An extreme decline effect in ocean acidification ecology

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

***Correspondence:** Jeff C. Clements, PhD

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

Current mailing addresss: 42 Senese Street, Moncton, NB E1E 0B8, Canada

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

Abstract

Ocean acidification – deceasing oceanic pH resulting from the uptake of excess atmospheric CO_2 – is expected 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 will be the most catastrophic. Recent studies documenting a lack of effect of experimental ocean acidification on fish behaviour, however, call this dire prediction into question. Here, we critically assess the past decade of ocean acidification research regarding direct effects on fish behaviour. Using a meta-analysis, we provide quantitative evidence that the research to date on this topic is strongly characterized by a phenomenon known as the "decline effect", where large effects have all but disappeared over a decade. The decline effect in this field cannot be explained biologically, but is strongly associated with well-known biases to which the process of science is generally prone. We contend that ocean acidification does not have as much of a direct impact on fish behaviour as previously thought, 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 | global change biology | scientific process

Introduction

Ground-breaking scientific discoveries are often followed by attempts to replicate and build upon the research. In many instances, however, follow-up studies fail to replicate initial effects. Indeed, this inability to replicate initial results is characteristic of many scientific fields and has fuelled the so-called 'reproducibility crisis' in science (Baker 2016).

The tendency for initial scientific findings—which can show outstanding effects—to lose strength over time is referred to as the 'decline effect' (Schooler 2011). This phenomenon was first described in the 1930s, and has since been described in a range of scientific disciplines (Schooler 2011). It captures the concept of inflated initial reports that overestimate reality; the real magnitude of an effect is only revealed once follow-up studies accumulate. In such instances, the early inflation of effect sizes is the key problem, not the subsequent decline; the 'decline effect' could therefore equally be referred to as the 'early inflation effect'.

Here, we present the most striking example of the decline effect in ecology using research on ocean acidification and fish behaviour—a growing field that has underscored profound and concerning effects on ecosystem resilience. We provide evidence that initial effects of acidification on fish behaviour appear drastically overestimated, and present quantitative evidence for the biases causing the decline effect over the past decade. Ways to mitigate the issues, applicable to any scientific field, are proposed.

Fishy effects

Over the past 15 years, biologists have documented substantial negative effects of ocean acidification on marine biota (Kroeker et al. 2010). With more than 300 papers published per year from 2006 to 2015, the exponential growth of studies in this field is unprecedented in the marine sciences (4). In recent years, however, there has been increasing skepticism and uncertainty around the severity of ocean acidification effects on marine organisms (Browman 2016; Clark et al. 2020).

Some of the most profound reports are those concerning fish behaviour, whereby a series of sentinel papers in 2009 and 2010 published in prestigious journals reported outstanding effects of laboratory-simulated ocean acidification (Munday et al. 2009, 2010; Dixson et al. 2010). 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 (Yong 2009; Bllack 2011; Dixson 2017) and through a presentation at the White House (Roberts 2015), perhaps partly due to the charismatic nature of the species studied (clownfish, *sensu* Finding Nemo). Not only were the findings alarming, the extraordinarily clear and strong results left little doubt that the effects were real, and a multimillion-dollar investment of research funding was initiated to quantify the impact. In recent years, however, an increasing number of papers have reported a lack of ocean acidification effects on fish behaviour, calling into question the universality and reliability of initial effects (Fig. 1a). To shed light on this phenomenon of global relevance, we investigated whether or not the presence of a decline effect existed in ocean acidification studies concerning fish behaviour (see Supplementary File 1 for methods).

Based on a systematic literature review (n = 95 studies), we found clear evidence for a decline effect in ocean acidification studies on fish behaviour. Strong effects of ocean acidification, as concluded by the authors of the studies, have decreased dramatically over time (Fig. 1a, b). Furthermore, we found that the mean effect size magnitude (absolute [unsigned] *lnRR*) was disproportionately large in early studies and has all but disappeared over time (Fig. 1c).

