An extreme "decline effect" in ocean acidification effects on fish behaviour

Jeff C. Clements[a]*, Josefin Sundin[a,b], Timothy D. Clark[c], Fredrik Jutfelt[a]

[a] Department of Biology, Norwegian University of Science and Technology, Høgskoleringen 5, NO-7491 Trondheim, Norway

[b] Department of Aquatic Resources, Swedish University of Agricultural Sciences, Drottningholm, Sweden.

[c] School of Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia

*Correspondence: Jeff C. Clements, PhD

Department of Biology, Norwegian University of Science and Technology, Høgskoleringen 5, 7491 Trondheim, Norway

Current address: Fisheries and Oceans Canada, Gulf Region, 343 Université Avenue, Moncton, NB E1C 9B6, Canada

Email:jeffery.clements@dfo-mpo.gc.ca, jefferycclements@gmail.com

Abstract

Ocean acidification – decreasing oceanic pH resulting from the uptake of excess atmospheric CO_2 – has the potential to affect marine life in the future. Among the possible consequences, a series of studies on coral reef fishes suggested that the direct effects of acidification on fish behaviour may be extreme and have broad ecological ramifications. Recent studies documenting a lack of effect of experimental ocean acidification on fish behaviour, however, call this prediction into question. Here, we explore the consistency and robustness of scientific evidence over the past decade regarding direct effects of ocean acidification on fish behaviour by testing for a "decline effect". Using a meta-analysis, we provide quantitative evidence that the research to date on this topic is characterized by a decline effect, where large initial effects have all but disappeared over a decade. The decline effect in this field cannot be explained by three likely biological explanations, including increasing proportions of studies examining (1) cold-water species, (2) non-olfactory associated behaviours, and (3) non-larval life stages. Furthermore, the vast majority of studies with large effect sizes in this field tend to be characterized by low sample sizes, yet are published in high impact journals and have a disproportionate influence on the field in terms of citations. We contend that ocean acidification has a negligible direct impact on fish behaviour, and we advocate for improved approaches to minimize the potential for a decline effect in future avenues of research.

Keywords: animal behaviour | bias | carbon dioxide | climate change | global change biology | scientific process

Introduction

Publications for new hypotheses or ground-breaking scientific discoveries are often followed by attempts to replicate and build upon the initial research. In many instances, however, follow-up studies fail to replicate initial effects, and/or report smaller effect sizes. The tendency for initial scientific findings—which can show strong effects with large effect sizes—to lose strength over time is referred to as the 'decline effect' [1]. This phenomenon was first described in the 1930s, and has since been documented in a range of scientific disciplines [1], including ecology and evolution [2,3]. It captures the concept of initial reports with large effect sizes that overestimate reality. In such instances, the early, large effect sizes are the key problem, not the subsequent decline. The decline effect could therefore equally be referred to as the 'early inflation effect'. Nonetheless, this process can be problematic by delaying accurate scientific understanding of a given phenomenon, and can have applied ramifications, for example, to policy making [4].

Over the past 15 years, biologists have documented substantial impacts of ocean acidification on marine biota [5]. With more than 300 papers published per year from 2006 to 2015, the exponential growth of ocean acidification studies represents one of the fastest expanding topics in the marine sciences [6], and underscores the perceived risk of ocean acidification to ecosystem resilience. In recent years, however, there has been increasing skepticism and uncertainty around the severity of ocean acidification effects on marine organisms [6,7].

Some of the most striking effects of ocean acidification are those concerning fish behaviour, whereby a series of sentinel papers in 2009 and 2010 published in prestigious journals reported large effects of laboratory-simulated ocean acidification [8–10]. Since their publication, these papers have remained among the most highly cited regarding acidification effects on fish behaviour. The severe negative impacts and drastic ecological consequences outlined in those studies were highly publicized in some of the world's most prominent media outlets [11–13] and were used to influence policy through a presentation at the White House [14]. Not only were the findings alarming, the extraordinarily clear and strong results left little doubt that the effects were real, and a multimillion-dollar international investment of research funding was initiated to quantify the broader impacts of ocean acidification on a range of behaviours. In recent years, however, an increasing number of papers have reported a lack of ocean acidification effects on fish behaviour, calling into question the reliability of initial reports.

Here, we present a striking example of the decline effect in ecology over the past decade in research on the impact of ocean acidification on fish behaviour. We find that initial effects of acidification on fish behaviour have all but disappeared over the past five years, and present evidence that common biases influence reported effect sizes in this field. Ways to mitigate these biases and reduce the time it takes to reach a "true" effect size, broadly applicable to any scientific field, are discussed.

Results and Discussion

Declining effects

Based on a systematic literature review and meta-analysis ($n = 91$ studies), we found evidence for a decline effect in ocean acidification studies on fish behaviour (Fig 1a, b). Generally, effect sizes in this field have decreased by an order of magnitude over the past decade, from mean effect size magnitudes >5 in 2009-2010 to effect size magnitudes <0.5 after 2015 (Fig. 1a, b; Table S1). Mean effect size magnitude (absolute lnRR) was disproportionately large in early studies, hovered at moderate effect sizes from 2012–2014 and has all but disappeared in recent years (Fig. 1a, b).
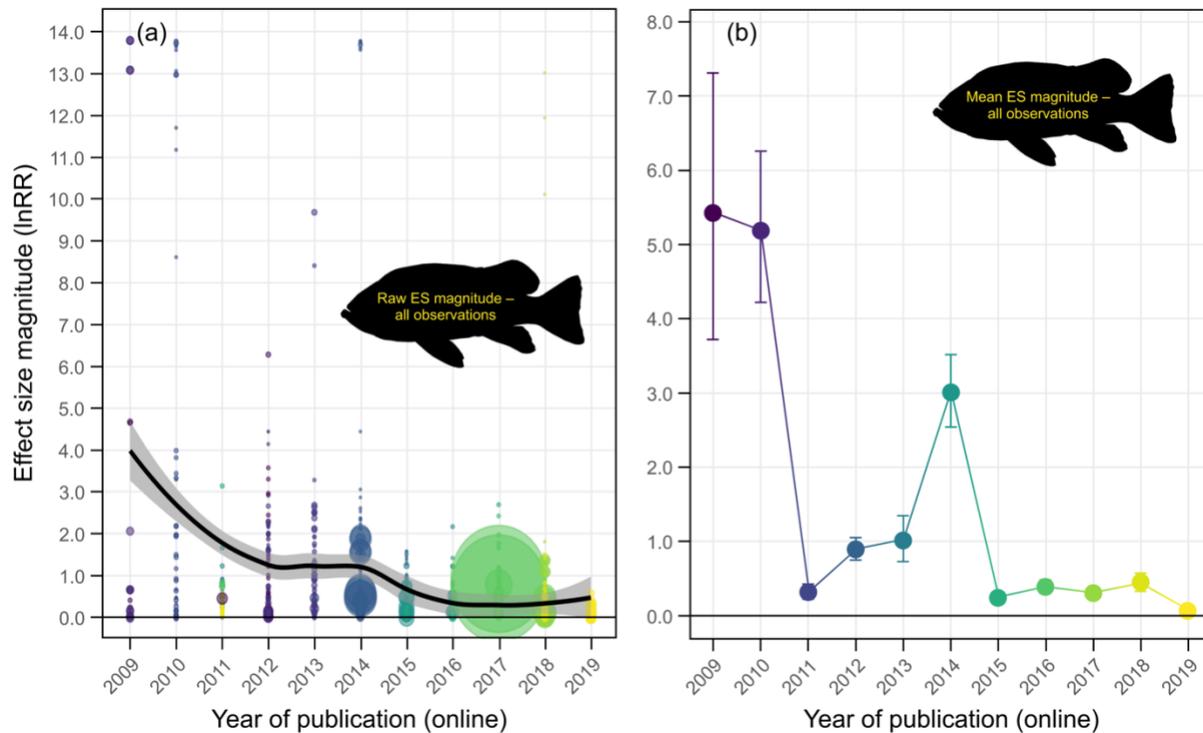
Fig. 1. The decline effect in ocean acidification research on fish behaviour. (a) Trend in raw effect sizes (lnRR) for each experiment in our dataset plotted as a function of year of publication (online) and colour-coded according to study. Data are fit with a Loess curve with 95% confidence bounds. **(b)** Mean effect size magnitude (absolute lnRR magnitude ± upper and lower confidence bounds) for each year of publication (online) in our dataset. Mean effect size magnitudes and confidence bounds were estimated using Bayesian simulations and a folded normal distribution. Note: colours are aesthetic in nature and follow a gradient according to year of publication.

