THE *BOUCHARD-120* AND CHALK POINT SPILL RESPONSES: OBJECTIVES AND PERFORMANCE METRICS

S.P. TULER, T.P. SEAGER, R. KAY
Social and Environmental Research Institute
278 Main Street, Room 404, Greenfield, MA 01301USA

I. LINKOV
Intertox
83 Winchester Street, Suite 1, Brookline, MA 02446 USA

Abstract

Oil spills present a chronic threat to the environmental security of most major ports. While mitigation of the risk of oil spills should include prevention, major oil spills remain periodic occurrences. Consequently, spill preparedness and response are critical aspects of minimizing the damage caused by spills. Nonetheless, any major spill response engages multiple stakeholder and public groups that may have different objectives. Currently, spill managers must balance conflicts in the midst of a crisis using *ad hoc* or heuristic approaches that may be difficult to justify or communicate. Public expectations are particularly challenging to manage. In some cases, the spill response may be perceived as a failure despite the response agency's best efforts. A systematic approach to stating varied spill objectives and tracking progress may result in better management and communication and improve the credibility of spill managers. This research studies two separate spill incidents to reveal the different types of objectives held by engaged personnel and the ways that they assess the progress of the response. A total of 30 interviews are conducted and interpreted using a grounded theory approach to reveal salient objectives. Where possible, metrics relating to these objectives are elicited and the results for each spill compared. Although the quality of the study metrics is not examined in detail, we find that some spill objectives are more readily stated in terms of performance metrics than others, suggesting that spill managers may benefit from greater guidance on how to gauge progress or set goals in areas such as protection of public health and safety or mitigation of sociopolitical or economic impacts.

I. Linkov, G. Kiker, R. Wenning (eds.) Environmental Security in Harbors and Coastal Areas, 209-228.
© 2006 Springer. Printed in the Netherlands.

1. Introduction

Oil spill response is defined as encompassing all activities involved in containing and cleaning up oil spills in ways that achieve the following overarching goals [4]:

- Maintaining safety of human life.
- Stabilizing the situation to preclude it from worsening.
- Minimizing adverse environmental and socioeconomic impacts by coordinating all containment and removal activities to carry out a timely, effective response.

To achieve these goals, oil spill response efforts must include a variety of participants. They can include: federal, state, and local officials (e.g., US Coast Guard, US Environmental Protection Agency, US Department Of Energy, local fire chiefs, harbormasters, state environmental officials), the responsible party and its contractors, non-local clean-up crews hired by private contractors, environmental and community advocates at the national, state, and local levels, and community residents who have an important stake in the response, (i.e., business- and home-owners, and beach associations). Because so many different and interested parties are affected, there is a potential for conflicting ideas about how a response should be organized and implemented to achieve these goals. Different groups may have different assessments of oil spill response success because they have different objectives and some may be in conflict with others. Response strategies are always dependent on the priorities placed on protecting specific resources in the context of a particular spill. Therefore, decisions made about priorities are to a very large extent political or social as well as technical. Consequently, measuring the success of any response (for example, in accordance with the Governmental Performance Results Act of 1993) is a significant challenge. To capture potentially disparate views and to facilitate management decisions, multiple performance metrics must be employed. However, good metrics for capturing the nuances of the decision process are not always available.

This chapter addresses the question of what objectives and performance metrics are used by key stakeholders to assess two recent oil spill responses:

1. The *Bouchard-120* spill response that began on April 27th, 2003 as the tugboat *Evening Tide* ran its tanker aground and released No. 6 home heating fuel just at the entrance to Buzzards Bay, Massachusetts. (See www.buzzardsbay.org for further details.)
2. The Chalk Point spill response that began on April 7, 2000, when an intrastate pipeline that transports oil from the Potomac Electric Power Company's (PEPCO) Chalk Point electrical generating facility to residents in Prince George's County released No. 2 and No. 6 home heating fuel oil into Swanson Creek and subsequently the Patuxent River. (See www.darrp.noaa.gov/northeast/chalk_point/index.html for further details.)

The full details of each case study are more completely described in Tuler et al. [6].

2. Oil Spill Response

Within several hours of a reported spill, an Incident Command System (ICS) incorporating federal, state, and local authorities is mobilized at the scene. Among the critical government participants in management of any major spill are representatives for the Responsible Party (RP), the Federal On-Scene Coordinator (FOSC), State On-Scene Coordinator (SOSC), and National Oceanic and Atmospheric Administration's (NOAA) Scientific Support Coordinator (SSC). The role of an SSC is to provide the FOSC scientific advice and information (such as weather and spill movement forecasts). In the initial response, the SOSC is responsible for notifying local first responders by telephone about the spill. As a coastal spill, the *Bouchard-120* fell under the jurisdiction of the USCG. By contrast, the Chalk Point spill was classified as an inland spill, over which the USEPA has jurisdiction. Other key players involved in the response and clean-up are various local, first-response officials; non-profit local advocacy groups; and contractors hired by the RP. Over a dozen government agencies may be involved in any major spill response.