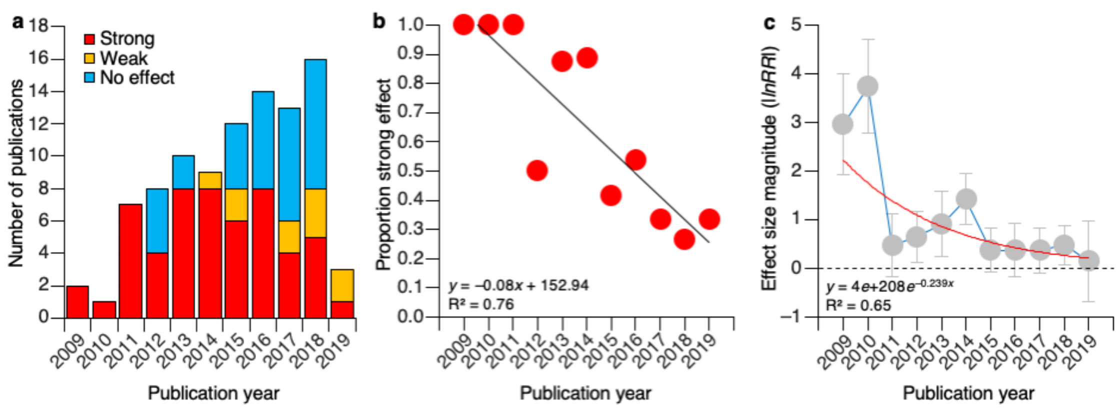

Fig. 1. The decline effect in ocean acidification research on fish behaviour. (a) Number of papers testing for effects of ocean acidification on fish behaviour over the past 11 years reporting strong effects (red) , weak effects (yellow), and no effects (blue). **(b)** Proportion of articles concluding a strong effect as a function of time (publication year). **(c)** Mean effect size magnitude (absolute *lnRR*) as a function of time (publication year). For **(c)**, error bars denote 95% confidence intervals and the dashed line indicates an effect size of 0. Note that using mean effect size magnitude results in an over-estimate of the 'true' effect size (see Supplementary File 1 for further details).

Outstandingly 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, and the magnitudes of effect sizes have been statistically similar to zero in four of the past five years (Fig. 1c). Furthermore, the decline effect persisted when we accounted for two potential biological explanations: an increasing number of studies on (potentially) less-sensitive cold-water species, and an increasing number of studies measuring baseline behaviours (i.e., not behaviours in response to a cue) (Fig. 2; see Supplementary File 1 for methods). Together, these findings show that ocean acidification studies on fish behaviour are strongly characterized by the decline effect, perhaps to the most extreme level found in the biological literature to date.

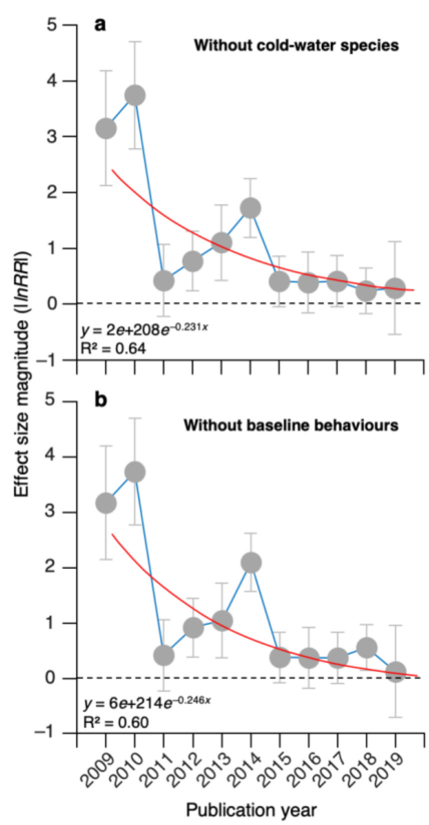

Fig 2. The decline effect cannot be explained by increasing number of studies on cold-water species, nor an increasing number of studies on baseline behaviours, over time. Mean effect size magnitude (absolute *lnRR*) as a function of time with cold-water species **(a)** and baseline behaviours (i.e., no stimulus or cue) (**b)** removed. Data are presented as weighted means and 95% confidence intervals.

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 (Browman 2016). The key thing to note is that if science concerning ocean acidification and fish behaviour was operating properly and early effects were true, the relationships presented in Figs. 1 and 2, would be flat lines. It also appears that the decline effect discovered herein for this field is not explainable by two likely biological culprits. Thus, the data presented here provide one of the strongest examples to date of a new and emerging field being prone to biases. Below, we underscore and quantitatively assess the roles of two potential biases: methodological bias (low sample sizes) and publication bias (selective publishing).

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 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 (Columb & Atkinson 2016). 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 (Head et al. 2015). In our analysis, we found that almost all of the studies with the largest effect size magnitudes had sample sizes below 30 fish. Indeed, 96% of the studies with a mean effect size magnitude above one had a sample size below 30 (Fig. 3a) and, when binned, only sample size bins below 30 had a mean effect size magnitude significantly greater than 0 (Fig. 3b). Encouragingly, however, we also found that sample sizes of studies have generally increased over time (Fig 3c,d), suggesting that the observed decline effect can at least partly be explained by increasing statistical power. This highlights the self-correcting nature of science and is indicative of maturation in this field.