The large effect size magnitudes from early studies on acidification and fish behaviour are not present in the majority of studies in the last five years (Fig. 1b; Table S1). This decline effect could be explained by a number of factors, including biological. For example, cold-water fish in temperate regions experience a higher degree of temporal variability in carbonate chemistry parameters over large spatial areas [15]. Therefore, they may be less sensitive to changes in seawater CO_2 as per the Ocean Variability Hypothesis [16]. As such, if an increasing number of studies on cold-water species over time was responsible for the decline effect, removing cold-water species from the dataset (i.e., only including warm-water species) should result in the decline effect trend disappearing. This was not the case, as the decline effect persisted when only warm-water species were considered (Fig. 2a). In the same vein, the strongest ocean acidification effects on fish behaviour have undoubtedly been reported for chemical cue (herein 'olfactory') responses, and increasing numbers of studies on non-olfactory behaviours could explain the decline effect. If this was true, removing non-olfactory behaviours from the dataset should negate the decline effect trend. Again, this was not the case (Fig. 2b). Finally, early studies of ocean acidification and fish behaviour used larval fish, which are typically considered to be more sensitive to environmental perturbations than juveniles and adults. If a greater proportion of studies used less sensitive life stages through time, then removing those life stages and focusing exclusively on larvae should abolish the decline effect. Once again, this was not the case (Fig. 2c). These analyses show that ocean acidification studies on fish behaviour exhibit a decline effect that is not explainable by three biological processes commonly considered important drivers of acidification effects (Fig. 2a-c; Table S1).
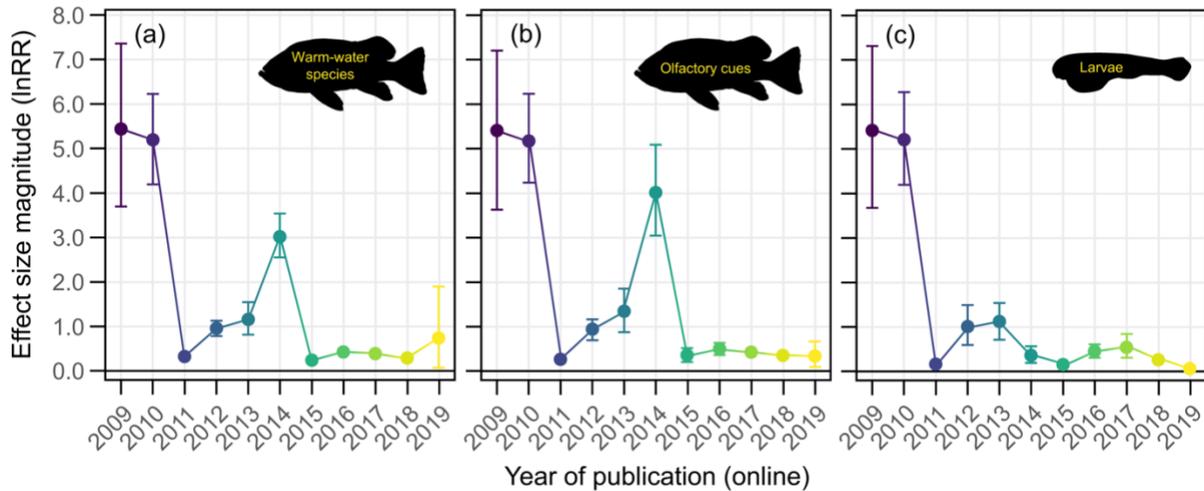
Fig 2. The decline effect cannot be explained by three commonly-considered biological drivers of acidification effects. Mean effect size magnitude (absolute lnRR ± upper and lower confidence bounds) as a function of time for datasets that only included experiments with **(a)** warm-water species, **(b)** olfactory-associated behaviours, and **(c)** larval life stages. Mean effect size magnitudes and confidence bounds were estimated using Bayesian simulations and a folded normal distribution. Note: colours are aesthetic in nature and follow a gradient according to year of publication.

While we were able to test and exclude three biological factors, there are other potential factors that could drive the decline which are not readily testable from our database. For example, while we were able to partially test for the influence of background CO_2 variability by comparing cold- and warm-water species, most studies do not report the actual background CO_2 levels that the experimental animals (and their ancestors) have historically experienced. As such, we are unable to account for the historic CO_2 acclimation conditions of animals used in experiments. The impact of this with respect to the observed decline effect could stem from an increasing proportion of studies using captive bred fish from recirculating aquarium systems with high CO_2 levels, as compared to fish from wild populations experiencing natural CO_2 levels. This is an unlikely explanation for the decline effect, however, given that the earliest studies conducted in 2009–2010 reporting high effect sizes were conducted with both captive-bred and wild-caught fish [7–9, 13]. Furthermore, recent replication attempts of those initial studies using wild-caught fish have failed to replicate the large effect sizes [7]. Nonetheless, we recommend that future studies provide better background CO_2 information for the fish used in their experiments, and use best practices for measuring and reporting carbonate chemistry [15].

Biased behaviour in a maturing field?

It is clear that the ocean acidification field, and indeed science in general, is prone to many biases including methodological and publication biases [6]. The key thing to note is that if science was operating properly from the onset, and early effects of ocean acidification on fish behaviour were true, the relationships presented in Figs. 1 and 2 would be flat lines showing consistent effect sizes over time. It is also evident that the decline effect discovered herein is not explainable by three likely biological culprits (outlined above). Thus, the data presented here provide a textbook example of a new and emerging "hot topic" field likely being prone to biases. Below, we underscore and assess the roles of three potential biases: (1) methodological biases (low samples size), (2) selective publication bias, and (3) citation bias.

Methodological biases. Methodological approaches for individual studies, and biases therein, can contribute to the early inflation of effects. Such biases can come in the form of experimental protocols, the chosen experimental design and sample size, and the analytical/statistical approach employed. Experimenter biases can also contribute to inflated effects.