In both the case study spills, failure to immediately recognize the magnitude of the spill may have complicated or delayed response efforts. In addition, strong winds and rough weather forced oil past containment booms in both cases. Consequently, oil migrated to areas that responders initially thought could be protected. In the case of the Bouchard spill, oil was "blown back" onto shores that had at first avoided oiling or had been recently cleaned.

Both response teams made extensive efforts to engage and inform the public, although these were not necessarily always perceived as successful. In the Bouchard case, summary briefings by various groups attending the incident command meetings were held several times daily initially and then twice daily once the situation was better under control. Both teams held public meetings held over the course of the response, established call centers and built or maintained websites. In the case of the Chalk Point spill, public meetings and briefings were supplemented by five issues of a newsletter that was circulated to over 27,000 area residents.

3. Research Methods

Understanding how people assess spill response efforts requires understanding of their goals and objectives. To explore these issues in the context of the study spills and responses we conducted a series of interviews and investigated published reports to gather information about:

- Roles and experiences of key responders and other interested and affected parties.
- Their concerns about spill impacts and response.
- Their views about the response.
- Their views about response performance metrics.

The interviews were designed to be semi-structured and open-ended, but they were *not* designed to systematically elicit information about relative priorities among objectives or performance metrics. Instead, at this time we were more interested in learning about the range of objectives and metrics among key participants. Initial research interviewees were identified in articles about the spill. Subsequent interviewees were selected in a snowball sample on the basis of suggestions by others. Interviewees were selected to represent different points of view, their experiences with the spill, and willingness to be interviewed.

To identify the information relevant to study objectives, we used a grounded theory approach, in which important concepts emerge *inductively* during the data analysis rather than in advance of the investigation [1, 2]. In grounded theory, data are categorized with respect to relevant similar characteristics in a process called "coding." At first, a relatively large number of categories are developed. Then, through iteration these categories are grouped into more abstract categories of conceptual relevance to the analysis; data and categories are grouped according to their relationships with each other. For example, all statements related to "reducing bird injuries" or "protecting nesting habitat" can be grouped into a category named "response should protect bird populations." This is referred to as *axial coding* in a grounded theory framework. In this way common themes among the coded objectives are identified. We then extracted all performance metrics expressed by the interviewees that were related to each of the objectives. In addition, we compiled all interviewee comments regarding appropriate uses of performance metrics for assessing oil spill response efforts and characteristics of "good" performance metrics, although we did *not* attempt to assess the quality of any of the suggested metrics either according to the interviewees own criteria for "good" metrics or other norms. (See Seager et al. [4] for a discussion of such criteria).

4. Results

We found that many different objectives for the response to the oil spill were important to our interviewees. We grouped them into endpoint, process, and resource-based metrics (following Seager et al. [4]) related to the following broad categories identified in Tables 1 through 9:

Table 1. Metrics related to protecting public health and safety.

Endpoint	Number people killed or injured.
	Number of mishaps during hours worked .
	Presence of contaminants (e.g. PAHs) in water samples?
	Concentrations of oil in fish tissues.
	Number life threatening situations to human health.
	Toxins in smoke plume if do in situ burning.
Process	(None identified).

Resource	Number of IRAC team members OSHA/HAZMAT trained.

Table 2. Metrics related to protection of cultural resources (Chalk Point).

Endpoint	Number of critical sites protected.
	Soil concentrations, smell or residual presence of oil on artifacts.
Process	Did trench digging affect sites?
	Experts contacted early for input about sites potentially at risk?
	Command responsive to requests for protection of sites?
	Were less destructive response actions chosen (e.g., sorbents and booms rather than burning)?
Resource	Number of GIS, hard maps, laptops and accuracy of location information.
	Amount of boom deployed

Table 3. Metrics related to mitigating social nuisances (Bouchard).

Endpoint	Presence of stained rocks, oil on beach and complaints?
Process	(None identified).
Resource	(None identified).

Table 4. Metrics related to protection of environment.