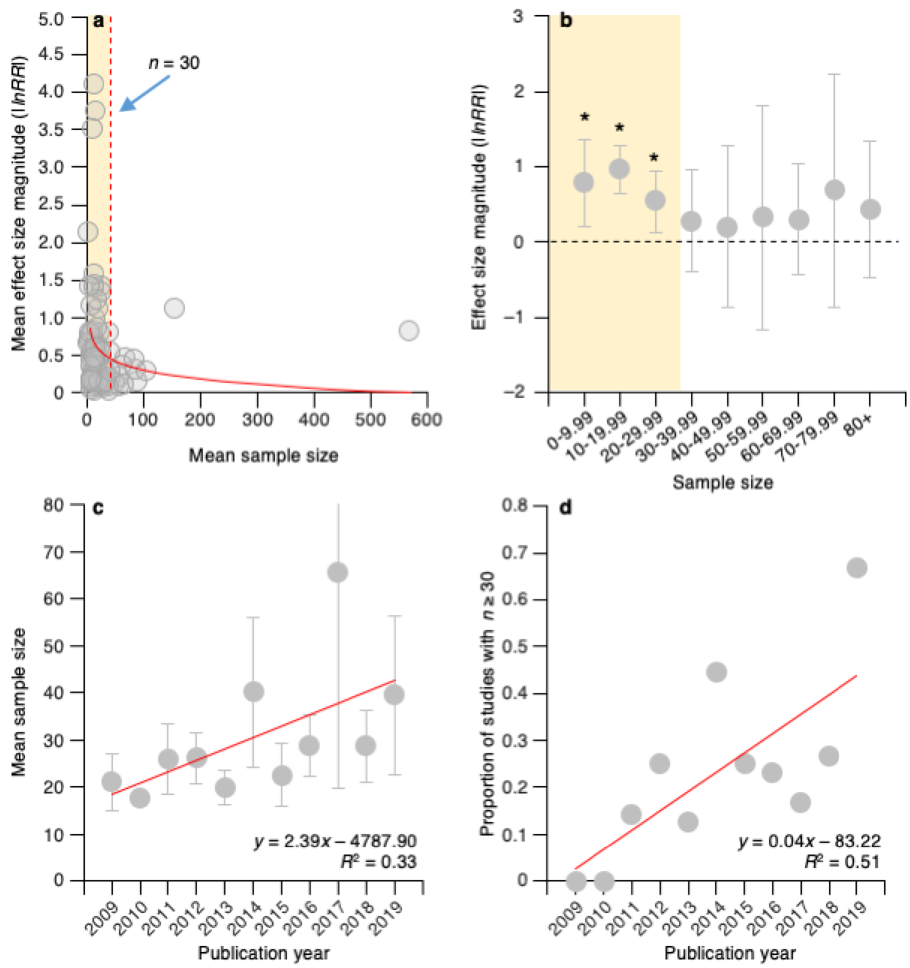

Fig. 3. Extreme effects may be false positives due to low sample size, and the decline effect is at least partially explained by increasing statistical power over time. (a) Mean effect size magnitude (absolute *lnRR*) as a function of mean sample size. Each point represents a single study. The vertical dashed line represents the arbitrary threshold after which extreme effects are not observed (n = 30). **(b)** Mean effect size magnitude (absolute *lnRR*) as a function of sample size bins. Asterisks denote mean effect size magnitudes that are significantly different from 0 (interpret with caution, as effect size magnitudes are overestimates of true effect size). **(c)** Mean study sample size (± standard error) as a function of publication year. **(d)** The proportion of studies with a sample size above 30, after which extreme effects are not typically observed.

Experimenter/observation bias during data collection is known to seriously skew results in behavioural research (Marsh & Hanlon 2007). Indeed, it appears that clear statements of blinded observation or other means of reducing experimenter bias have only become prevalent in recent years. 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 (Duarte et al. 2015).

Publication biases. Another prominent explanation for the decline effect is publication bias, as results showing strong effects are often published more readily, and in higher-impact journals, than studies showing weak or null results. Publication bias 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 drastic effects, even though they may not be true (Ioannidis 2005). While it is difficult to quantify whether authors selectively publish only their strongest effects, we were able to quantify mean 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 very high impact factors (Fig. 4a). In addition, the average impact factor of journals publishing ocean acidification research on fish behaviour has generally decreased over time (Fig. 4b). Intriguingly, a temporary increase in mean effect size magnitude in 2014 was accompanied by a temporary increase in the average journal impact factor (compare Fig. 1b and Fig. 4b), providing strong evidence that high impact journals selectively publish studies reporting extreme effects. This is a troubling finding because it means that studies reporting extreme effects of ocean acidification on fish behaviour will be highly cited within this field even though those findings are likely to be false positives (as evidenced in our sample size analysis above). Consequently, the one-two punch of low sample sizes and the preference of journals to publish extreme effects has led to a likely incorrect interpretation that ocean acidification will catastrophically impair fish behaviour and thus have wide ranging ecological consequences.

