A framework for practical and rigorous impact monitoring by field managers of marine protected areas

Anthony B. Rouphael · Ameer Abdulla · Yasser Said

Received: 17 July 2010 / Accepted: 22 November 2010 © Springer Science+Business Media B.V. 2010

Abstract Monitoring is a crucial component of conservation in marine protected areas (MPAs) as it allows managers to detect changes to biodiversity and to infer cause of change. However, the complexities of sampling designs and associated statistical analyses can impede implementation of monitoring by managers. Two monitoring frameworks commonly used in marine environments are statistical testing and parameter estimation. For many managers these two approaches fail to help them detect change and infer causation for one or more reasons: the complexity of the statistical test, no decision-making structure and a sampling design that is suboptimal. In collaboration with marine park rangers in Egypt, we instigated a monitoring framework to detect impacts by snorkelers in a pragmatic but scientifically rigorous way. First, we used a literature review to define causal criteria to facilitate inference. This was essential because our sampling design was suboptimal due to a lack of baseline data and there was only one impact site. Second, we established a threshold level of coral damage that if exceeded would trigger management to reduce the impact of snorkelers. This provided a clear decision-making structure. Third, we estimated effect sizes with confidence intervals to detect change. For the field managers, this approach to detection was easier to understand than assessing a null hypothesis and provided critical information for decision making. At no stage during the short study period did snorkelers cause damage that exceeded the threshold and thus mitigation was not required. In situations of technical and financial constraints this framework will increase the implementation of effective impact monitoring for many activities in MPAs and enhance management of marine biodiversity.

Keywords Effect size · Levels-of-evidence · Marine protected areas · Monitoring framework · Null hypothesis testing · Parameter estimation · Sampling design
Introduction

Background

The majority of nations are signatories to the Convention of Biological Diversity and its Work Program on Protected Areas, which requires governments to take action to conserve biodiversity, improve fish yields, support tourism and protect cultural values (De Fontaubert et al. 1996). Field managers of Marine Protected Areas (MPAs) can protect species and habitats by introducing fishing gear restrictions, excluding damaging activities from sensitive areas, restricting visitor numbers and modifying visitor behaviour (Hawkins and Roberts 1994; Marion and Rogers 1994; Salm and Clark 2000). Increasingly, government and funding agencies are demanding that managers demonstrate they are meeting their management objective in relation to biodiversity conservation (Hocking et al. 2000). Monitoring can help managers achieve this objective by detecting change in biodiversity and to identify cause of change (Fabricius and De’ath 2004; Underwood 1989). Only with this information can managers act decisively to prevent negative impacts, such as a decline in habitat area, or reliably conclude management is having a positive impact on, for example, abundances of exploited species.

Unfortunately, detecting and inferring cause of change to biodiversity in MPAs is not simple. Many potentially destructive human activities are legally permitted in some MPAs (Day 2002; Preen 1998) and it is not always clear how biodiversity responds to these activities (Hatcher et al. 1989). Human impact to marine biodiversity is rarely diagnostic because it cannot always be distinguished from impact caused by natural agents of disturbance (Riegl and Velimirov 1991). In addition, populations of species and distribution of habitats are naturally variable (Connell and Sousa 1983; Connell et al. 1997; Hatcher et al. 1989). Consequently, separating the effects of humans to biodiversity from natural variability is not straightforward (Green 1979; Osenberg and Schmitt 1996; Underwood 1996). Indicator variables have been promoted as a pragmatic approach for dealing with these challenges in MPAs (Pomeroy et al. 2005, 2007). Pomeroy et al. (2007) defined an indicator as “a unit of information measured over time that allows you to document changes in specific attributes of your MPA”. Used in isolation, however, indicator variables have major limitations to interpretation (Hatcher et al. 1989; Rouphael and Hanafy 2007; Underwood and Chapman 1999). Foremost are the unlikely assumptions that ecological variables are invariant prior to disturbances and change predictably in response to human activity (Marsh 1995; Oliver 1995). Hence, simply measuring an indicator variable is unlikely to allow managers to determine with confidence the cause of change.

A more scientifically robust approach to detect change in biodiversity and to infer causation is the use of formal scientific frameworks (Fabricius and De’ath 2004; Underwood 2000). There are two frameworks commonly used to monitor biodiversity in marine environments (Benedetti-Cecchi 2001). The first is point-source null hypothesis testing (hereafter significance testing) combined with a before/after, control/impact (BACI) sampling design (Green 1979; Underwood 2000). Under more controlled situations, a true experimental design involving random assignment of treatments to replicate experimental units might be available (Manly 2001). With this framework, detection of change is based on a probabilistic test used to falsify a null hypothesis. Interpretation is based on how a response variable at impact locations changed from before to after the start of a disturbance, and in relation to control locations. Within the second framework, detection is based on estimating the difference in a response variable between impact and control locations with a measure of uncertainty such as a confidence interval (CI) (Beyers 1998; Fabricius and De’ath 2004; Suter 1996). Within this framework (hereafter parameter estimation), the difference between locations can be assessed irrespective of statistical significance. Proponents of this framework recommend that the cause of impact be inferred using multiple lines-of-evidence (Beyers 1998; Fabricius and De’ath 2004; Suter 1996), which is a process of making a conclusion based on several pieces of circumstantial evidence.