Endpoint	Number of oiled birds, eggs or other wildlife.
	Number of miles of shoreline impacted or cleaned.
	Amount of oil or globules on shore.
	No re-oiling or residual oil causing chronic toxicity to something.
	Number of fish, birds or other wildlife killed or injured (per unit search area).
	Number of “appropriate” (not exotics) animals rehabilitated and released.
	How far sheen at surface extended out [miles]?
	How long oil stayed?
	Presence of odors of oil.
	How much grass was destroyed or acres of marsh were impacted.
	Dead and stressed organisms found (rather than estimated).
	Time to achieve background levels/concentrations of contaminant or clean-up standards and recover from clean-up related damage.
	Areas protected (e.g., by redirecting or containing oil).
	Observe water blowing over booms.
	Degree of change to beaches and sandbars from clean-up actions.
	Types of animals and vegetation present after spill cleanup.
Process	Did getting required permits delay response action?
	Time for wildlife rehabilitation efforts to set up operations.
	Rate of bird handling at rehabilitation center.
	Gallons of oil and pounds of contaminated debris recovered and disposed of.
	Accuracy of cataloguing and enumeration of findings.
	Monitoring stations established.
	Time to deploy booming and double-booming in sensitive areas.
	Immediacy of rehabilitator organization’s response to call for assistance.
	Change of helicopter flight patterns in response to requests from biologists to not disturb nesting birds.
	Oil direct to sacrificial (rather than sensitive) areas.
	Oil being captured in open water before it hits the beach.
	Area covered in search and recovery.

Resource	Amount of oil containment boom deployed.
	Number of volunteers.
	Number of floating resources to pick up oil in open water oil.
	Number of bodies to manage different aspects of response, inc. SCATs.
	Is there a 'bird searcher' on each team?
	Number sandbags deployed.
	Number people on cleanup crews to deal with oiled beaches

Table 5. Metrics related to economic impacts.

Endpoint	Lost rental income, tourism dollars, property values, wages to fisher/watermen.
	Duration of beaches closures.
	Recreational opportunities that were lost that are now back to what they were?
	Acres and duration of shellfish areas closed or acres of closures reopened.
	Acres of shellfish beds lost and number of lost fishing days.
	Costs of laboratory work, other research studies and money spent on response.
	Increase in crime rate [in southern MD due to influx of people from cleanup crews].
	Number of dead fish, ducks & geese.
Process	Kept track of all costs.
	Local municipalities reimbursed by responsible party?
Resource	(None identified).

Table 6. Metrics related to coordinated and effective response framework.

Endpoint	Clear chain of command established and Incident Command System used?
	Are pre-identified areas potentially being affected?
	Did we have information to keep Governor's office and other state Senators and Reps abreast of what was going on?
	Understanding of whether oil is still stored offshore re-contaminating cleaned up beaches?
	Is oil coming on shore several days later?
	Accurate accounting of volume oil spilled and on shore?
	Number areas cleaned as of today? Number of miles of shoreline impacted right now?
	Bad feelings among locals responders toward the Unified Command staff?

	RP is responsive?
	Number of hours to set-up stable incident command center. Location easily accessible?
	Key people became involved early? Local and federal responders notified quickly?
	Clear understanding of rights the state trustees have as a state agency?
	Is Incident Commander able to reach key people, does their phone number work, did they respond? Clear communication protocols and reliable technology working?
	How quickly decisions made? Are decisions correct (in hindsight)?
	Is there conflict or chaos in command center? Chauvinistic behavior? Cooperation?
	Did Unified Command resist information that did not conform to their expectations?
	Number and frequency of meetings, daily reports for morning meetings?
	Informed of meetings in advance (i.e., lead time)?
	Presence of watchdog to see what's going on?
	Modelers able to get 24 hours ahead of spill with accurate projections?
	Experts consulted for input on response strategies?
	Are there clear protocols and schedules? Plans communicated day in advance?
Process	Time it takes to implement tasks, such as boom deployment?
	Are players familiar with each other? Frequency of resource and personnel changes. Time taken to re-staffing response people after contractor fired.
	Pick-up and shipping schedule for waste generated by clean-up organized?
	Resources placed in the proper locations? Response organized by segments?
	System established to track progress? Are crews visiting hard hit areas every day and recording information in a unified way? Is all pertinent data gathered and recorded?
	Ability to revise objectives and activities based on monitoring effort?
	Accurate information obtained from the wildlife surveys and SCAT teams?
	Systematic, 'non-political' approach used to deploy clean-up crews.
	Attend to short, medium, and long term needs simultaneously?
	Follow 'best response' protocol and integrated command system?
	Coordination of volunteers performed appropriately and quickly?
	Clarity to all parties about stages of response effort?
	Equipment and personnel demobilized when no longer needed?
	Clear standards for sign-off established.

Resource	Number of supervisors, Spanish speaking supervisors assigned per section.
	Number of radios; availability of GIS and computers; phones available and working.
	Number of teams of trained observers walk coastline and make observations of extent and coverage area of oil
	Number crews trained, hours worked.
	Type of oil.