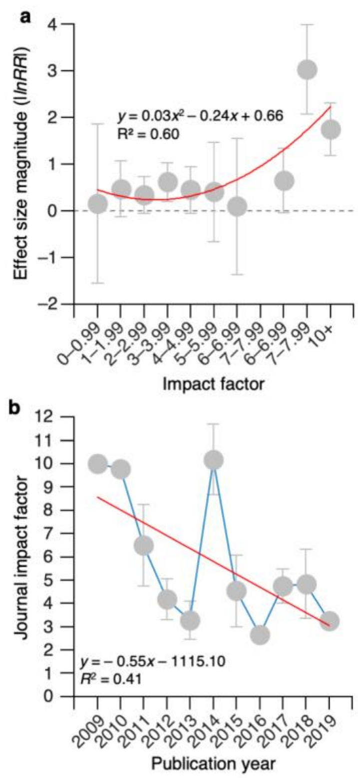

Fig. 4. Strong effects are restricted to high impact journals, and the average impact factor of journals publishing ocean acidification research on fish behaviour has declined over time. (a) Mean effect size magnitude (absolute *lnRR*, ± 95% CI) as a function of journal impact factor bin. **(b)** Mean journal impact factor (± standard error) as a function of publication year. Note the increase in impact factor for 2014 in **(b)**, which is associated with a concurrent increase in mean effect size magniotude in the same year (see Fig. 1c).

Being on our best behaviour

Our results provide strong evidence that the dramatic reports of ocean acidification affecting fish behaviour were likely due to methodological limitations and biases in early studies (e.g., low sample sizes, experimenter biases). Furthermore, the proliferation and persistence of this idea has been aided by publication bias, driven by the selective publication of outstanding effects by authors and journals. As such, we call on journals, journal editors, peer-reviewers, and researchers to take steps to proactively address this issue, 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 must 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).

Journals, researchers, editors, and reviewers can take additional steps to ensure that only unbiased empirical results are obtained and published. First and foremost, we suggest that journals adopt the practice of pre-registration to ensure that all negative results are published in a timely manner. This practice would minimize publication bias and reduce the risk of early, flawed studies being disproportionately influential in a given field (Gonzales & Cunningham 2015). Researchers should also seek, develop, and adhere to best practice guidelines for experimental setups (Jutfelt et al. 2017) to minimize the potential for experimental artefacts to influence results. Properly blinded observations and the use of technologies such as automated tracking (Dell et al. 2014) and biosensors (Clements & Comeau 2019) can also reduce observer bias and increase trust in reported findings (Traniello & Bakker 2015). When automated methods are not possible, video recordings of experiments from start to finish can greatly increase transparency (Clark 2017). 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 (Hofseth 2018).

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.

Finally, being critical and sceptical 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 (Camerer et al. 2018), 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 fully trusted.

Final remarks

Does ocean acidification affect the biology of marine animals? In many instances, most probably yes. Our data demonstrate, however, that more than a decade of ocean acidification research on fish behaviour is strongly characterized by the decline effect. In a broader sense, our data reveal that the decline effect is real and warrants exploration with respect to other biological and ecological phenomena and a wider array of scientific disciplines. The early exaggeration of effects can have real impacts on the process of science; 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 in Scopus and Google Scholar by J. Clements. up until March 23, 2019 using two primary keyword strings: '*ocean acidification fish behavio(u)r*' and '*elevated co2 fish behavio(u)r*'. The abstract of each article was then screened for relevance and inclusion criteria. Articles were included in the database if they quantitatively assessed the effect of elevated pCO_2 (i.e., ocean acidification) on a behavioural trait of a marine fish; we excluded papers that measured the effect of elevated pCO_2 on freshwater fishes and invertebrates. The reference lists of each included article were then screened for additional papers that may have been missed using the online search, which were subsequently added to the database. Once the database was established by J. Clements, it was cross-checked by J. Sundin. and any additional relevant papers were added. Final checks were conducted by both J. Clements and J. Sundin. This approach resulted in a total of 95 peer-reviewed articles assessing the effect of ocean acidification on fish behaviour, comprising the most comprehensive database for this field to date.

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 2.