Experimental designs and protocols can introduce unwanted biases during the experiment whether or not the researchers realise it. For example, experiments with small sample sizes are more prone to statistical

3

errors (i.e., Type I and Type II error) and studies with larger sample sizes should be trusted more than those with smaller sample sizes [18]. While we did not directly test it in our analysis, studies with small sample sizes are also more susceptible to statistical malpractices such as p-hacking and selective exclusion of data that do not conform to a pre-determined experimental outcome, contributing to inflated effects [19]. In our analysis, we found that almost all of the studies with the largest effect size magnitudes had mean sample sizes (per experimental treatment) below 30 fish. Indeed, 87% of the studies (13 of 15 studies) with a mean effect size magnitude >1.0 had a mean sample size below 30 fish (Fig. 3). Likewise, the number of studies reporting an effect size magnitude >0.5 sharply decreased after a mean sample size exceeding 30 fish (Fig. 3). Sample size is of course not the only attribute that describes the quality of a study, but the effects detected here certainly suggest that studies with $n < 30$ fish per treatment may yield spurious effects and should be weighted accordingly.
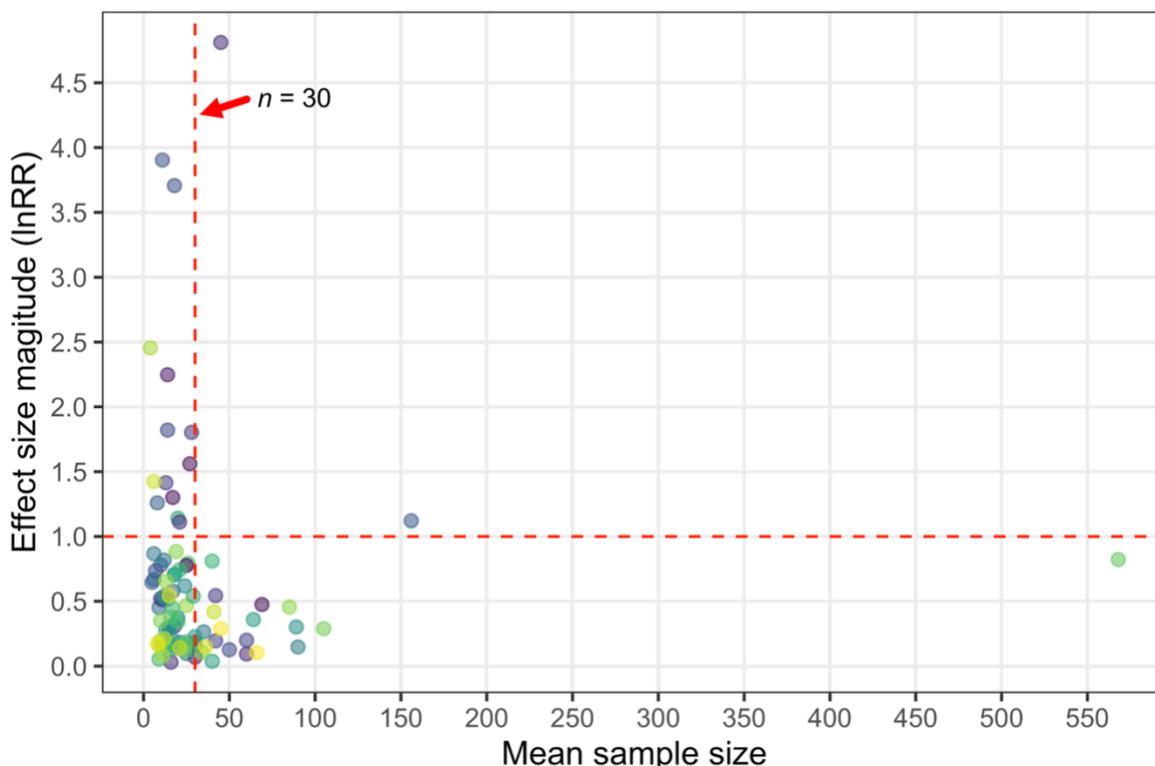
Fig. 3. Studies with large effect sizes tend to have low samples sizes. Mean effect size magnitude (absolute lnRR) for each study as a function of the mean sample size of that study (i.e., sample size per experimental treatment). Note that mean effect size for a given study is a weighted effect size magnitude, but is simply computed as the mean of individual observations for a given study. Vertical red dashed line denotes a sample size of 30, while the horizontal red dashed line represents a lnRR magnitude of 1.

Experimenter/observation bias during data collection is known to seriously skew results in behavioural research [20]. For example, non-blinded observations are common in life sciences, but are known to result in higher reported effect sizes and more significant *p*-values than blinded observations [21]. Most publications assessing ocean acidification effects on fish behaviour, including the initial three studies reporting large effect sizes, do not include statements of blinding for behavioural observations. Furthermore, given that statements of blinding can be misleading [22], there has been a call for video evidence in animal behaviour research [23]. Moreover, the persistence of inflated effects beyond initial studies can be perpetuated by confirmation bias, as follow-up studies attempt to confirm initial inflated effects and capitalise on the receptivity of high-profile journals to new (apparent) phenomena [24]. While our analysis does not empirically demonstrate that experimenter bias contributed to the decline effect, it is possible that conscious and unconscious experimenter biases may have contributed to large effect sizes in this field.

Publication and citation bias. Another prominent explanation for the decline effect is selective publication bias, as results showing strong effects are often published more readily, and in higher-impact journals, than studies showing weak or null results. Indeed, publication bias has been suggested as perhaps the most parsimonious explanation for the decline effect in ecology and evolution, as studies showing no effect can be difficult to publish [2]. This can be attributed to authors selectively publishing impressive results in prestigious journals (and not publishing less exciting results), and also to journals—particularly high impact journals—selectively publishing strong effects. This biased publishing can result in the proliferation of studies reporting strong effects, even though they may not be true [25], and can fuel citation bias [26]. To determine if studies testing for effects of ocean acidification on fish behaviour exhibited signs of publication bias and citation bias, we assessed relationships between effect size magnitude, journal impact factor, and Google Scholar citations (Fig. 4). Examining average citations per year and the total number of citations since 2020, four papers stood above the rest: the initial three studies in this field [8–10] and the sentinel paper proposing $GABA_A$ neurotransmitter interference as the physiological mechanism for observed behavioural effects [27] (Fig. 4a,b). While it is difficult to quantify whether authors selectively published only their strongest effects early in this field, we were able to quantify effect size magnitudes as a function of journal impact factor. We found that the most striking effects of ocean acidification on fish behaviour have been published in journals with high impact factors (Fig. 4c). In addition, these studies have had a stronger influence (i.e., higher citation frequency) on this field to date than lower impact studies with weaker effect sizes (Fig. 4b,c). Similar results have been reported in other areas of ecology and evolution, perhaps most notably in studies regarding terrestrial plant responses to high CO_2 [28].

Together, our results suggest that large effect sizes among studies assessing acidification impacts on fish behaviour generally have low sample size, but tend to be published in high impact journals and are cited more [26]. Consequently, the one-two punch of low sample sizes and the preference of high impact journals to publish large effects has seemingly led to an incorrect interpretation that ocean acidification will result in broad impacts on fish behaviour and thus have wide-ranging ecological consequences – an interpretation that persists in studies published today (Table S2).

Being on our best behaviour

Our results suggest that large effects of ocean acidification on fish behaviour were at least in part due to methodological factors in early studies (e.g., low sample sizes). Furthermore, the proliferation and persistence of this idea has likely been aided by the selective publication of large effect sizes by authors and journals, particularly at the onset of this field, and the continued high frequency of citations for those papers. It is important to note, however, that low sample size and selective publication cannot fully explain the strong decline effect detected here, and other biases and processes may be at play [7,29]. Nonetheless, we call on journals, journal editors, peer-reviewers, and researchers to take steps to proactively address the issues of low sample size and selective publication, not only in the ocean acidification field, but also more broadly across scientific disciplines.