Table 7. Metrics related to effective and timely response.

Endpoint	Number of days company (PEPCO) shut down.
	Presence/absence of sheen, oil in water, tarballs, oil on shoreline, oil in sediments.
	Areas impacted to be cleaned up to the ecological state the environment was before the spill happened. Number of areas to be signed-off compared to total.
	Number of days until endpoints achieved?
	Is oil being contained? Patchiness of oil? Solid sheet of heavy oil.
	What it looks like a year later.
	Number of public meetings organized? Members of the public voice support?
	Good working relationships with all parties involved?
	Are response actions having the desired effect?

Process	How often beach is searched—both oiled and non-oiled?
	Digging holes to look for oil on shoreline?
	Quarterly checks to see if oil is present or not on beaches
	Breach of water over boom?
	Protection equipment put in the right place at the right time?
	Quality of contractor work. Number of times a orders were given but still not done.
	Number of public meetings, newsletters published by RP.
	Time it takes to get response in order. Lead resources mobilized?
	Number/frequency of fly-overs for real time aerial photography.
	Clear chain of command established? Communication to appropriate people?
	Accurate reporting and counting of crews in field. How quickly SCATs out in field?
	Are clean-up crews assigned effectively to do a good job?
	Is all pertinent data gathered and recorded? Data sheets available and sufficient?
	Local officials set up task force for spill response? Muster all forces in town?
	Basic training for volunteers on bird collection and rehabilitation conducted?
	Cleanup of impacted areas organized to be manageable and able to monitor?
	Access established for recovery and clean-up crews through private property?
	Do efforts correspond to tides?
	Recovery or rehabilitation of wildlife conducted? How soon have experts been called and set up triage and rehab centers?
Resource	Number of teams/people/supervisors in the field? Number of volunteers?
	Number of people working at one time. Hours worked.
	Number of monitors in field to give direction and warning to clean up crews
	Resources adequate for planned tasks?
	Cost of response.
	Types of skills represented on team? A ‘bird searcher’ on each SCAT?
	Pounds of sorbent material. Number of packets of baby oil for oil removal.
	Amount of oil containment boom.

Table 8. Metrics related to achieving legal regulatory requirements.

Endpoint	Achieve termination endpoints? Shoreline back to conditions prior to spill?
	No oil should come off to the touch.
	Evaluate response with respect to endpoints achieved.
	For sandy beach no visible oil, no odor of oil. For marshes no sheen.
	For groin (jetties between properties), riprap no sheen or no oil available when touched.
Process	Number of days until endpoints achieved.
	All procedures followed (e.g., NIMS).
Resource	(None identified).

Table 9. Metrics related to addressing public concerns, needs, and support.

Endpoint	Public reimbursements for private property losses.
	Number calls from public. Members of the public voice support?
	Public comments from critics and local residents (re: response effort, not outcomes)?
	Level of staining Complaints about stained rocks? Residual oil on shore (tar balls)?
	Was spirit of state regulations for public involvement met?
	Level of public trust?

Process	Relationships and trust with local officials developed?
	Immediacy of public meetings (number days after spill). Number of public meetings.
	People given examples of what was impacted and what kinds of cleanup was going on?
	Number of fliers and informational packets delivered door to door, visuals for media.
	Public provided the kinds of information it wants? Incorrect information disclosed?
	A timeframe for ending the cleanup established?
	Ongoing monitoring and addressing of issues post-spill?
	People have a place or someone to go to with concerns?
	Public receives assurances that beaches will be cleaned up to the level of their expectation?
	Is a forum provided to public so they can hear what's going on and give their feedback?
	Feelings: level of conflict/anger or happiness.
	Establish and keep up to date website for public information.
	Quality of questions from media
	Able to 'stay on message' during public meetings, press conferences, etc.?
	Unified Command accessible for public questions and comments?
	Amount of oil removed manually from shoreline.
Resource	Number hours agencies spent on public outreach (meetings).
	Number of pamphlets distributed to inform public of hazards.
	Number of stakeholders involved in setting clean-up standards.
	Number of dispatch teams arranged to reach-out to various stakeholders.
	Frequency of information postings on website.

Interviewees in both cases wanted to protect sensitive habitats and populations of threatened and endangered species (although the methods were not designed to tell us about the relative importance or frequency within the respective case study populations). They were concerned with mitigating impacts from clean-up actions, including economic impacts related to lost recreation, tourism, fisheries, and to towns from their efforts during the response. Furthermore, they were concerned about the timely gathering and use of relevant, accurate, and credible information for decision making. Strong and flexible leadership that can learn from past experiences was important to them so that the response could be well-planned and implemented. And, they shared objectives related to addressing public concerns (e.g., providing accurate information to the public).