Qualitative data collection

From each of the 95 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. Finally, we qualitatively scored the strength of the overall effect that ocean acidification had on behaviour for each study, based on the authors conclusions and the reported results. Strength was scored as either ‘Strong Effect’, ‘Weak Effect’, or ‘No Effect’. A study was categorized as having a ‘Strong Effect’ when ocean acidification affected all or a majority of behaviours assessed in the study, and if the authors concluded a unanimous effect of acidification. In contrast, a study was categorized as having ‘No Effect’ of acidification when none of the behaviours assessed were affected by acidification, and the authors concluded that acidification did not affect behaviour. A study was categorized as showing a ‘Weak Effect’ if a minority of behaviours were affected by ocean acidification and the authors concluded that acidification had some, but weak, effects on behaviour.

Quantitative data collection

Alongside qualitative data, we also collected quantitative data from each study with the exception of 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 effects of ocean acidification, and discarded acidification effects when they interacted with other variables explored in a given study (temperature, salinity, pollution, noise, gabazine, etc.).

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 (Schneider et al. 2012). 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. These data were not obtainable from two 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) (Maneja et al, 2012; Devine et al. 2013; Poulton et al. 2017) 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 (Schunter et al. 2016, 2018). Where means and variance were measurable but observed to be zero, we estimated both as 0.0001 in order to calculate effect size (Munday et al. 2009, 2010; Dixson et al. 2010; Lönnstedt et al. 2013; Munday et al. 2013, 2014; Bender et al. 2015; Pimentel et al. 2016; Rodriguez-Dominguez et all. 2018).

The data were used to generate effect sizes and their variance estimates for each observation. 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. This effect size metric is commonly used in ocean acidification research (Harvey et al. 2013; Kroeker et al. 2013; Brown et al. 2018; Clements & Darrow 2018) and is appropriate for both continuous and ratio-type (i.e., proportions and percentages) response variable data that are commonly used in behavioural studies (Hintze 2007; Pustejovsky 2018). Effect size variance was calculated as:

$$v = \frac{(S_E)^2}{n_E \bar{X}_E^2} + \frac{(S_C)^2}{n_C \bar{X}_C^2}$$

where *S* and *n* are the standard deviation and sample size, respectively, for a given experimental treatment (denoted by the subscripts C [control] and E [experimental, i.e., elevated pCO_2]); $\bar{X}_E$ and $\bar{X}_C$ are defined as above. We chose *lnRR* because it 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 (Hintze 2007; Pustejovsky 2018) (while other effect sizes incorporating variance into their calculations are not due to different variance structures of proportion and percentage data). 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 lateralization 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.

Individual effect sizes and their associated variance were obtained for each included measurement from each study using the *metafor* package (Veichbauer 2010) in R v. 3.5.1 (R Core

Team 2018). 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 can make prey fish more difficult for predators to capture, but can 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 test for the decline effect (hereafter 'absolute effect size'). 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 (Paulus et al. 2013), the use of absolute effect size values is informative for understanding the strength of effects ignoring directionality (Garamszegi et al. 2006).

Meta-analysis

Testing for the decline effect

To assess whether or not ocean acidification research on fish behaviour is characterized by the decline effect, we used two analytical approaches. First, we assessed the relationship between the proportion of articles reporting a 'Strong Effect' (see definition above) of acidification on fish behaviour over time (time = publication year; defined as the year in which a given article was first published online and made available to the scientific community). For this approach, the decline effect would be evidenced by a negative relationship between 'Strong Effect' proportion and time.

Second, we assessed the relationship between mean absolute *lnRR* as a function of time (publication year as defined above). For this analysis, mean effect sizes for each year (2009–2019) and their associated variance were derived from weighted random effects models in *metafor*, which give a higher weighting to studies with higher sample sizes and lower variance (Hedges & Olkin 1985) (see individual effect size variance formula above). We accounted for non-independence associated with multiple data points from a single study by using three-level meta-analytical models (Nakagawaa et al. 2015; Noble et al, 2017) to calculate mean effect sizes, including 'measure nested within study' as a random variable. Like the first analytical approach, the decline effect would be evidenced by a negative relationship between mean absolute *lnRR* and time. A handful of individual effect sizes (n = 13 of 785) were omitted from weighted mean effect size computations due to outstandingly large variance estimates, which preclude *metafor* from calculating mean effect sizes for a category of interest; individual effects sizes with a variance estimate >10 were excluded and all such instances are highlighted in the 'Notes' column of Supplementary File 2.

Explaining the decline effect

Since a decline effect was detected in our analysis, we explored two potential explanatory factors that might drive the observed effect: 1. Biological explanations including climatic region and the presence/absence of cues or stimuli; 2. studies with small sample sizes exhibiting larger effects than those with larger sample sizes, and 3. publication bias due to high impact journals publishing large effects.