To this end, we strongly argue that future ocean acidification studies on fish behaviour should employ a sample size greater than 30 fish per treatment in order to be considered reliable. It is the combined responsibility of researchers, journal editors, and peer-reviewers to ensure that submitted manuscripts abide by this guideline. To achieve this, authors should report exact sample sizes clearly in the text of manuscripts; however, from our analysis, 34% of studies did not do this adequately (see raw data in Supplementary File 2). In addition, for other fields, we suggest that studies with higher sample sizes be published alongside, if not very soon after, an original novel finding to ensure that such a finding is robust. Ideally, researchers would conduct pilot studies with varying sample sizes to determine an adequate sample size threshold; however, time and financial constraints can make this difficult. While adequate sample sizes will vary across topics and fields, ensuring studies with large sample sizes are published early alongside those with smaller sample sizes can strive toward reducing the amount of time it takes to truly understand a phenomenon.
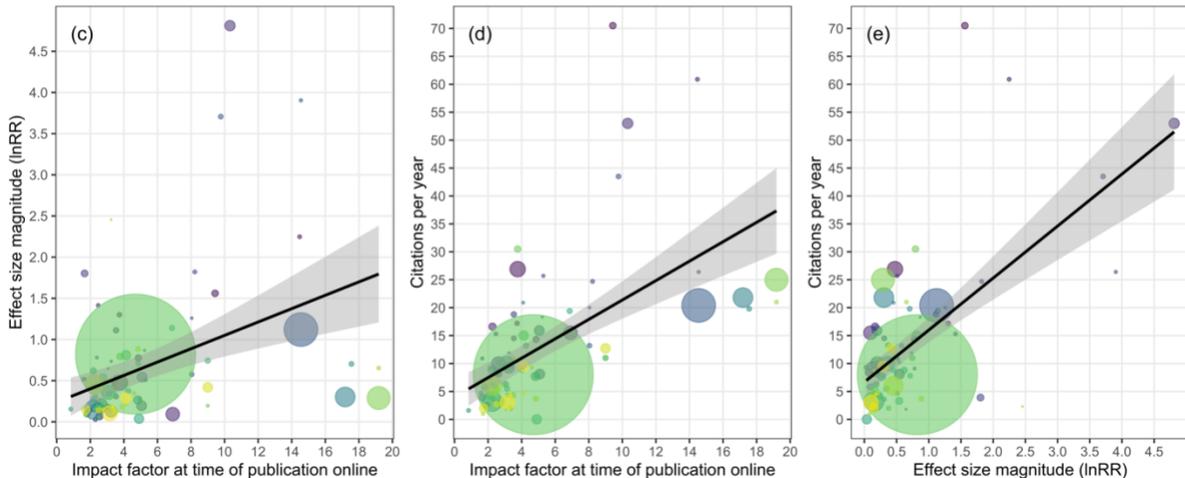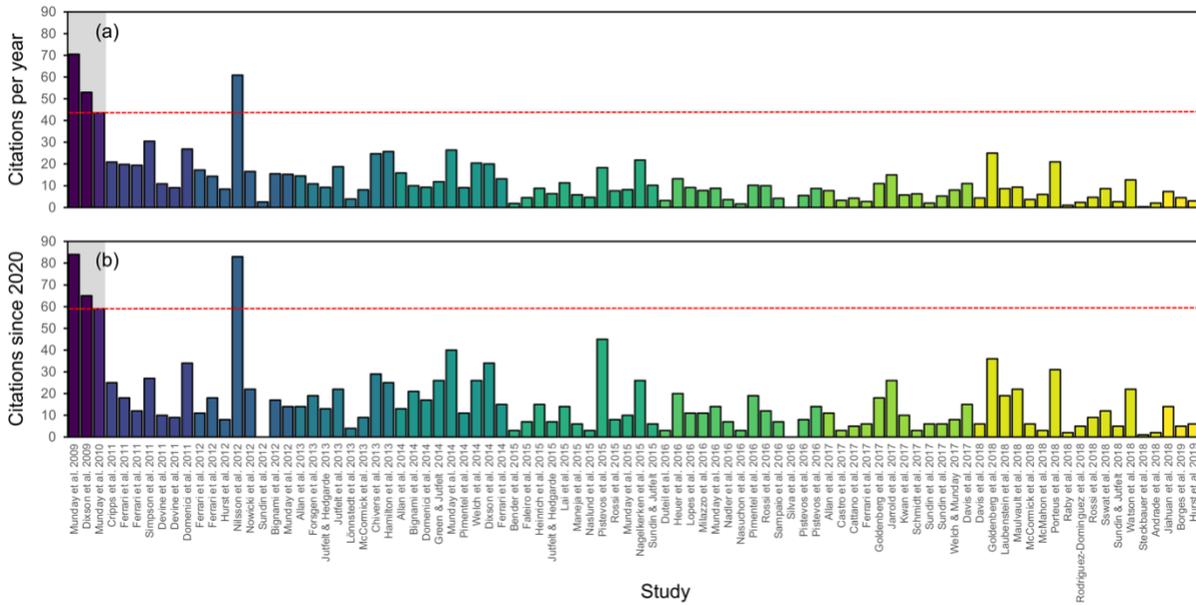
Fig. 4. Strong effects are published in high impact journals, and these studies are cited more than small effect studies in lower impact journals. (a-b) Google Scholar citation metrics as of September 10, 2021 for each of the studies included in our meta-analysis, including average citations per year **(a)** and total citations since 2020 **(b)**. The initial three studes spearheading this field are denoted by the grey background, and the red dashed line represents the lowest citation metric among those three studies. Studies are ordered chronologically along the x-axis, and colour-coded by year published online. **(c)** Mean effect size magnitude for each individual study as a function of journal impact factor (at time of online publication). **(d)** The number of citations per year for each study as a function of journal impact factor (at time of online publication) **(e)** The number of citations per year for each study as a function of mean effect size magnitude for that study. Note that, for panels (a) and (c), mean effect size for a given study is not a weighted effect size magnitude, but is simply computed as the mean of individual observations for a given study. Data are fit with linear curves and 95% confidence bounds, and points are colour-coded by study; the size of data points represents the relative mean sample size of the study.

Journals, researchers, editors, and reviewers can take additional steps to limit biases in published research. First and foremost, we suggest that journals adopt the practice of registered reports to ensure that studies not detecting an effect are published in a timely manner. Herein, journals should provide authors with the ability to submit proposed methodologies and have them formally peer reviewed prior to studies even being conducted. If methodologies are deemed sound (or revised to be so) and "accepted" by reviewers, journals should commit to publishing the results regardless of their outcome. While not a silver-bullet solution, this practice could help to minimize selective publication bias and reduce the risk of early, flawed studies being

disproportionately influential in a given field [30]. Researchers should also seek, develop, and adhere to best practice guidelines for experimental setups [31] to minimize the potential for experimental artefacts to influence results. Properly blinded observations [21] and the use of technologies such as automated tracking [32] and biosensors [33] can also reduce observer bias and increase trust in reported findings [34]. When automated methods are not possible, video recordings of experiments from start to finish can greatly increase transparency [23]. Editors and the selected peer reviewers should closely consider and evaluate the relevance and rigor of methodological approaches, which can help increase accuracy and repeatability [35]. When selecting peer-reviewers for manuscripts, editors should also be aware that researchers publishing initial strong effects may be biased in their reviews (i.e., selectively accepting manuscripts that support their earlier publications) and ensure a diverse body of reviewers for any given manuscript. While we do not empirically demonstrate this bias in our analyses, it is important to recognize and mitigate the potential for it to prolong inaccurate scientific findings.