At the same time, there are a few interesting differences, which are reflective of the particular contexts of each spill. First, mitigating impacts to cultural resources was identified in the Chalk Point case and not the Bouchard case. This reflects the presence of significant artifacts in the region affected by the Chalk Point spill. It may also be a reflection of whom we interviewed—or more accurately, did not interview in the each

case. However, it is noteworthy that no one in the Bouchard case made mention of historical or other culturally significant artifacts.

The mitigation of social nuisance impacts was not a category of objectives that emerged from our analysis of the Chalk Point interviews. However, some factors that can be related to social nuisance impacts were raised in the context of other objectives. For example, addressing the potential for an "increase in crime rate [in southern MD due to influx of people from cleanup crews]" might be considered by some to be a social nuisance, but the interviewee was clearly talking about this in the context of economic impacts.

While people in both cases expressed objectives (and performance metrics) related to establishing a coordinated and effective response framework, there were some differences in emphasis. These differences suggest the importance of the particulars of experience that inform people's views. Critical comments emerged about the integration of non-federal officials and responders into the response effort. In the Bouchard case this concern was raised about local officials and local first responders, but not in regard to state-level responders. In the Chalk Point case we found the opposite.

In the Bouchard interviews, the roles and participation of local officials and local residents were a concern. In fact, the way in which local responders along Buzzards Bay were contacted and integrated into the response was a source of contention and criticism among those we interviewed. In addition, there was a concern with how local volunteers were brought into the response effort in the Bouchard case. On the other hand, among the Chalk Point interviewees neither of these two concerns was strongly articulated; volunteers were not even mentioned in these interviews. Instead, there was a concern with whether expert input (e.g., about cultural resources) was obtained and how well federal and state agencies' were coordinated (e.g., clarity about jurisdictions). In the Chalk Point case a federal official spoke of the importance of involving local stakeholders in decisions about clean-up goals, but the metrics suggested were about one-way communication and outreach to the public (e.g., number of pamphlets). However, there are some indications that involvement of local officials would have been useful in Chalk Point in terms of understanding local nuances, such as the effect of currents that would impact oil removal efforts. There was also a concern expressed from *state* officials about their involvement in the response effort and much said about the quality of leadership and the coordination and communication among responding federal and state agencies at Chalk Point. Another difference among objectives relates to the kind of people that expressed a concern with gaining public support for the response. Only state and federal officials spoke of this as an objective. This is a very instrumental perspective that one often finds from officials involved with hazard management.

Interviewees in both cases also expressed the importance of minimizing the costs to the responsible party, although this objective might conflict with others. All objectives identified as important during the interviews are related to the protection and promotion of what people value, such as protection of critical habitats and promotion of decisions

based on the best information available. *Why* they value certain things can differ; they may value certain outcomes for intrinsic reasons (e.g., value of species for their own sake) or instrumental reasons (e.g., they allow other things to be accomplished) and the relative weights given to each may also differ. Overall, a large number of performance metrics were suggested by the interviewees. However, a number of interviewees also pointed out the shortcomings of some of them. For example, one interviewee noted that the number of volunteers taking part in the response may not be a good measure because there are constraints on using volunteers (e.g., OSHA regulations) and thought that a better metric might be the number of people calling to volunteer.

However, performance metrics are best defined with respect to specific objectives, which is why we have organized the elicited performance metrics with respect to each objective in Tables 1 through 9. By coding performance metrics in the context of particular objectives it is apparent that many metrics were discussed for some objectives (e.g., mitigating ecological impacts) and few in relation to other objectives (e.g., protection of public and worker health and safety). This should not be taken to mean that those objectives associated with more performance metrics are more important than those with fewer performance metrics; this conclusion is not justified based on the evidence available. Instead, some objectives may be easier to gauge than others or there may be consensus about which metrics to use. For example, protecting health and safety can be assessed by counting OSHA-reportable injuries and fatalities and work-hours without accidents, whereas degradation of cultural resources may be difficult to capture quantitatively. Another reason may be an artifact of how we grouped objectives into larger categories. Furthermore, one might argue that many of the metrics suggested for assessing the prevention and mitigation of ecological impacts and economic impacts could also be measures for whether or not public concerns were addressed. However, they were not always expressed as such. Instead, addressing public concerns was often measured in terms of what effort was made for public outreach and involvement: How many meetings? How many leaflets distributed? Our inability to systematically identify all performance metrics with respect to each objective is a shortcoming of using interviews. This shortcoming is to be addressed by further research; we are developing methods that will ask respondents to rank-order objectives and to rank-order performance metrics for each objective.