Biological explanations

If observed, the decline effect could potentially be driven by two biological characteristics of the studies included in the analysis. First, an increasing number of studies on temperate and/or cold-water species could explain the decline effect if the number of such studies have increased over time and if temperate species are tolerant to ocean acidification (while tropical and subtropical species in the early studies are sensitive). Second, the decline effect could be explained by an increasing number of studies measuring baseline behaviours in the absence of a behavioural stimulus, if baseline behaviours are not altered by acidification but behaviours requiring a stimulus or cue are (which are characteristic of early studies). To account for the 'climate region' explanation, we simply excluded temperate and cold-water species from the dataset and tested whether or not the decline effect persisted for subtropical and tropical species only. Climate region was obtained from Fishbase (Froese & Pauly 2019) for each species; if a species was not found in FishBase then the climate region was obtained directly from the article. Similarly, to account for the 'no stimulus' explanation, we determined whether or not each experiment in each article included a stimulus, removed those that did not contain a stimulus from the dataset, and tested whether or not the decline effect persisted when only behaviours in the presence of a stimulus were included. If the decline effect persisted when cold-water species and experiments without a stimulus were removed, this would indicate that the decline effect could not be explained by these two biological variables.

Sample size

Correlations between sampling effort and effect size can be indicative of observer bias[15]. Herein, if large effects are only observed when sample sizes are low, it is probable that the observed large effects may be false positives (i.e., are driven by Type I Error). Thus, if observer bias was driving a decline effect, we would predict two things: 1. the strongest effects being observed when sample sizes are low; and 2. a positive relationship between sample size and time (publication year). For 1., we assessed the relationship between the mean effect size for each study and the average sample size for that study. Average sample size was calculated as the average of all sample sizes across treatments and was used because individual studies often had varied sample sizes between experiments or treatments. Additionally, for 1., we calculated weighted mean effect sizes (absolute *lnRR* as above) for sample size bins (0-9.99, 10-19.99, 20-29.99, … 70-79.99, 80+) to determine which categories of sample size had mean effect sizes statistically different from 0 (see the <u>Statistical analysis</u> section below). For 2., we calculated the average sample size for each publication year and assessed the relationship between average sample size and time. In addition, if 1. was true from the data, we calculated the proportion of articles having a sample size above an observed threshold of sample size whereby extreme and significant effects no longer occurred. We then assessed the relationship between publication year and the proportion of articles at or above that threshold.

<u>Publication bias driven by larger effects in high-impact journals</u>

In new and emerging fields, the early inflation of effect sizes can be driven by publication bias resulting from the tendency for high-impact journals to publish novel and ground-breaking results showing strong and seemingly undisputable effects (Sterne et al. 2001). If this were true for our analysis, two things would be evident: 1. Higher impact journals would have higher mean effect sizes; and 2. there would be a negative relationship between mean impact factor and time (publication year). We therefore explored both of these relationships to provide evidence for or against the idea that the decline effect could be driven by publication bias due to initial large effects in high-impact journals. For 1., we derived mean *lnRR* (mean of study-specific averages, as above) for each of 11 impact factor bins: 0–0.99, 1–1.99, 2–2.99, … , 9–9.99, and 10+, and assessed the relationship between effect size and impact factor. For 2., we calculated the average journal impact factor for each year and assessed the relationship between impact factor and time; 2017 impact factors were used for studies published in 2018 and 2019 because 2018 and 2019 impact factors were unavailable at the time of analysis. For both relationships, impact factor was defined as the journal impact factor for a given article during the year that it was published online.

Statistical analysis

For all categorical analyses using mean effect sizes (absolute *lnRR*), effect sizes were deemed statistically significant from 0 if their 95% CI did not overlap with zero. Note, however, that statistical significance needs to be interpreted with caution, as using absolute effect sizes (i.e., unsigned, positive effect sizes) results in an overestimate of the true effect size.

References

Allen, C. & Mehler, D.M.A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS Biol.* **17**, e3000246.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454.

Bender, D., Champ, C.M., Kline, D., Diaz-Pulido, G. & Dove, S. (2015). Effects of "Reduced" and "Business-As-Usual" CO_2 emission scenarios on the algal territories of the damselfish *Pomacentrus wardi* (Pomacentridae). *PLoS One* **10** e0131442.

Black, R. (2011). Acid oceans turn 'Finding Nemo' deaf. *BBC News* https://www.bbc.com/news/science-environment-13605113.

Browman, H.I. (2016). Applying organized scepticism to ocean acidification research. *ICES J. Mar. Sci.* **73**, 529–536.