Finally, being critical and skeptical of early findings with large effects can help avoid many of the real-world problems associated with inflated effects. Interestingly, a recent study showed that experienced scientists are highly accurate at predicting which studies will stand up to independent replication versus those that will not [36], lending support to the idea that if something seems too good to be true then it probably is. The earlier that scepticism is applied, the less impact inflated results may have on the scientific process and the public perception of scientists. Ultimately, independent replication should be established before new results are to be trusted.

Final remarks

Our results demonstrate that more than a decade of ocean acidification research on fish behaviour is characterized by the decline effect. While the field has seemingly settled in a good place with respect to realistic effect sizes, it has taken 10 years to get there. Furthermore, studies continue to cite early studies with unreasonable effect sizes to promote that acidification is predicted to broadly impact fish behaviour and ecology (e.g., Table S2), suggesting that a shift in mindset is still needed for many in this field. In a broader sense, our data reveal that the decline effect warrants exploration with respect to other biological and ecological phenomena and a wider array of scientific disciplines, particularly pertaining to global change effects. The early exaggeration of effects can have real impacts on the process of science and the people who do it [37]; following the steps outlined here can help to mitigate those impacts, sooner get to a real understanding of a phenomenon, and progress towards increased reproducibility.

Materials and methods

Literature search

Peer-reviewed articles assessing the effects of ocean acidification on fish behaviour were searched for through Scopus and Google Scholar by J. Clements up until December 21, 2018 using two primary keyword strings: '*ocean acidification fish behavio(u)r*' and '*elevated co2 fish behavio(u)r*'. The search was conducted using the free software "Publish or Perish" [38] selecting a time period spanning 2009 – 2018 and the maximum number of results that the software allows (1,000 results), ignoring citations and patents. The keyword search resulted in a total of 4,411 results, with 2,508 papers remaining for initial screening after duplicates were removed (Fig. 5, Table S3). The titles and abstracts of each article were then screened for initial relevance and inclusion criteria. Articles were included in the database if they included statements of quantitatively assessing the effect of elevated CO_2 (i.e., ocean acidification) on a behavioural trait of a marine fish; we excluded review articles and papers that measured the effect of elevated CO_2 on freshwater fishes and invertebrates. This initial screening resulted in a total of 93 papers being retained from the database search for further evaluation. Five papers were subsequently excluded from the meta-analysis due to a lack of appropriate data for estimating effect size (i.e., variance and/or sample sizes were not a part of the behavioural metric, or specific behavioural data were not presented), resulting in a total of 88 papers. A cited reference search of the 93 articles was subsequently conducted on March 23, 2019 (just prior to conducting the data analysis) by searching the reference lists and lists of citing articles (on the article's webpage) and three additional relevant papers were added, for a total of 91 papers included in the

meta-analysis. While we did not solicit a call for grey literature, which can be important for meta-analyses [3], such literature online would have been captured in the Google Scholar search; however, no relevant grey literature was uncovered in this search. Final checks of the 91 papers were conducted by both J. Clements and J. Sundin. Results of the literature search are provided in Fig. 5 below. Further details can be found in Table S3, and full search results for each step can be found in Supplementary File 2.
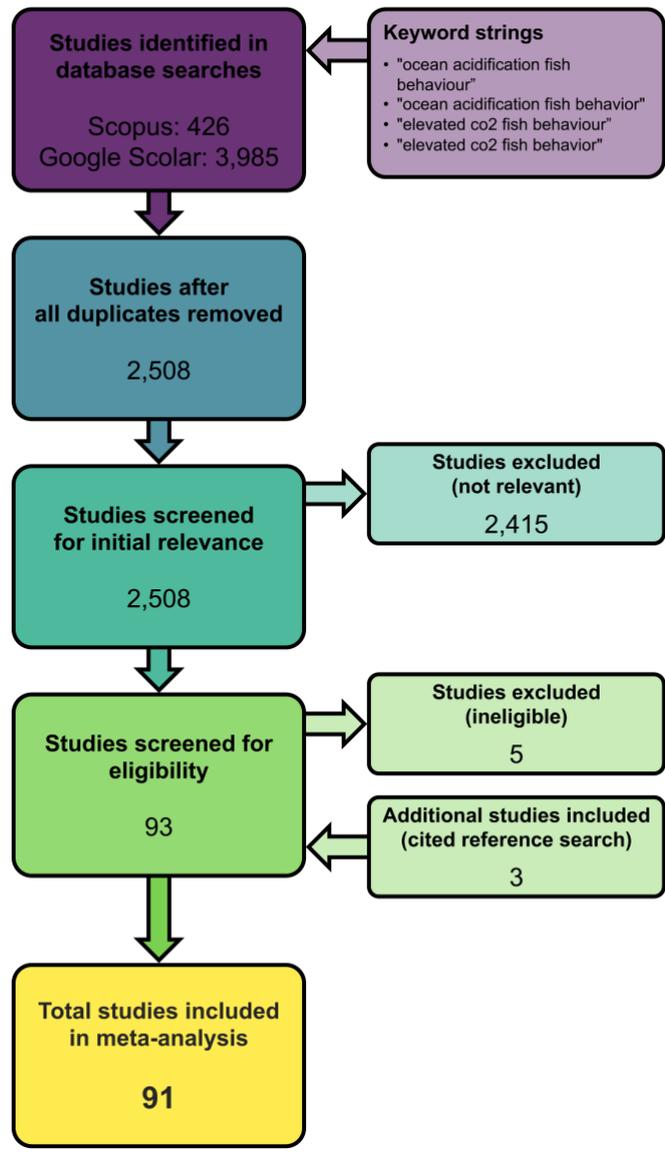
Fig. 5. PRISMA flow diagram. Values represent the numbers of records found and retained at each stage of the literature search. Papers were considered "relevant" if they included an empirical test of ocean acidification on the behaviour of a marine fish. Off-topic papers and topical review papers were excluded, as were topical papers on freshwater species and invertebrates. Relevant studies were deemed "ineligible" if they did not contain data from which effect sizes could be calculated (this included data that did not have an associated sample size or variance, or relevant papers that did not report the behavioural data). Details of relevance and exclusion can be found in Supplementary File 2.

Data collection

We collected both qualitative and quantitative data from each study. All raw data (both qualitative and quantitative) can be found in Supplementary File 3.

Qualitative data collection. From each of the 91 articles, we collected general bibliographic data, including authors, publication year, title, journal, and journal impact factor. For publication year, we recorded the year that the article was published online as well as the year that the article was included in an issue. Journal impact factor was recorded for the year of publication as well as the most current year (2017); papers published in 2018 and 2019 were assigned to the impact factor for 2017 since 2018 and 2019 data on impact factor were unavailable at the time of analysis. Impact factors were obtained from InCites Journal Citation Reports® (Clarivate Analytics). We also recorded other qualitative attributes for each study, including the species and life stage studied, and the behavioural metric(s) measured.

Quantitative data collection. Alongside qualitative data, we also collected quantitative data from each study with the exception of the five studies that were excluded due to unreported data, or other issues with data reporting and/or the nature of the data reported (i.e., if effect sizes could not be calculated from the type of data reported; see Supplementary File 2). For applicable studies, we collected the mean, sample size, and variance associated with control and ocean acidification treatments. We considered all ocean acidification treatments in our analysis; however, we only included data for independent main effects of ocean acidification, and interactive effects of acidification with other factors (temperature, salinity, pollution, noise, gabazine, etc.) were ignored.