In this paper, we begin by comparing the needs of MPA managers in regards to biodiversity
impact monitoring. We then review statistical testing and parameter estimation, focusing on their strengths and weaknesses in context of the needs of managers. We believe that on their own each will not always be successful in many situations relating to MPAs and may be detrimental if managers do not understand the statistical approaches or where resources are invested in poorly designed protocols with limited capacity for inference and decision making. We propose a monitoring framework for managers that combine the strengths of significance testing and parameter estimation while minimising the risk of inappropriate implementation. We describe the characteristics of this framework and key steps to implement it, and use a case study from an Egyptian MPA to illustrate how we used this monitoring framework. We believe this framework will increase the implementation of effective impact monitoring for a range of activities in MPAs and enhance management of marine biodiversity.

Managers and monitoring frameworks

In proposing this framework, we focused on the needs of field managers to undertake biodiversity impact monitoring in MPAs. Field managers face challenges in implementing monitoring programs due to limited access to technical expertise, financial constraints and staff limitations (Abdulla et al. 2008; Buckley et al. 2008; Marshall and Abdulla 2009; Pomeroy et al. 2005). We make the distinction between scientists and managers of MPAs because the latter are not always marine scientists (Alder 1996). Instead, many come to the role as environmental generalists or administrators and do not always have a strong background in marine ecology, sampling design and inferential statistics. Unlike scientists, field managers will rarely be in a position to make decisions based on the acceptance or rejection of a null hypothesis (Stewart-Oaten 1996). Further, rarely will they have the authority to stop a human activity legally permitted in a MPA. Instead, they are best positioned to contribute evidence of impact to members of a management board to make decisions on when or how to respond to threats to biodiversity.

Monitoring frameworks currently in use

Strengths and weaknesses of significance testing and parameter estimation have been debated extensively previously (Di Stefano 2004; Fabricius and De’ath 2004; Suter 1996; Underwood 2000), but not in context of the needs of field managers of MPAs. This section provides the basis for why we believe both frameworks, used individually, will commonly fail to meet the needs of managers of MPAs.

Significance testing

As stated earlier, with this framework detection of change is based on test of a null hypothesis. As an example, a manager predicts (proposes a hypothesis) that the mean length of grouper (e.g. *Epinephelus* sp.) in areas where commercial fishing is permitted will be equal or less than that in adjacent no-take zones. This is based on the theory that fishing removes many large groupers, reducing the mean length of grouper in the fishing grounds and no-take zones are effective at preventing fishing-related mortality (Russ and Alcala 1989). The mean length of grouper is compared between areas open to fishing and replicate no-take zones, which share similar environmental conditions but fishing is not permitted. An analysis of variance (ANOVA) is performed and the null hypothesis, a prediction that there is no difference in mean length, is rejected with 95% confidence. This result provides support for the original prediction (Underwood 1997), which in turn gives us confidence in our theory relating to the effectiveness of no-take zones.

A major strength of this approach is that it provides a clear structure for evaluating competing theories, which are the basis of predictions, and forces managers to consider carefully, before initiating monitoring, how data will be interpreted (Underwood 1997). When the statistical test assumptions are met, managers can predict, with a specified level of confidence, the risk of rejecting a true null hypothesis. Also, with power analysis, the error of accepting a false null hypothesis can be controlled (Cohen 1988). Another benefit is that multivariate approaches to null hypothesis testing (Anderson 2001; Clarke 1993) can
conveniently summarise data comprising large numbers of response variables (Wonnacott 1987).

Unfortunately, significance testing poses difficulties for managers. Foremost is the mechanics behind the statistical approaches and concepts (e.g. ANOVA, statistical power, P values and F statistic) are easily misunderstood by inexperienced practitioners (Perry 1986; Suter 1996; Yoccoz 1991). Selecting appropriate statistical models and ensuring the correct error terms are used in the F ratio tests are not always easy. Calculating power for statistical models involving combinations of fixed, random and nested terms is complicated (Benedetti-Cecchi 2001; Underwood 1997). The need to transform data to fulfil assumptions of parametric approaches can further complicate the analysis and interpretation (Bence et al. 1996; Clarke et al. 2006; McDonald et al. 2000). Also, with frequent monitoring and multiple tests of related response variables there is the risk of inflating the Type I error rate, further complicating interpretation (Manly 2001; Underwood 1997).