5. Discussion

When we consider the kinds of metrics elicited in the two case studies, one observation is that the performance metrics suggested for each case differed for some objectives. For example, making the best possible decisions with limited information, uncertainties, and time pressures was identified as an important objective by many people we interviewed for the Chalk Point case. While this category of objectives was also raised in the Bouchard interviews, it seems to have generated much more attention from the Chalk Point interviewees; this may be a reflection of their perceptions of inadequacies during the Chalk Point spill response. While they found it difficult to articulate specific measurable performance metrics for them, it is nevertheless

important as they view clear chain of command, strong leadership, clarity about responsibilities, and organizational jurisdictions to be closely associated with good outcomes.

Some interviewees argued that it is not always useful to define performance in terms of endpoints because they are difficult to measure in the short-term. For example, one person stated that the point was whether a comprehensive *effort* was made to recover all the wildlife. Many of the proposed metrics have to do with process, as one way of measuring "good" decision making is by the way that decisions are made. In these cases much of the attention is on measuring the quality of the effort. Endpoints are related to effectiveness, but these cannot be measured directly. For example, the "quality of decisions" is difficult to assess on the basis of decision outcomes because there can be many intervening factors that affect outcome. However, *how* a decision was made or information gathered and validated can be more appropriate. It is also apparent with objectives that are associated with the *conduct* of response effort: establishing a well-coordinated response, meeting legal requirements, and implementing an effective and timely response. On the other hand, performance metrics related to mitigating economic impacts were almost entirely related to endpoints, rather than processes or resources related. This may be a reflection of the belief that economic costs are easier to measure than impacts to ecological impacts. In fact, estimating economic impacts can be very difficult.

In some cases the endpoint metrics elicited assess indirect or interim indicators of the state of an ecological system, habitat, or population. For example, impacts to ruddy ducks were based on extrapolations of empirical field data. Similarly, some performance metrics are direct measure of economic impacts, such as lost tourism dollars and lost fisherman income, while others are indirect measures: length of shellfish bed closures, acres of shellfish bed closures, number dead ducks and geese, duration of beach closures and presence of polyaromatic hydrocarbons (PAHs) in water samples. Furthermore, endpoints suggested to evaluate the effectiveness of response were often indicators of long-term impacts. A difficulty reflected by many of the suggested metrics is that measures of effectiveness or ultimate impacts can rarely be made at the time of response because the systems are dynamic. This is a challenge in assessing progress toward goals that take a long time to realize [3]. In the case of oil spills, a desire to assess the effectiveness of the response is really about wanting to know about long-term conditions and how they differ from pre-spill conditions. To gauge long-term impacts, indirect indicators observable in the short-term were suggested.

Many of the suggested performance metrics are based on what can be counted or observed (e.g., did it occur or not occur). Metrics that are easy to measure may be more appealing from a bureaucratic perspective. How many gallons of oil and contaminated debris were removed? How many leaflets were distributed to local residents? How many birds were found dead or oiled? Were endpoints defined in plans achieved? Were state and federal standards for contaminants met? Of course, it might not be so easy to count such things accurately, claimed some of the interviewees. In fact, the focus was not always on *quantitative* metrics that some may assume can be accurately

counted. For example, a state official in Massachusetts asked "was the *spirit* of state regulations for public involvement met?" which can only be answered subjectively. Problems raised by subjectivity and definitional clarity are also apparent in another example. One metric suggested was that "appropriate" response actions be used. However, the definition of what is appropriate is influenced by the type of oil spilled. For example, if the oil floats, there is less need to worry about it affecting nearby sunken ships. However, if the oil sinks, those sites may be of more concern. Appropriateness can also depend on the kind of artifacts or sites that are at risk. Chemical dispersants may not pose a risk to future research on the site when its age is already known, but they may make it difficult or impossible to use carbon dating for some sites (e.g., prehistoric sites). Similarly, effective oil removal techniques may impact the archeological integrity of a site, so what may seem appropriate with respect to one objective may be inappropriate with respect to another.