Where possible, precise means and variance were collected from published tables or published raw data; otherwise, means and variance were estimated from published graphs using ImageJ 1.x [39]. Sample sizes were obtained from tables or the text, or were back-calculated using degrees of freedom reported in the statistical results. We also recorded the type of variance reported and, where possible, used that to calculate standard deviation, which was necessary for effect size calculations. Again, these data were not obtainable from five papers, due to either the nature of the data (i.e., no variance associated with the response variable measured, or directional response variables measured in degrees; the latter due to computational issues arising from such metrics) [40–42] or from the paper reporting an effect of ocean acidification but not adequately providing the means and/or variance in neither the paper or supplementary material [43,44]. Where means and variance were measurable but observed to be zero, we estimated both as 0.0001 in order to calculate effect size [8–10,17,45–49]. The data were used to generate effect sizes and variance estimates for each observation.

Meta-analysis

Testing for the decline effect. To assess whether or not a decline effect was evident in ocean acidification research on fish behaviour, we used two approaches: (1) visualizing the trend of effect size magnitudes for all experiments in the dataset over time; and (2) computing weighted mean effect size magnitudes for each year in our dataset and assessing the trend in mean effect size magnitudes over time.

Visualizing the decline effect using raw effect size magnitudes. First, we computed the effect size magnitude for each individual observation in our dataset and simply visualized the trend in these effect sizes over time (i.e., Fig. 2a). The effect size of choice was natural logarithmic transformed response ratio, lnRR, which is calculated as:

$$lnRR = ln\left(\frac{\bar{X}_E}{\bar{X}_C}\right)$$

where $\bar{X}_E$ and $\bar{X}_C$ are the average measured response in the experimental and control treatments, respectively. We chose lnRR because it is commonly used in ocean acidification research [50–53] and is appropriate for both continuous and ratio-type response variable data (i.e., proportions and percentages, which were abundant in our dataset) that are commonly used in behavioural studies [54,55]. Using lnRR does have drawbacks, however. Mainly, lnRR cannot be calculated when a response variable has a positive value for one treatment group and a negative value for the other. As such, we excluded measures of relative behavioural lateralization (a measure of left-right turning preference) from our analysis, as well as any index

metrics that spanned positive and negative values. For response variables that were reported as a 'change in' behaviour from a specific baseline (and could therefore have both positive and negative values), we only included instances in which the response variable values for the control treatment and elevated CO_2 treatment were both of the same directionality (i.e., both positive or both negative changes). For all such instances, the rationale for omissions and/or inclusion are provided in the 'Notes' column in Supplementary File 2.

Once calculated, the individual effect sizes were transformed to the absolute value due to the inherent difficulty in assigning a functional direction to a change in behaviour, as many behavioural changes can be characterized by both positive and negative functional trade-offs. For example, increased activity under elevated pCO$_2$ could make prey fish more difficult for predators to capture, but could also make prey more noticeable to predators. Therefore, rather than prescribing arbitrary functional directionality to altered behaviour, we simply elected to use absolute value (i.e., unsigned value) of lnRR to visualize the decline effect. It is important to note that such a transformation only provides a measure of effect size magnitude. Thus, the absolute effect size overestimates, and is therefore a conservative estimate of, the true effect size, but can still be used to test for declining effect size magnitudes over time (and can thus be used to test for the decline effect). Although this can complicate *true* population-level inferences [56], the use of absolute effect size values is informative for understanding the strength of effects ignoring directionality [57].

Assessing weighted mean effect size magnitudes by year. Although useful for visualizing a trend in effect sizes over time, the first approach above is not analytically rigorous. Properly analyzing trends in effect sizes should including a weighted component whereby individual effect sizes are weighted according to their precision (i.e., measurements with a larger sample size and lower variance should be given more weight than those with a lower sample size and higher variance) [58]. As such, we computed weighted mean effect size magnitudes (and their associated uncertainty; i.e., upper and lower confidence bounds) for each year represented in our dataset, and assessed the trend in these effect sizes over time.

Weighted mean effect size magnitudes (lnRR) and their confidence bounds were computed using the "transform-then-analyze" approach as suggested by [59], with R code adapted from [60] to avoid biased estimates of effect size magnitude. Briefly, this method estimates the mean effect size for each level of a moderator of interest (i.e., each year in our dataset) by assuming a normal distribution and subsequently transforming the mean effect size using a folded normal distribution to estimate a mean effect size magnitude. Uncertainty around the mean effect size magnitude was estimated in a Bayesian fashion using the MCMCglmm() function from the *MCMCglmm* package [61], applying the entire posterior distribution of mean estimated to the folded normal distribution as per [60]. Annotated R code and all raw data files for the analysis are provided in the supplementary material for reproducibility (Supplementary files 4–11).

Assessing biological explanations for the decline effect. Since a decline effect was detected in our analysis, we explored three biological factors that might explain the observed decline effect: 1. climate (cold-water *vs.* warm-water species); 2. behaviour type (olfactory *vs.* non-olfactory behaviours); and 3. life stage (larvae *vs.* juveniles and adults).

Because early studies were focused on warm-water fish from tropical coral reefs, the observed decline effect could potentially be driven by an increasing number of studies on less sensitive cold-water species over time. Cold-water fish in temperate regions experience a higher degree of temporal variability in carbonate chemistry parameters over large spatial areas [15]. Therefore, they may be less sensitive to changes in seawater CO_2 as per the Ocean Variability Hypothesis [16]. If an increasing number of studies on cold-water species over time was responsible for the decline effect, removing cold-water species from the dataset (i.e., only including warm-water species) should result in the decline effect trend disappearing. In the same vein, the strongest effects of ocean acidification on fish behaviour have undoubtedly been reported for olfactory responses, and an increasing number of studies on non-olfactory behaviours could explain the decline effect. If this was true, removing non-olfactory behaviours from the dataset should negate the decline effect trend, so we therefore tested for the influence of non-olfactory behaviours by removing them from the dataset and re-running the analysis. Finally, larvae are typically considered to be more sensitive to acidification than juveniles and adults, and removing less sensitive life stages from the

dataset would remove the decline effect trend if this explanation was responsible for the decline (i.e., if studies using less sensitive life stages had increased proportionally over time). Therefore, to test whether or not the decline effect was due to these three biological factors, re-ran analysis the analysis described in **"Assessing weighted mean effect sizes by year"** above on three separate datasets; one with cold water species removed, one with non-olfactory responses removed, and one with juvenile and adult life stages removed.

Assessing evidence for selective publication bias, citation bias, and methodological bias. Alongside testing for the decline effect, we also wanted to determine whether publication bias and/or methodological bias may have contributed to the large effect sizes reported in this field, and whether there was any evidence for citation bias. In new and emerging topics, large effect sizes can be driven by authors and high-impact journals selectively publishing novel and ground-breaking results with large effect sizes [62]. If selective publication bias was evident among studies testing for effects of ocean acidification on fish behaviour, there would be a positive relationship between effect size magnitude and journal impact factor *sensu* [28]. Thus, to determine if selective publication bias could be present in this field, we visually assessed the relationship between the journal impact factor (for the year of online publication) and the mean effect size magnitude for each study. It is important to note here that we did not compute weighted mean effect size magnitudes for each study, but simply computed the mean of the raw effect size magnitudes as calculated in the section **"Visualizing the decline effect using raw effect size magnitudes"** above. To check for citation bias, we visually assessed the relationship between impact factor and the number of citations per year (according to Google Scholar on September 10, 2021) for each study, as well as the relationship between mean effect size magnitude and citations per year. If citation bias was present in this field, citations per year would be positively correlated with mean effect size magnitude. Furthermore, if selective publication bias was influencing citation bias, a positive relationship between impact factor and citations per year would be present. Finally, to assess if low sample sizes could contribute to large effect sizes (i.e., higher probability of Type 1 error), we plotted mean effect size magnitude for each study against the mean sample size of that study. If low sample size was influencing effect sizes among studies in this field, large effect sizes would cluster near the lower end of the sample size spectrum.