Even for managers familiar with this approach, statistical significance is routinely equated with management significance and non-significant results dismissed as ecologically unimportant (Stewart-Oaten 1996; Wonnacott 1987). These beliefs are misguided because statistical significance is, in part, a function of sample size (Johnson 1999; Wonnacott 1987). That is, statistically significant differences are more likely to be detected between two or more treatments (locations or populations) if replication is large and less likely, for the same treatments, if replication is small (Yoccoz 1991). Many, if not most, monitoring studies in MPAs involve inadequate replication, which greatly limits the capacity of managers to detect impacts (Abdulla 2000; Abdulla et al. 2008). Further, with an emphasis on the P value, managers commonly avoid discussing effect size, which is an essential piece of information for interpreting data and making management decisions (Stewart-Oaten 1996). An effect size can be defined quantitatively in a number of ways (Keppel 1991), but the simplest definition is the raw difference score between two treatment means ($\bar{x}_1 - \bar{x}_2$). To illustrate, if a survey of the gastropod Trochus in replicate no-take zones and areas open to harvesting reported 18/200 m$^2$ and 10/200 m$^2$, respectively, the effect size between the two treatments is eight Trochus.

With significance testing, the ability to conclude reliably the cause of an observed change in biodiversity is typically gained through the way in which measurements/samples are collected in space and time (the sampling design) (Underwood 1997). A randomised experimental sampling design is considered one of the most robust approaches to unambiguously infer cause and effect relationships in ecological systems (Downes et al. 2002; Underwood 1997). With an experimental sampling design, measurements are taken at a number of times at locations before the application of a treatment, say snorkeling or fishing, and at a number of locations, where the treatment is not applied. Data collected prior to the application of the treatment are called baseline data. The locations where samples are collected, but do not receive the treatment are called controls. Control locations are chosen to be as similar as possible in all respects to the impact location, except for the presence of the putative impact (Downes et al. 2002). Data obtained during the baseline survey and from controls are used to describe variability in a response variable outside the influence of the treatment. Importantly, the treatment should be applied to randomly selected experimental units (e.g. locations, reefs or zones) in order to factor out natural processes that may be influencing the response variable in the study area. Further, more than one location receiving the treatment and one control location are necessary in order to develop precise estimates of the treatment effect and control condition, and so that individual location differences are not confused with the treatment effect. Unfortunately, one or more components of an experimental design will usually be absent when managers assess the effects of human activities in MPA (Abdulla 2000; Abdulla et al. 2008). Treatment effects are not randomly allocated to experimental units because the location and size of MPAs are still largely influenced by the socio-economic goals of fishermen (Halpern and Warner 2003; McNeill 1994) or divers are taken to locations determined by aesthetic, economic and safety considerations (Tabata 1989). In other situations there may be no baseline data because
ecological monitoring was not undertaken prior to the establishment of a MPA (Francis et al. 2002). In addition, there may be no replication of treatment effects because there is, for example, a single no-take zone.

Parameter estimation

With this framework, detecting change is based on measuring effect sizes with some measure of uncertainty (Fabricius and De’ath 2004; Stewart-Oaten 1996; Suter 1996). To illustrate, mean densities of seagrass leaves are estimated and compared through time from two bays with mooring buoys and two bays where anchoring is permitted. A CI is calculated for each mean, allowing the assessment of its level of precision. Precision relates to the confidence we have in an estimate of a parameter (Andrew and Mapstone 1987), with wide CIs equating to low precision. More formally, a CI is defined as a range of plausible values for the parameter being estimated (Cumming and Finch 2005). For each survey site, the effect size (±CI) is then interpreted. A strength of this framework is that estimation of effect size focuses on the magnitude of treatment differences, rather than on the probability that differences were due to random chance (Suter 1996). In addition, an effect size can be understood by even those not trained in statistics. However, a weakness with parameter estimation is that when done in the absence of a decision-making structure there is no clear guidance on when managers should act based on the results of their monitoring (Underwood 1997, 2000). Further, without specifying an effect size considered biologically important, it is difficult to a priori determine the number of replicates needed to obtain precise estimates of means, effect sizes and other parameters (Andrew and Mapstone 1987).

With the parameter estimation framework, inference is developed based on carefully structured arguments, a technique referred to as weight of evidence (Suter 1996), causal argument (Beyers 1998; Fabricius and De’ath 2004) or levels-of-evidence (Downes et al. 2002; McArdle 1996). This approach has been used successfully in disciplines where manipulative experimentation is unlikely for ethical reasons, such as assessing the effects of diseases on humans. Its formal use in ecological impact assessment is uncommon and comparatively recent (Beyers 1998; Downes et al. 2002; Fabricius and De’ath 2004). Hill (1965) categorised different types of causal argument into nine criteria for studies into the effects of diseases on humans. Table 1 lists each of Hill’s causal criterions and how they relate to ecological impact assessment (Beyers 1998). With levels-of-evidence there is a need to seek evidence not only to support the impact prediction, but evidence to rule out plausible alternative predictions, such as that the observed difference was due to natural processes (Beyers 1998; Downes et al. 2002).

A strength of levels-of-evidence is that it provides a highly structured approach to facilitate inference (Beyers 1998; Downes et al. 2002). Fabricius and De’ath (2004) also argued that it is transparent and easy for decision makers to understand. A weakness of this method is that the evidence is circumstantial because it is based on correlations (Downes et al. 2002), which does not necessarily imply causation.

