning in ocean acidification experiments: a test with the intertidal isopod *Paradella dianae*. Marine and Freshwater Behaviour and Physiology, 46: 33–44.

Nilsson, G. E., Dixson, D. L., Domenici, P., McCormick, M. I., Sørensen, C., Watson, S. A., and Munday, P. L. 2012. Near-future carbon dioxide levels alter fish behaviour by interfering with neurotransmitter function. Nature Climate Change, 2: 201–204.

Nisumaa, A -M., Gattuso, J -P., Bellerby, R. G. J., Delille, B., Geider, R. J., Middelburg, J. J., Orr, J. C., *et al.* 2010. EPOCA/EUR-OCEANS data compilation on the biological and biogeochemical responses to ocean acidification. Earth Systems Science Data, 2: 167–175.

Pörtner, H. O. 2008. Ecosystem effects of ocean acidification in times of ocean warming: a physiologist's view. Marine Ecology Progress Series, 373: 203–217.

Riebesell, U., and Gattuso, J -P. 2015. Lessons learned from ocean acidification research. Nature Climate Change, 5: 12–14.

Riebesell, U., Fabry, V. J., Hansson, L., and Gattuso, J. P. 2010a. Guide to Best Practices for Ocean Acidification Research and Data Reporting. Publications Office of the European Union, Luxembourg. 260 pp.

Riebesell, U., Lee, K., and Nejstgaard, J. C. 2010b. Pelagic mesocosms. *In* Guide to Best Practises for Ocean Acidification Research and Data Reporting, pp. 95–112. Ed. by U. Riebesell, V. J. Fabry, L. Hansson, and J. P. Gattuso. Publications Office of the European Union, Luxembourg.

Riebesell, U., Zondervan, I., Rost, B., Tortell, P. D., Zeebe, R. E., and Morel, F. M. M. 2000. Reduced calcification of marine phytoplakton in response to increased atmospheric CO_2. Nature, 407: 364–367.

Rost, B., Zondervan, I., and Wolf-Gladrow, D. 2008. Sensitivity of phytoplankton to future changes in ocean carbonate chemistry: current knowledge, contradictions and research directions. Marine Ecology Progress Series, 373: 227–237.

Russell, B. D., Thompson, J. I., Falkenberg, L. J., and Connell, S. D. 2009. Synergistic effects of climate change and local stressors: CO_2 and nutrient-driven change in subtidal rocky habitats. Global Change Biology, 15: 2153–2162.

Schulz, K. G., Bellerby, R. G. J., Brussaard, C. P. D., Büdenbender, J., Czerny, J., Engel, A., Fischer, M., *et al.* 2013. Temporal biomass dynamics of an Arctic plankton bloom in response to increasing levels of atmospheric carbon dioxide. Biogeosciences, 10: 161–180.

Wernberg, T., Smale, D. A., and Thomsen, M. S. 2012. A decade of climate change experiments on marine organisms: procedures, patterns and problems. Global Change Biology, 18: 1491–1498.

Wood, H. L., Spicer, J. I., and Widdicombe, S. 2008. Ocean acidification may increase calcification rates, but at a cost. Proceedings of the Royal Society B: Biological Sciences, 275: 1767–1773.

Handling editor: Howard Browman