Acknowledgements

Funding

Data Accessibility Statement

All statistical results, raw data, R code, and original datafiles uploaded to R are available as supplementary material.

Competing interests

We declare we have no competing interests.

Supplementary material

Supplementary File 1. Supplementary tables (Fig S1–S3).
Supplementary File 2. Full, detailed literature search results.
Supplementary File 3. Raw data collected from each study.
Supplementary File 4. Annotated R code.
Supplementary Files 5–11. Raw data files for analysis in R (see annotated code for specific usage of data files).

References

1. Schooler J. Unpublished results hide the decline effect. Nature. 2011;470: 437–437. doi:10.1038/470437a
2. Jennions MD, Møller AP. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proc R Soc B. 2002;269: 43–48. doi:10.1098/rspb.2001.1832
3. Sánchez-Tójar A, Nakagawa S, Sánchez-Fortún M, Martin DA, Ramani S, Girndt A, et al. Meta-analysis challenges a textbook example of status signalling and demonstrates publication bias. eLife. 2018;7: e37385. doi:10.7554/eLife.37385
4. Koricheva J, Kulinskaya E. Temporal instability of evidence base: A threat to policy making? Trends Ecol Evol. 2019;34: 895–902. doi:10.1016/j.tree.2019.05.006
5. Kroeker KJ, Kordas RL, Crim RN, Singh GG. Meta-analysis reveals negative yet variable effects of ocean acidification on marine organisms: Biological responses to ocean acidification. Ecol Lett. 2010;13: 1419–1434. doi:10.1111/j.1461-0248.2010.01518.x
6. Browman HI. Applying organized scepticism to ocean acidification research. ICES J Mar Sci. 2016;73: 529.1-536. doi:10.1093/icesjms/fsw010
7. Clark TD, Raby GD, Roche DG, Binning SA, Speers-Roesch B, Jutfelt F, et al. Ocean acidification does not impair the behaviour of coral reef fishes. Nature. 2020;577: 370–375. doi:10.1038/s41586-019-1903-y
8. Munday PL, Dixson DL, Donelson JM, Jones GP, Pratchett MS, Devitsina GV, et al. Ocean acidification impairs olfactory discrimination and homing ability of a marine fish. Proc Natl Acad Sci USA. 2009;106: 1848–1852. doi:10.1073/pnas.0809996106
9. Dixson DL, Munday PL, Jones GP. Ocean acidification disrupts the innate ability of fish to detect predator olfactory cues. Ecol Lett. 2010;13: 68–75. doi:10.1111/j.1461-0248.2009.01400.x
10. Munday PL, Dixson DL, McCormick MI, Meekan M, Ferrari MCO, Chivers DP. Replenishment of fish populations is threatened by ocean acidification. Proc Natl Acad Sci USA. 2010;107: 12930–12934. doi:10.1073/pnas.1004519107
11. Black R. Acid oceans turn "Finding Nemo" deaf. In: BBC News [Internet]. 2011 [cited 8 Nov 2019]. Available: https://www.bbc.com/news/science-environment-13605113
12. Yong E. Losing Nemo - Acid oceans prevent baby clownfish from finding home. In: Discover Magazine [Internet]. 2009 [cited 8 Nov 2019]. Available: http://blogs.discovermagazine.com/notrocketscience/2009/02/02/losing-nemo-acid-oceans-prevent-baby-clownfish-from-finding-home/#.XcW48zJKhTZ
13. Dixson DL. Increasingly acidiic oceans are causing fish to behave badly. Sci Am. 2017;316: 42–45. doi:10.1038/scientificamerican0617-42
14. Roberts KB. Sea change: UD's Dixson discusses ocean acidification at White House briefing. In: UDaily [Internet]. 2015 [cited 8 Nov 2019]. Available: http://www1.udel.edu/udaily/2016/dec/ocean-acidification-120415.html
15. Kwiatkowski L, Orr JC. Diverging seasonal extremes for ocean acidification during the twenty-first century. Nature Clim Change. 2018;8: 141–145. doi:10.1038/s41558-017-0054-0
16. Baumann H. Experimental assessments of marine species sensitivities to ocean acidification and co-stressors: how far have we come? Can J Zool. 2019;97: 399–408. doi:10.1139/cjz-2018-0198
17. Munday PL, Pratchett MS, Dixson DL, Donelson JM, Endo GGK, Reynolds AD, et al. Elevated CO_2 affects the behavior of an ecologically and economically important coral reef fish. Mar Biol. 2013;160: 2137–2144. doi:10.1007/s00227-012-2111-6
18. Columb M, Atkinson M. Statistical analysis: sample size and power estimations. BJA Educ. 2016;16: 159–161. doi:10.1093/bjaed/mkv034
19. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. 2015;13: e1002106. doi:10.1371/journal.pbio.1002106
20. Marsh DM, Hanlon TJ. Seeing what we want to see: Confirmation bias in animal behavior research. Ethology. 2007;113: 1089–1098. doi:10.1111/j.1439-0310.2007.01406.x
21. Holman L, Head ML, Lanfear R, Jennions MD. Evidence of Experimental Bias in the Life Sciences: Why We Need Blind Data Recording. PLoS Biol. 2015;13: e1002190. doi:10.1371/journal.pbio.1002190

22. Berg J. Addendum to "Editorial Retraction of the Report 'Environmentally relevantconcentrations of microplastic particles influence larval fish ecology,' by O. M.Lönnstedt and P. Eklöv." Science. 2017;358: 1549. doi:10.1126/science.aar7766

23. Clark TD. Science, lies and video-taped experiments. Nature. 2017;542: 139–139. doi:10.1038/542139a

24. Duarte CM, Fulweiler RW, Lovelock CE, Martinetto P, Saunders MI, Pandolfi JM, et al. Reconsidering ocean calamities. BioScience. 2015;65: 130–139. doi:10.1093/biosci/biu198

25. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2: e124. doi:10.1371/journal.pmed.0020124

26. Jannot A-S, Agoritsas T, Gayet-Ageron A, Perneger TV. Citation bias favoring statistically significant studies was present in medical research. J Clin Epidemiol. 2013;66: 296–301. doi:10.1016/j.jclinepi.2012.09.015

27. Nilsson GE, Dixson DL, Domenici P, McCormick MI, Sørensen C, Watson S-A, et al. Near-future carbon dioxide levels alter fish behaviour by interfering with neurotransmitter function. Nature Clim Change. 2012;2: 201–204. doi:10.1038/nclimate1352

28. Murtaugh PA. Journal quality, effect size, and publication bias in meta-analysis. Ecology. 2002;83: 1162–1166. doi:10.1890/0012-9658(2002)083[1162:JQESAP]2.0.CO;2