Table 1  Hill’s causal criteria and description in context to ecological impact assessment (after Beyers 1998)

<table>
<thead>
<tr>
<th>Causal criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strength of association</td>
<td>A large proportion of individuals are effected in the exposed area relative to control areas</td>
</tr>
<tr>
<td>Consistency of association</td>
<td>The association has been observed by other investigators at other times and places</td>
</tr>
<tr>
<td>Specificity of association</td>
<td>The effect is diagnostic of exposure</td>
</tr>
<tr>
<td>Temporality</td>
<td>Exposure must precede the effect in time</td>
</tr>
<tr>
<td>Biological gradient</td>
<td>The risk of effect is a function of magnitude of exposure</td>
</tr>
<tr>
<td>Biological plausibility</td>
<td>A plausible mechanism of action links cause and effect</td>
</tr>
<tr>
<td>Experimental evidence</td>
<td>A valid experiment provides strong evidence of causation</td>
</tr>
<tr>
<td>Coherence</td>
<td>Similar stressors cause similar effects</td>
</tr>
<tr>
<td>Analogy</td>
<td>The causal hypothesis does not conflict with existing knowledge of natural history and biology</td>
</tr>
</tbody>
</table>
acknowledge that each causal argument is weak independently, but argue that when combined may provide strong support for a conclusion (Downes et al. 2002). However, rarely will all criterion, listed in Table 1, be useful for any one monitoring program. For instance, the criterion specificity of association will not apply unless the assessment relates to an activity that has a unique effect on the environment.

A monitoring framework for managers

Although both frameworks have limitations for many managers of MPAs, elements of both can be combined to form a monitoring framework that will meet the needs of managers yet remain scientifically robust. We propose three key components of an impact monitoring framework for field managers. First, detecting change in biodiversity

---

**Fig. 1** Key steps to implement the proposed monitoring framework

---

**Step 1: Pre-monitoring planning**
- Define monitoring objective(s) and select response variables
- Use literature review to define threshold to be used as early warning trigger for management action
- Develop sampling design and state its limitation for inference
- Use literature review to establish level-of-evidence by defining causal criteria
- Consult and agree with stakeholders on the threshold and causal criteria

**Step 2: Start monitoring**
- Undertake monitoring and measure response variables
- Calculate effect sizes (e.g., difference between the putative impact and control sites) and confidence intervals during each survey period

**Step 3: Compare effect size(s) with threshold during each survey period**

- **Step 3a: Effect size threshold not exceeded**
  - Continue monitoring if activity is ongoing

- **Step 3b: Effect size threshold exceeded**
  - Evaluate causal criteria
  - Forward all evidence, supportive and unsupportive of impact to management board

**Step 5: Action following impact assessment**
- If management board concludes impact, instigate management action to alleviate impact or enhance recovery and continue monitoring.
- If management board concludes that the impact is not ecologically important review threshold level and continue monitoring

**Step 4: Assessment**
- Management board evaluates evidence for impact (taking into account precision of the data)
  - and / or
- Commission an experienced marine scientist to provide a more detailed impact assessment
should be based on estimating the magnitude of an effect (an effect size) between impact and control locations. An effect size is a variable that is easy to interpret by non-specialists and a CI is an intuitive way to communicate the level of uncertainty associated with an effect size. Second, inference should be based on both a sampling design and levels-of-evidence. Although we believe the sampling design should be the foundation for inference, a randomised experimental design will normally be unavailable to managers for reasons already described. Therefore, inference will need to be supported using levels-of-evidence. Third, a management threshold should be used to determine when action is required to mitigate potential deleterious change in biodiversity. A threshold is a pre-determined effect size that if exceeded during monitoring would result in the monitoring data and other evidence of impact being reported to a management board or other decision-making group. Managers need to state the level of change to biodiversity from human activity that would constitute a concern for management (Oliver 1995; Stankey et al. 1985; Underwood 2000). Theoretically, early-warning of detrimental change would allow decision makers sufficient time to assess the ecological, social and economic importance of an impact and to consider a suitable management response. These three features provide the basis for a monitoring framework that can detect change early, is inferentially robust, has minimal risk of wrong implementation and the results can be communicated easily to non-specialists. We present a case study where we have trialled this framework in order to illustrate how these three key features were combined to develop a monitoring framework to assess if snorkeling at an Egyptian coral reef was leading to an unacceptable level of damage to corals.

To assist the reader to conceptualise the approach we took, steps taken to implement the framework are shown in Fig. 1. Most of these steps are consistent with robust monitoring programs and are explained in detail elsewhere (Downes et al. 2002; Field et al. 2007; Green 1979; Keough and Mapstone 1995), so are not repeated here. For purposes of this paper, we focus only on those steps used to detect change, facilitate inference and to determine when the monitoring data should be reported to decision makers.