Brown, N.E., Bernhardt, J.R., Anderson, K.M. & Harley, C.D.G. (2018). Increased food supply mitigates ocean acidification effects on calcification but exacerbates effects on growth. *Sci. Rep.* **8**, 9800.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T-H., Johannesson, M., Kirchler, M. *et al.* (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behav.* **2**, 637–644.

Clark, T.D. (2017). Science, lies and video-taped experiments. *Nature* 542, 139–139.

Clark, T.D., Raby, G.D., Roche, D.G., Binning, S.A., Speers-Roesch, B., Jutfelt, F. & Sundin, J. (2020). Ocean acidification does not impair the behaviour of coral reef fishes. *Nature* **577**, 370–375.

Clements, J.C. & Comeau, L.A. (2019) Use of high frequency, non-invasive electromagnetic biosensors to detect ocean acidification effects on shellfish behaviour. *J. Shellfish Res.* **38**, 811–818.

Clements, J.C. & Darrow, E.S. (2018). Eating in an acidifying ocean: a quantitative review of elevated CO_2 effects on the feeding rates of calcifying marine invertebrates. *Hydrobiologia* **820**, 1–21.

Columb, M.O. & Atkinson, M.S. (2016). Statistical analysis: sample size and power estimations. *BJA Educ.* **16**, 159–161.

Noble, D.W., Lagisz, M., O'dea, R.E. & Nakagawa, S. (2017). Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. *Molec. Ecol.* **26**, 2410–2425.

Dell, A.I., Bender, J.A., Branson, K., Couzin, I.D., de Polavieja, G.G., Noldus, L.P.J.J. *et al.* (2014). Automated image-based tracking and its application in ecology. *Trends Ecol. Evol.* **29**, 417–428.

Devine, B.M. & Munday, P.L. (2013). Habitat preferences of coral-associated fishes are altered by short-term exposure to elevated CO_2. *Mar. Biol.* **160**, 1955–1962.

Dixson, D.L. (2017). Increasingly acidic oceans are causing fish to behave badly. *Sci. Am.* **316**, 42–45.

Dixson, D.L., Munday, P.L. & Jones, G.P. (2010). Ocean acidification disrupts the innate ability of fish to detect predator olfactory cues. *Ecol. Lett.* **13**, 68–75.

Duarte, C.M., Fulweiler, R.W., Lovelock, C.E., Martinetto, P., Saunders, M.I., Pandolfi, J.M. *et al.* (2015). Reconsidering ocean calamities. *BioScience* **65**, 130–139.
Froese, R. & Pauly, D. (2019). FishBase. www.fishbase.org
Garamszegi, L.Z. (2006). Comparing effect sizes across variables: generalization without the need for Bonferroni correction. *Behav. Ecol.* **17**, 682–687.
Gonzales, J.E. & Cunningham, C.A. (2015). The promise of pre-registration in psychological research. *American Psychological Association* https://www.apa.org/science/about/psa/2015/08/pre-registration.
Harvey, B.P., Gwynn-Jones, D. & Moore, P.J. (2013). Meta-analysis reveals complex marine biological responses to the interactive effects of ocean acidification and warming. *Ecol. Evol* **3**, 1016–1030.
Head, M.L., Holman, L., Lanfear, R., Kahn, A.T. & Jennions, M.D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol.* **13**, e1002106.
Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Academic Press, New York, NY.
Hintze, J.L. (2007) *NCSS statistical system user's guide IV: Multivariate analysis, clustering, metaanalysis, forecasting / time series, operations research, mass appraisal.* NCSS, Kaysville, UT.
Hofseth, L.J. (2018). Getting rigorous with scientific rigor. *Carcinogenesis* **39**, 21–25.
Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Med.* **2**, e124.
Jutfelt, F., Sundin, J., Raby, G.D., Krång, A.-S. & Clark, T.D. (2017). Two-current choice flumes for testing avoidance and preference in aquatic animals. *Methods Ecol. Evol.* **8**, 379–390.
Kroeker, K.J., Kordas, R.L., Crim, R., Hendriks, I.E., Ramajo, L., Singh, G.S. *et al.* (2013). Impacts of ocean acidification on marine organisms: quantifying sensitivities and interaction with warming. *Glob. Change Biol.* **19**, 1884–1896.
Kroeker, K.J., Kordas, R.L., Crim, R.N. & Singh, G.G. (2010). Meta-analysis reveals negative yet variable effects of ocean acidification on marine organisms: Biological responses to ocean acidification. *Ecol. Lett.* **13**, 1419–1434.
Lönnstedt, O.M., Munday, P.L., McCormick, M.I., Ferrari, M.C.O. & Chivers, D.P. (2013). Ocean acidification and responses to predators: can sensory redundancy reduce the apparent impacts of elevated CO_2 on fish? *Ecol. Evol.* **3**, 3565–3575.
Maneja, R.H., Frommel, A.Y., Browman, H.I., Clemmesen, C., Geffen, A.J., Folkvord, A. *et al.* (2012). The swimming kinematics of larval Atlantic cod, *Gadus morhua* L., are resilient to elevated seawater pCO_2. *Mar. Biol.* **160**, 1963–1972.
Marsh, D.M. & Hanlon, T.J. (2007). Seeing what we want to see: Confirmation bias in animal behaviour research. *Ethology* **113**, 1089–1098.
Munday, P.L., Cheal, A.J., Dixson, D.L., Rummer, J.L. & Fabricius, K.E. (2014). Behavioural impairment in reef fishes caused by ocean acidification at CO_2 seeps. *Nat. Clim. Change* **4**, 487–492.
Munday, P.L., Dixson, D.L., Donelson, J.M., Jones, G.P., Pratchett, M.S., Devitsina, G.V. & Døving, K.B. (2009). Ocean acidification impairs olfactory discrimination and homing ability of a marine fish. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1848–1852.