Several interviewees observed that just because something can be measured does not mean it is relevant to understanding the success of a spill response or important to many stakeholders. For example, several interviewees in both cases suggested that the amount of boom deployed could be used as a metric to assess how well a shoreline was protected from oil contamination. While it is easy to measure, it may not be a good indicator of whether the shoreline is actually impacted. As we were told by many interviewees, deployed boom was not always effective when currents or winds were strong. Another stated that "we tried to prevent oiling by putting out booming, but it's not a very accurate predictor of whether the outcome [of shoreline protection] will be achieved. Once any oil gets on a beach you still need to clean it up." Furthermore, deploying booms may have been a good decision at the time, but weather conditions can shift and cause them to fail. That is, the metric does not have a causal relationship between the state of the system and the variables that are under a decision maker's control. Similarly, an interviewee suggested that the amount of money spent—for numbers of crew, manhours, and amount of boom deployed—may be very appealing politically as a measure of performance, but questioned whether that is "doing the job to the most effective way" and suggested "there may be more cost-effective ways to look at the whole picture. Money spent is not a good measure." Nonetheless, assessment of short-term success and long-term success can be very different in natural systems [3]. Therefore, interim gauges are essential to provide more immediate feedback to decision makers. Provided the causal linkages between the interim measures (such as effort or process efficiency measures) and endpoint measures (such as recovery of wildlife populations) are well understood, the interim indicators can be an improvement on endpoint measures.

Accuracy, consistency, and reproducibility are frequently a problem—particularly with qualitative metrics. We learned that the amount of oil removed from beaches might be a relevant metric for assessing response, but some argued that the measurement is not meaningful because it is very difficult to quantify the actual amount of oil on the beach because it is mixed with sand and rocks and in sorbent material: "Especially in recovered number you are getting a mixture of oil and water…skimmers never perform as the manufacturers claim." Similarly, different people may have different ideas about what is "impacted." One way around this is training personnel so that "when they come

back they say ‘we’ve got 200 yards of shoreline heavily impacted’ what one calls ‘heavily impacted’ is the same thing… . We try to get them to know what the shoreline types are so they can describe it in the proper terminology.” However, qualitative or semi-quantitative descriptions may present a moving target and personnel continue to learn on the job, as one subject explained: “Getting the shoreline reasonably back to where it should be…it was a subjective judgment call with each section of the shoreline based on the criteria and adhering to the criteria… We refined the process as we went along over time because you have more of a sense of perspective after you’ve seen a bunch of different segments of the shoreline.”

In several cases, objectives were posed as a “yes” or “no” question, such as “Are members of the public happy?” or “Was there trust from the public?” or “Are there good working relationships with all parties involved?” Some of the might be rephrased to suggest a more sliding scale, such as “What is the level of trust from the public?” However, spill managers would need new approaches to assessing such qualities in the midst of the response. In other cases, it maybe difficult to trace measures directly to the impact of the spill, such as number of lost fishing days suggested by a Chalk Point interviewee: are fishing days “lost” due to impacts of the oil spill or because of foul weather or other reasons? Similarly, natural mortality may be “tallied in as a result of the spill.” Typically, there is a paucity of accurate baseline data on which to base comparisons or assessments. For example, “there wasn’t an established criterion of how many ppm of hydrocarbons in shellfish is dangerous…There are spills here from these boats all summer long…each municipality needs to know what the baseline is and what the safe baseline is…what’s allowable and what’s standard?”

According to some federal and state responders, the cause-and-effect relationships between decisions and performance metrics can be very difficult to communicate to the public. This issue was raised repeatedly in regard to the impacts of clean-up efforts. Many interviewees from both case studies provided examples of when, according to them, it would have been better to cease clean-up actions because more severe impacts would result than if none or no further actions were taken. However, articulating these trade-offs to the public was difficult. There may be a disconnect between scientific metrics and measures of public perception: “We took sample after sample after sample [testing for presence of oil contaminants in fish] and everything was fine...the only problem was convincing people that everything was fine.” Similarly, the public may not understand the details of response actions. In one case, “people were reporting that clean-up crews [were] being negligent, because of throwing rocks back into the water. But that was what they were instructed to do after wiping them off—and here is the rationale for doing that. After we explained this to them they were accepting of that process.” Ultimately, some interviewees’ feelings could be summed up succinctly by the comment made by one that “a lot of what we get to judge spills by is how the public feels about it after the response… public perception is our reality.”

The causes of good or poor performance may be multiple and difficult to disentangle. In other words, while measurement of a performance metric may suggest a poor response the response might have been “good” in the context in which the decision or action was made. For example, one person suggested that a relevant metric for

assessing success is the amount of oil recovered in open water before it hits the shore. However, in the Bouchard spill response, very little was picked up in the open water because of 1) weather conditions and 2) poor coordination in the early stages. While the second reason might suggest a poorly organized response, the first reason is beyond the control of the responders; it would be unfair to judge the response as poorly implemented because of harsh weather conditions that prevented a higher rate of offshore recovery. Similarly, in the Chalk Point spill there was a concern about the amount of oil recovered or removed. Several interviewees thought that *in situ* burning might have been a better response option, but the time it would take to get all the permits and necessary information in place precluded its use. The problem was not the lack of coordination, but rather other external factors limiting this option.