**Methods**

**The case study and location**

Snorkeling, a popular recreational activity in the Egyptian Red Sea, has been associated with impacts to reefs worldwide (Allison 1996; Harriott 2002). It is a potential source of chronic mechanical damage to corals in shallow water (Plathong et al. 2000). Monitoring the effects of the snorkeling in Egyptian MPAs is important because it is one of the fastest growing recreational activities in the Egyptian Red Sea (Hawkins and Roberts 1994). To trial the monitoring framework, we undertook a study at a single coastal resort near Wadi Gemal Hamata National Park (WGHNP). The study area was a section of fringing reef immediately adjacent to the Shams Alam Resort, about 300 km south of Hurghada, Egypt. The resort lies at the northern boundary of the WGHNP. Preliminary observations revealed that snorkeling was intense on the fringing reef adjacent to the resort (hereafter the snorkel location), but less so at reefs >300 m either side of it.

**Sampling design, sampling approach and response variable**

We were greatly constrained in the type of sampling design available. A complete experimental sampling design was impossible because the treatment (i.e. snorkeling) was not randomly assigned to reefs and there was only one snorkel (impact) location. A BACI design was impossible because there was no opportunity to collect baseline data prior to the commencement of snorkeling. However, multiple control locations were established, which improved our capacity to separate the effects of snorkeling on corals from that caused by natural agents of disturbance (Underwood 1993, 1996). We established two controls located 500 m and 1 km south of the resort. To assess the
relationship between the intensity of snorkeling and amount of damage, the snorkel location was
stratified into three zones of increasing distance from where snorkelers entered the water. Four
replicate transects were placed near the entry point, and at 100 and 200 m further along the reef.
Transects were placed parallel to the reef crest in water depths ranging from 0.5 to 3 m. Tran-
sects were placed in the controls in the same way. Sampling was limited to the reef crest because
we predicted that only corals in this habitat were at risk from snorkelers because the water depth
over the reef flat was normally too shallow for snorkelers to access. Further, impacts were pre-
dicted not to extend 3 m below sea level because this is about the maximum depth at which fins are
likely to contact the reef when snorkelers tread water (Plathong et al. 2000). A modified version
of the point intercept transect method (Musso and Inglis 1998) was used in which divers recorded
the substratum type beneath points spaced 50 cm along each transect (a 20 m surveyor tape) laid on
the seafloor. When a point overlayed a coral with a branching growth form, we estimated the pro-
portion of branches that were broken. We focused on measuring physical damage to branching corals
as the primary response variable because this growth-form is most sensitive to mechanical dis-
turbance (Marshall 2000) and on the assumption that an increase in physical damage to a colony
would be a pre-cursor to whole colony mortality (Rouphael and Hanafy 2007). We collected data
for this study during a 3-month period, with sam-
pling undertaken once per month during the first
week of March, April and May 2007. We chose
a sampling frequency of 4 weeks as snorkeling
has potential to lead to a rapid increase in coral
damage (Plathong et al. 2000). A lack of funding
did not permit this study to continue beyond May
2007.

Defining causal criteria

Consistent with levels-of-evidence, we used causal
criteria to increase our capacity to infer that
snorkeling, not a natural agent of disturbance, was
the cause for an observed breach of our man-
agement threshold (defined below). A literature
review of snorkeling impact studies was under-
taken to help us define causal arguments for and
against the conclusion that snorkeling was the
cause of elevated levels of coral damage at the
snorkel location. We did this by constructing a
table with four columns. The first column listed
Hill’s nine causal criteria. In the second column,
we summarised findings of papers relevant to each
criterion. For example, with the criterion bio-
logical gradient we summarised what researchers had
observed in relation to the level of coral dam-
age with the intensity of snorkeling activity. In
the third and fourth columns, we described the
expected outcomes during our monitoring pro-
gram that would or would not support the notion
that snorkeling had caused the exceedance of the
threshold, respectively. Following Downes et al.
(2002), we defined each criterion before the start
of monitoring because seeking evidence during
or after the completion of monitoring to inter-
pret results is more likely to contribute to biased
conclusions.

Defining a management threshold

A literature review was used to help us define a
management threshold. The goal was to define a
quantitative threshold for the response variable
that would act as an early warning trigger of po-
tential deleterious change, such as a decline in the
density of living coral colonies. To achieve this, we
reviewed published literature to determine what
level of damage a branching coral colony could
sustain before it resulted in total colony mortality.
The threshold would be selected to represent a
level of coral colony damage below that which
would result in total colony mortality.

Detecting change

The method used to detect change in this study
was estimation of effect sizes ±95% CIs, which
were compared with the management threshold.
After each survey, means and standard errors
were estimated for the response variable at the
snorkel and two control locations. To compare
the monitoring data with the threshold during
each sampling period, we first calculated effect
Environ Monit Assess

sizes (e.g. $\bar{x}_1 - \bar{x}_2$) for each of the following three treatment combinations: between the snorkel location and control 1; between the snorkel location and control 2 and between the snorkel location and the control locations combined (data pooled for both controls). To recall, the effect size is defined here as the raw score difference between any two treatments (e.g. snorkel location versus control locations) and the CI indicates the precision of the estimate. We then calculated the 95% CI for each effect size and compared these with the threshold level using error graphs. A 95% CI for an effect size assuming normally distributed errors and two large samples can be defined as $(\bar{x}_1 - \bar{x}_2) \pm (1.96 \text{SE}.\text{diff})$. The standard error (SE) of the difference between two treatments is defined as $\text{SE}.\text{diff} = \sqrt{\text{S.E.}_1^2 + \text{S.E.}_2^2}$. Note that a more cumbersome equation is used for small samples (Fowler et al. 1998). Lastly, although we used effects sizes based on means, there may be circumstances when medians are appropriate, such as when the assumption of normality cannot apply.