Munday, P.L., Dixson, D.L., McCormick, M.I., Meekan, M., Ferrari, M.C.O. & Chivers, D.P. (2010). Replenishment of fish populations is threatened by ocean acidification. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12930–12934.

Munday, P.L., Pratchett, M.S., Dixson, D.L., Donelson, J.M., Endo, G.G.K., Reynolds, A.D. & Knuckey, R. (2013). Elevated CO_2 affects the behavior of an ecologically and economically important coral reef fish. *Mar. Biol.* **160**, 2137–2144.

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L.Lagisz, M. & Senior, A.M. (2015). Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods Ecol. Evol.* **6**, 143–152.

Paulus, F.M., Krach, S., Albrecht, A.-G. & Jansen, A. (2013). Potential bias in meta-analyses of effect sizes in imaging genetics. *Schizophr. Bull.* **39**, 501–503.

Pimentel, M.S., Faleiro, F, Marques, T., Bispo, R., Dionísio, G., Faria, A.M. *et al*. (2016). Foraging behaviour, swimming performance and malformations of early stages of commercially important fishes under ocean acidification and warming. *Clim. Change* **137**, 495–509.

Poulton, D.A., Porteus, C.S. & Simpson, S.D. (2017). Combined impacts of elevated CO_2 and anthropogenic noise on European sea bass (*Dicentrarchus labrax*). *ICES J. Mar. Sci.* **74**, 1230–1236.

Pustejovsky, J.E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *J. School Psychol.* **68**, 99–112.

R Core Team. (2018) *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

Roberts, K.B. (2015). Sea change: UD's Dixson discusses ocean acidification at White House briefing. *UDaily* http://www1.udel.edu/udaily/2016/dec/ocean-acidification-120415.html.

Rodriguez-Dominguez, A., Connell, S.D., Baziret, C. & Nagelkerken, I. (2018). Irreversible behavioural impairment of fish starts early: Embryonic exposure to ocean acidification. *Mar. Pollut. Bull.* **133**, 562–567.

Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature* **470**, 437.

Schunter, C., Welch, M.J., Nilsson, G.E., Rummer, J.L., Munday, P.L. & Ravasi, T. (2018). An interplay between plasticity and parental phenotype determines impacts of ocean acidification on a reef fish. *Nat. Ecol. Evol.* **2**, 334–342.

Schunter, C., Welch, M.J., Ryu, T., Zhang, H., Berumen, M.L., Nilsson, G.E. *et al.* (2016). Molecular signatures of transgenerational response to ocean acidification in a species of reef fish. *Nat. Clim. Change* **6**, 1014–1018.

Sterne, J., Egger, M. & Smith, G.D. (2001). Investigating and dealing with publication and other biases. In: *Systematic reviews in health care: Meta-analysis in context*, Second Edition [eds Egger, M., Smith, G.D. & Altman, D.G.] BMJ Publishing Group, London, pp. 189–208.

Traniello, J.F.A. & Bakker, T.C.M. (2015). Minimizing observer bias in behavioural research: blinded methods reporting requirements for *Behavioural Ecology and Sociobiology*. *Behav. Ecol. Sociobiol.* **69**, 1573–1574.

Veichbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48.

Yong, E. (2009). Losing Nemo - Acid oceans prevent baby clownfish from finding home. *Discover Magazine* http://blogs.discovermagazine.com/notrocketscience/2009/02/02/losing-nemo-acid-oceans-prevent-baby-clownfish-from-finding-home/#.XcW48zJKhTZ.