29. Enserink M. Sea of doubts. Science. 2021;372: 560–565. doi:10.1126/science.372.6542.560 2

30. Gonzales JE, Cunningham CA. The promise of pre-registration in psychological research. In: American Psychological Association [Internet]. 2015 [cited 8 Nov 2019]. Available: https://www.apa.org/science/about/psa/2015/08/pre-registration

31. Jutfelt F, Sundin J, Raby GD, Krång A-S, Clark TD. Two-current choice flumes for testing avoidance and preference in aquatic animals. Reynolds J, editor. Methods Ecol Evol. 2017;8: 379–390. doi:10.1111/2041-210X.12668

32. Dell AI, Bender JA, Branson K, Couzin ID, de Polavieja GG, Noldus LPJJ, et al. Automated image-based tracking and its application in ecology. Trends Ecol Evol. 2014;29: 417–428. doi:10.1016/j.tree.2014.05.004

33. Clements JC, Comeau LA. Use of high frequency, non-invasive electromagnetic biosensors to detect ocean acidification effects on shellfish behavior. J Shellfish Res. 2019;38: 1–8. doi:10.2983/035.038.0300

34. Traniello JFA, Bakker TCM. Minimizing observer bias in behavioral research: blinded methods reporting requirements for *Behavioral Ecology and Sociobiology*. Behav Ecol Sociobiol. 2015;69: 1573–1574. doi:10.1007/s00265-015-2001-2

35. Hofseth LJ. Getting rigorous with scientific rigor. Carcinogenesis. 2018;39: 21–25. doi:10.1093/carcin/bgx085

36. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behav. 2018;2: 637–644. doi:10.1038/s41562-018-0399-z

37. Clements JC. Reproducibility and mental health. Nature. 2020;582: 300. doi:https://doi.org/10.1038/d41586-020-01642-9

38. Harzing A-W. Publish or Perish on macOS. 2019. Available: https://harzing.com/resources/publish-or-perish

39. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9: 671–675. doi:10.1038/nmeth.2089

40. Maneja RH, Frommel AY, Browman HI, Clemmesen C, Geffen AJ, Folkvord A, et al. The swimming kinematics of larval Atlantic cod, *Gadus morhua* L., are resilient to elevated seawater pCO_2. Mar Biol. 2013;160: 1963–1972. doi:10.1007/s00227-012-2054-y

41. Devine BM, Munday PL. Habitat preferences of coral-associated fishes are altered by short-term exposure to elevated CO_2. Mar Biol. 2013;160: 1955–1962. doi:10.1007/s00227-012-2051-1

42. Poulton DA, Porteus CS, Simpson SD. Combined impacts of elevated CO_2 and anthropogenic noise on European sea bass (*Dicentrarchus labrax*). Juanes F, editor. ICES Journal of Marine Science. 2017;74: 1230–1236. doi:10.1093/icesjms/fsw003

43. Schunter C, Welch MJ, Nilsson GE, Rummer JL, Munday PL, Ravasi T. An interplay between plasticity and parental phenotype determines impacts of ocean acidification on a reef fish. Nat Ecol Evol. 2018;2: 334–342. doi:10.1038/s41559-017-0428-8

14

44. Schunter C, Welch MJ, Ryu T, Zhang H, Berumen ML, Nilsson GE, et al. Molecular signatures of transgenerational response to ocean acidification in a species of reef fish. Nature Clim Change. 2016;6: 1014–1018. doi:10.1038/nclimate3087

45. Munday PL, Cheal AJ, Dixson DL, Rummer JL, Fabricius KE. Behavioural impairment in reef fishes caused by ocean acidification at CO2 seeps. Nature Clim Change. 2014;4: 487–492. doi:10.1038/nclimate2195

46. Lönnstedt OM, Munday PL, McCormick MI, Ferrari MCO, Chivers DP. Ocean acidification and responses to predators: can sensory redundancy reduce the apparent impacts of elevated CO $_2$ on fish? Ecol Evol. 2013;3: 3565–3575. doi:10.1002/ece3.684

47. Bender D, Champ CM, Kline D, Diaz-Pulido G, Dove S. Effects of "Reduced" and "Business-As-Usual" CO_2 emission scenarios on the algal territories of the damselfish *Pomacentrus wardi* (Pomacentridae). Sotka E, editor. PLoS ONE. 2015;10: e0131442. doi:10.1371/journal.pone.0131442

48. Pimentel MS, Faleiro F, Marques T, Bispo R, Dionísio G, Faria AM, et al. Foraging behaviour, swimming performance and malformations of early stages of commercially important fishes under ocean acidification and warming. Climatic Change. 2016;137: 495–509. doi:10.1007/s10584-016-1682-5

49. Rodriguez-Dominguez A, Connell SD, Baziret C, Nagelkerken I. Irreversible behavioural impairment of fish starts early: Embryonic exposure to ocean acidification. Marine Pollution Bulletin. 2018;133: 562–567. doi:10.1016/j.marpolbul.2018.06.004

50. Harvey BP, Gwynn-Jones D, Moore PJ. Meta-analysis reveals complex marine biological responses to the interactive effects of ocean acidification and warming. Ecol Evol. 2013;3: 2782–2782. doi:10.1002/ece3.728

51. Kroeker KJ, Kordas RL, Crim R, Hendriks IE, Ramajo L, Singh GS, et al. Impacts of ocean acidification on marine organisms: quantifying sensitivities and interaction with warming. Glob Change Biol. 2013;19: 1884–1896. doi:10.1111/gcb.12179

52. Brown NEM, Bernhardt JR, Anderson KM, Harley CDG. Increased food supply mitigates ocean acidification effects on calcification but exacerbates effects on growth. Sci Rep. 2018;8: 9800. doi:10.1038/s41598-018-28012-w

53. Clements JC, Darrow ES. Eating in an acidifying ocean: a quantitative review of elevated CO_2 effects on the feeding rates of calcifying marine invertebrates. Hydrobiologia. 2018;820: 1–21. doi:10.1007/s10750-018-3665-1

54. Hintze J. NCSS statistical system user's guide IV: Multivariate analysis, clustering, metaanalysis, forecasting / time series, operations research, mass appraisal. Kaysville, UT: NCSS; 2007.

55. Pustejovsky JE. Using response ratios for meta-analyzing single-case designs with behavioral outcomes. Journal of School Psychology. 2018;68: 99–112. doi:10.1016/j.jsp.2018.02.003

56. Paulus FM, Krach S, Albrecht A-G, Jansen A. Potential bias in meta-analyses of effect sizes in imaging genetics. Schizophrenia Bulletin. 2013;39: 501–503. doi:10.1093/schbul/sbt035

57. Garamszegi LZ. Comparing effect sizes across variables: generalization without the need for Bonferroni correction. Behavioral Ecology. 2006;17: 682–687. doi:10.1093/beheco/ark005

58. Hedges L, Olkin I. Statistical methods for meta-analysis. New York: Academic Press; 1985.

59. Morrissey MB. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. J Evol Biol. 2016;29: 1882–1904. doi:10.1111/jeb.12950

60. García-Roa R, Garcia-Gonzalez F, Noble DWA, Carazo P. Temperature as a modulator of sexual selection. Biol Rev. 2020;95: 1607–1629. doi:10.1111/brv.12632

61. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm *R* package. J Stat Soft. 2010;33: 1–22. doi:10.18637/jss.v033.i02

62. Sterne J, Egger M, Smith G. Investigating and dealing with publication and other biases. In: Egger M, Smith G, Altman D, editors. Systematic reviews in health care: Meta-analysis in context, Second Edition. London: BMJ Publishing Group; 2001. pp. 189–208.