To interpret the monitoring data in relation to the threshold, we developed decision rules based on Fox (2001) and Di Stefano (2004). Figure 2 illustrates these rules conceptually. In Case 1, the effect size $\pm 95\%$ CI clearly exceeds the threshold and thus the monitoring findings along with other evidence of an impact would be reported to decision makers. In Case 2, the effect size $\pm 95\%$ CI is below the threshold and no further action would be taken at that time. In Case 3, the effect size $\pm 95\%$ CI overlaps the threshold. In this case, additional sampling with greater replication would be undertaken to obtain a more precise estimate of the effect size. If the new effect size $\pm 95\%$ CI remained overlapping the threshold, the findings would be reported to the decision makers.

**Results and discussion**

Defining causal criteria

Prior to the commencement of monitoring, we defined causal criteria to improve our capacity to infer reliably if snorkelers were the reason for coral damage exceeding the threshold at the snorkel location. For this study, only three criteria were immediately useful for this purpose. Table 2 provides a summary of evidence relating to each of the three causal criteria that if observed during the study would support the conclusion that snorkeling was contributing to an unacceptable level of coral damage at the snorkel location. Evidence supportive of an impact by snorkelers would include a greater abundance of broken corals at the impact location compared with the control locations (Plathong et al. 2000). This relates to the causal criteria strength of association. If all control locations had less damage than the snorkel location (consistency of association), this would strengthen the case that snorkeling, rather than a natural agent of disturbance, was the cause of impact. Lastly, a negative relationship between the amounts of coral damage with distance from the entry point of the snorkel location (Allison 1996) would also provide evidence that snorkeling was the cause of the damage (biological gradient).

After our literature review it was apparent that there would be limitations in our use of levels-of-evidence in this study. Foremost was that not all causal criteria were helpful at providing evidence for or against the conclusion that snorkeling was the cause of impact. Although Plathong et al. (2000) demonstrated that impacts can accrue rapidly after the commencement of snorkeling, the temporality criterion was unhelpful because we commenced our study after snorkelers started using the reef at Shams Alam. Specificity of association was not useful because damage caused

![Fig. 2 Decision rules for interpreting effect sizes ($\pm 95\%$ CI) in relationship to hypothetical management threshold (see text for details)](image)
<table>
<thead>
<tr>
<th>Causal criterion</th>
<th>Literature review findings</th>
<th>Evidence supportive of snorkeling impact</th>
<th>Evidence unsupportive of snorkeling impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strength of association</td>
<td>Allison (1996) reported 12 counts of broken coral at an intensively used snorkel location compared with one to two counts from locations receiving less snorkelers Plathong et al. (2000) reported that the mean percent of recent broken Acropora colonies ranged from ≈30 at snorkel trails to less than five at controls. Sixty percent of colonies at the snorkel trail showed evidence of old injury compared with 10% at control sites</td>
<td>An appreciably larger amount of broken coral at the snorkel location compared with control locations</td>
<td>An appreciably lower or equal amount of broken coral at the snorkel location compared with control locations</td>
</tr>
<tr>
<td>Consistency of association</td>
<td>Two studies reviewed demonstrated high levels of broken colonies and broken branches at snorkel locations relatively to areas experiencing few or no snorkelers (Allison 1996; Plathong et al. 2000)</td>
<td>All controls had less coral damage than the snorkel location</td>
<td>Only one control had less damage than the snorkel location</td>
</tr>
<tr>
<td>Biological gradient</td>
<td>Allison (1996) reported a positive relationship between the amount of broken coral and the amount of snorkeling activity</td>
<td>The abundance of broken coral is positively correlated with intensity of snorkeling activity</td>
<td>The abundance of broken coral is not positively correlated with intensity of snorkeling activity</td>
</tr>
</tbody>
</table>
by snorkelers cannot readily be distinguished from that caused by other agents of mechanical disturbance (Riegl and Velimirov 1991). The criterion experimental evidence was not used, but had potential to assist us with inference if our study had continued. Downes et al. (2002) suggested that experimental evidence of an impact could be derived if the response variable responded soon after the disturbance ceased. That is, if we temporarily stopped snorkeling at the snorkel location and the amount of damage declined with time, this would help support the conclusion that snorkeling was the cause of impact. Lastly, the criteria biological plausibility, coherence and analogy were not used. These criteria are more useful in situations when the mechanism of impact is not well understood. This is not the case with snorkeling because this activity is a well-documented source of mechanical disturbance to corals with a branching growth form (Allison 1996; Meyer and Holland 2008; Plathong et al. 2000).