We found that some people assessed the response to the Bouchard and Chalk Point spills on the basis of how well preparations such as planning workshops, clean up standards, spill training, pre-spill resource procurement, staging areas permissions, contingency and emergency plans, and call lists were made prior to the response. Their concern for preparedness was related to for a) gaining public support for the response and b) ensuring effective response effort and coordination.

Furthermore, some of the interviewees assessed the quality of the response based on how well it supported the needs of damage assessment and restoration related activities. For example, several interviewees in both case studies stated that an important performance metric for the response effort was how well it developed the data necessary for damage assessment

These findings suggest that people do not necessarily consider “response” as distinct from preparation, damage assessment, and restoration. In fact, these activities can overlap in time. Activities considered as part of the response can extend for many years after the spill event and the initial clean-up. Such is the case in the Bouchard case, where several years after the spill occurred some activities related to “response” are still being conducted.

6. Conclusions

The case studies of the *Bouchard-120* and Chalk Point spill responses suggest that spill managers and interested and affected parties can have multiple objectives. They may also use multiple measures for assessing achievement of or progress on those objectives. By and large, they were cognizant of the usefulness of assessing response efforts both during and after a spill response. Although these case studies also illustrate that objectives may not be entirely shared among interested and affected parties, in general making objectives explicit and tracking performance metrics related to those objectives may improve spill management and public communication. While we did not interview many local residents (beyond those working for a local advocacy group or local officials) for these case studies (because we intentionally selected people who were engaged with the response), a number of people spoke of their experiences with members of the public. In particular, local residents were described as wanting to know

very practical information that related to impacts and response: what were clean-up schedules? what were public health effects of consuming fish contaminated (or that might be contaminated) by oil? Although some of our interviewees characterized members of the public as wanting things that were impossible or that were not reasonable, e.g., remove *all* oil, a few members of the public can be characterized as quite sophisticated in their views.

We also found that while interviewees think performance metrics are important and useful, there may be differences about *when* they should be developed and used. For example, response performance metrics were developed by the Unified Command *during* the Bouchard spill when they "sat down as a group—three weeks into the spill—because first three weeks we are just responding… . We ask the bird wildlife and rehabilitation people, what are your success factors? They said we have a goal of 20% rehabilitated. We said OK, that is the objective of the bird people. We asked the safety people, what is your objective? They said, occupational safety and health said this many hours, no mishaps. So we started tracking that number and put them on a poster so people could track them… ." The hypothesis motivating our research project is that better oil spill response can be achieved if performance metrics are defined *prior* to a spill in a systematic and collaborative process. Nonetheless, establishing metrics as part of the response process may be an improvement on management without any metrics at all.

Spill response metrics can drive which strategies are chosen and help officials gain the legitimacy and trust they hope for—if they can show that their efforts measure well on the chosen metrics. Spill response metrics can also drive learning. Most generally, a measurable understanding of the different objectives of the parties directly engaged in oil spill response and the communities impacted by the spill is likely to improve understanding, communication, response and ultimately reduce the risk of adverse impacts due to oil spills.

7. Acknowlegements

This study was funded by the Coastal Response Research Center, a partnership between the University of New Hampshire and the National Oceanic and Atmospheric Administration under NOAA Grant Number: NA04NOS4190063. Project Number: 05-983.

8. References

1. Glaser B, Strauss A. 1967. The discovery of grounded theory: Strategies for qualitative research. Chicago: Aldine.

2. Glaser B. 1992. Basics of Grounded Theory. Mill Valley, CA: Sociology Press.

3. National Research Council 2000a. International Conflict Resolution After the Cold War. Washington, DC: National Academy Press.

4. NOAA. 2001. Characteristics of Response Strategies: A Guide for Spill Response Planning in Marine Environments. A joint publication of the American Petroleum Institute, National Oceanic and Atmospheric Administration, US Coast Guard, and US Environmental Protection Agency: Seattle WA. http://response.restoration.noaa.gov/book_shelf/910_response.pdf

5. Seager TP, Satterstrom FK, Tuler SP, Kay R, Linkov I. 2006. Typological Review of Environmental Performance Metrics (with Illustrative Examples for Oil Spill Response). Integrated Environmental Assessment & Management. In press.

6. Tuler SP, Kay R, Seager TP, Linkov I. 2006. Objectives and Performance Metrics In Oil Spill Response: The Bouchard-120 And Chalk Point Spill Responses. SERI Report 06-001. Social and Environmental Research Institute: Greenfield, MA. Available at www.seri-us.org/publications.html