Management threshold

Before the commencement of monitoring we established a management threshold for coral damage. If this threshold was exceeded during monitoring it would lead to the review of causal criteria and reporting of the results to the management board. To recall, the management threshold was defined as pre-defined effect size of coral damage between the snorkel and control locations that would constitute a change worthy of management attention. Unfortunately, defining effect sizes of important ecological change is not straightforward (Di Stefano 2004; Keough and Mapstone 1995). In our study, we wanted to develop a threshold effect size of coral damage that, if observed during a survey, would provide early warning of potential major ecological damage, namely a reduction in the density of living coral colonies. Consequently, we undertook a review of published studies that assessed the relationship between the level of coral damage and total colony mortality. The review revealed little published information on the level of damage a coral colony with a branching growth form could sustain before it suffered total mortality. A summary of the review is given here. Loya (1976) broke ‘some’ branches of 70 Stylophora pistillata colonies, of which 19 colonies, representing the smallest, suffered total mortality. Liddle and Kay (1987) damaged coral colonies of three branching species (Acropora palifera, Acropora millepora and Pocillopora damicornis) to assess their survival to mechanical damage. Horizontal dimensions of colonies ranged from 15 to 30 cm. Two months after all branches of 14 colonies of each species were removed, all remained alive. Rodgers et al. (2003) experimentally damaged 40 colonies of four species, three of which were characterised by a branching growth form. Eleven months after the colonies were damaged three had died, but two had died also among the control group.

Although our literature review suggested that colonies of some species and size classes can sustain considerable damage without being killed (Liddle and Kay 1987; Loya 1976; Rodgers et al. 2003), we were unable to establish a definitive level of colony damage that, if exceeded, would lead to total colony mortality. For this reason, we took a precautionary approach by adopting a moderate level of coral breakage as our threshold effect size, but acknowledging this threshold could be modified later based on the results of the monitoring program or newly published data. We considered that a 0.15 raw unit difference between the mean proportion of broken branches per colonies at the snorkel location relative to the control locations was a difference worthy of further investigation. To illustrate, this would occur if the mean proportion of broken branches per colony was 0.20 (or 20% of branches per colony) at the snorkel location and 0.05 (5%) averaged for the control locations (i.e. $0.2 - 0.05 = 0.15$). Consequently, a raw unit difference of 0.15 between the impact and control locations would act as the threshold, which if breached during a survey would trigger the review of the causal criteria and the reporting of these results to the management board. Importantly, triggering of further action would apply only in situations where the proportion of broken branches was greater at the snorkel location compared to the controls. Conceivably, a higher level of damage could be observed at the control locations if a natural agent of physical damage affected the control locations, but not the snorkel location.
Monitoring results

During the study, three surveys were undertaken in the first weeks of March 2007, April 2007 and May 2007. After each survey, effect sizes (±95% CI) for the proportion of broken branches were estimated for three treatment pairs: the snorkel location and control location 1; the snorkel location and control location 2 and the snorkel location and the average of the two controls. Each effect size was than compared to the threshold (Fig. 3). At no time during the study did any effect size (±95% CI) exceed or overlay the threshold (Fig. 3) and thus a review of the causal criteria and reporting to the management board was unnecessary. Effect sizes (±95% CI) relating to the proportion of damaged branches did not exceed 0.12 of branches per colony (mean ± 96% CI), so were below the 0.15 threshold. Another feature of these data was the broad lengths of some CI, suggesting that estimates of these effect sizes were imprecise. Given, however, that no CI overlayed the threshold (Fig. 3), it was unnecessary to resample with increased level of replication.

Hypothetically, if one or more effect size (±95% CI) had exceeded the threshold, the next step would have been to seek additional evidence supporting the conclusion that snorkeling had caused the breach of the threshold. That is, we would have looked at each of the three causal criteria defined before the start of monitoring to help us conclude if snorkeling was the likely cause of an exceedance of the threshold. To illustrate how causal criteria could be used following an exceedance, we provide one worked example using the biological gradient criteria based on the May data (see Fig. 4). To recall, the amount of coral damage caused by snorkeling should typically be negatively correlated with distance from the entry point (Table 2). To assess this assumption, we plotted the mean proportion of broken branches per colony from three sampling sites within the snorkel location, namely the entry point, and 100 and 200 m further away. We used the bar graph to visually assess if there was a trend of decreasing damage with distance from the entry point of the snorkel location. There was no discernible relationship between the proportion of broken branches and distance from the entry point during May (Fig. 4). Instead, the greatest mean proportion of damage branches at the snorkel location was found at 100 m, with the least amount found near the entry point (Fig. 4). Consequently, the criterion biological gradient supported the conclusion that snorkeling was not contributing to an unacceptable level of damage at the snorkel location during May.

Our statistical approach to detect change differed from more complicated methods using

![Fig. 3 March, April and May effect size estimates of the proportion of broken branches in relationship to the management threshold. This example shows three effect sizes (±95% CI) pairs: Control 1 (C1) versus the Snorkel Location (SL); Control 2 (C2) versus SL; and the pool data from both control sites (CP) versus SL. Note that none exceed the threshold and no additional action was required (see text for details).]
First, we did not construct CIs in order to test for statistically significant associations between effect sizes and the threshold. Rather, we used CIs only to indicate the precision of effect size estimates in order to improve the robustness of our conclusion relating to the effects of snorkelers at the study location and to provide an easily communicable level of confidence to decision makers. To reiterate, with this framework, decisions are not based on a statistical test of a null hypothesis. Instead, decisions are based on variety of evidence including the relationship between the effect size ±95% CI and threshold.

Second, we made no attempt to detect an impact based on the assessment of a BACI interaction as promoted by Green (1979). He argued that human impacts to natural environments should be assessed by testing for a statistical interaction based on a BACI design. With this approach, detection of change is normally undertaken in context of statistical testing using ANOVA; however, estimation approaches can also be used to assess interactions, but interpretation is not straightforward (Di Stefano 2004). In our study, the statistical component of the assessment was based purely on a spatial comparison that involved comparing effect sizes with the threshold at each survey period. We did not attempt to interpret interaction effects in this study for two reasons. As mentioned previously, we did not have baseline data, so a BACI sampling design was impossible. Second, defining a management threshold for interactions is complicated (Di Stefano 2004), which would have contributed an additional layer of complexity. Nevertheless, we believe that a spatial comparison of the amount of coral damage between the snorkel and two control locations, combined with the causal criteria, provided a robust assessment to indicate if management action was warranted during each survey period. Such action could include commissioning an experienced marine scientist to provide a more detailed site specific investigation, particularly if evidence of impact remained inconclusive.

Implications for managers and biodiversity conservation

Significance testing and parameter estimation are two monitoring frameworks used in marine environments to detect change in biodiversity and to infer the cause of change. In context of field managers of MPAs, both have one or two major drawbacks: computationally demanding, weak inferential capacity if the sampling design is sub-optimal and no decision-making structure. For these reasons we propose a simple yet scientifically rigorous monitoring framework for field managers of MPAs. Although we demonstrate the utility of this monitoring framework using a case study involving snorkeling, the framework could be adapted to monitor the effects of other activities common to MPAs, such as diving, reef walking and recreational fishing.

The monitoring framework described in this paper offers advantages to managers who lack skills and resources to implement technically complex approaches. This framework focuses on estimating the magnitude of change or difference between impact and control locations. As an effect size is easier to understand than a $P$ value, communicating results to decision makers is less problematic and uncertainty is captured intuitively using CIs. An obvious desirable attribute of this approach is that it is in the interest of the organisation whose activity is being monitored to ensure replication is large, thus ensuring the CIs are kept narrow and management thresholds are not breached unnecessarily due to high uncertainty.
This framework retains a major strength of statistical testing by having a clear decision-making structure. This is achieved by specifying in advance, a change in a response variable that would constitute a management concern and, thus, when action is warranted. Importantly, change is assessed in context of natural variation measured concurrently at control locations. It is not based on an unrealistic assumption that response variables (sensu indicator variables) change unambiguously in reaction to human activity. The framework used in this study has considerable potential for determining the cause of an impact because it relies on two inferential approaches: levels-of-evidence and a sampling design. Adopting both inferential approaches will normally be a necessity in context of MPAs because rarely will all elements of an experimental sampling design be fulfilled, nor will all causal criteria be usable during monitoring. The use of causal criteria has the added advantage of being transparent and easy for non-scientists to understand. Also, defining causal criteria prior to monitoring and in consultation with stakeholders should reduce the risk of conflict, especially if the results prove unfavourable to the organisation responsible for the activity being monitored. Lastly, this monitoring framework aims to detect change well before there is deleterious change to biodiversity. This is done by selecting appropriate response variables and setting thresholds at levels well below that which would constitute major ecological damage.

We end with some cautionary notes. First, this framework was adopted to help us detect a change worthy of management interest and to increase our confidence that snorkeling, not a natural agent of disturbance, was the cause of change. The framework was not used to make generalisations about the effects of snorkeling on corals or to assess specific management approaches to mitigate snorkeling impacts. Such investigations require experiments, which are complicated to undertake in marine environments (Barker and Roberts 2004; Campbell et al. 2001; Rouphael and Inglis 2002) and require a high level of expertise. Second, managers need to periodically review their management thresholds based on the findings of the monitoring results. This would have been essential if our study had continued because we were unable to establish a clear relationship between colony damage and whole colony mortality before monitoring commenced. Third, although we promote this framework as being less complicated than statistical testing, managers still require an understanding of CIs to ensure correct interpretation of data and communication of results. Four, with this framework concluding that an impact is due to human activity rather than a natural event is more subjective compared with statistical testing. However, this risk is reduced by establishing causal criteria before monitoring and constructing CI for parameter estimates. Lastly, this framework does not negate the need for scientifically robust sampling. Biased and confounded parameter estimates will still be uninterpretable irrespective of the framework used.

Acknowledgements The authors wish to acknowledge Jim Stoddart and Michael Marnane for reviewing an early draft. Mahmoud Hanafy is thanked for assisting with the logistics of the study.

References


