GREAT LAKES
POLICY RESEARCH NETWORK

FINAL REPORT
Testing a Framework for Assessing the Effectiveness of Programs and Other Measures under the Great Lakes Water Quality Agreement

Prepared by:

Dr. Carolyn Johns
Director, Great Lakes Policy Research Network
Associate Professor, Department of Politics and Public Administration
Ryerson University

Dr. Debora VanNijnatten
Associate Professor
Department of Political Science and North American Studies
Wilfrid Laurier University

Adam Thorn
PhD Candidate
Policy Studies Program
Ryerson University

October 30, 2015

CONTENTS

LIST OF ACRONYMS

AP	Actual Performance
CO	Collective Optimum
CSO	Combined Sewage Overflows
DFO	Department of Fisheries and Oceans
EPA	Environmental Protection Agency (U.S.)
GLEEM	Great Lakes Environmental Effectiveness Metric
GLPRN	Great Lakes Policy Research Network
GLRI	Great Lakes Restoration Initiative
GLWQA	Great Lakes Water Quality Agreement
GO	General Objective (Great Lakes Water Quality Agreement)
IDEP	Illicit Discharge Elimination Program
IJC	International Joint Commission
MOE	Ministry of Environment (Ontario)
MNR	Ministry of Natural Resources (Ontario)
MS4	Municipal separate storm sewer system (US)
NGO	Nongovernmental Organization
NPDES	National Pollutant Discharge Elimination System
NR	No-Regime Counterfactual
NRDC	Natural Resources Defense Council
PEI	Program Effectiveness Indicator
QQEF	Quantitative-Qualitative Effectiveness Framework
SOLEC	State of the Lakes Ecosystem Conference
USACOE	United States Army Corps of Engineers
WWTP	Wastewater Treatment Program

EXECUTIVE SUMMARY

Program effectiveness and performance measures are now core components of accountability regimes and progress reporting in public policy and public administration. Governments and international organizations around the globe are developing more sophisticated evaluation and reporting frameworks for a wide range of governance regimes.

The International Joint Commission (IJC) is a binational organization that 'prevents and resolves disputes between the United States of America and Canada under the 1909 Boundary Waters Treaty and pursues the common good of both countries as an independent and objective advisor to the two governments'.[1] Since 1972 the IJC has been responsible for reporting on progress by the governments toward restoring and protecting the integrity of the waters of the Great Lakes basin ecosystem under the Great Lakes Water Quality Agreement (GLWQA).

As part of its triennial reporting commitments under the 2012 GLWQA, the IJC has been engaged in efforts to improve evaluation and reporting, outcomes and progress related to the General and Specific Objectives in the Agreement.

This study tests an evaluation and assessment framework developed for the IJC by Hill and Eichinger (2013). The project had two goals:

1. Test the proposed framework through its application to an assessment question related to Great Lakes Water Quality Agreement objectives, and
2. Provide advice to the Commission on the framework's suitability for supporting the Commission's Great Lakes Triennial Assessment of Progress Report.

This report tests the recommended approach by collecting evaluation data using two on-line surveys related to two of the nine General Objectives in the GLWQA: General Objective (ii) and General Objective (vii).

The findings indicate the evaluation framework is sound and applicable across the two selected General Objectives in the GLWQA, and could be effectively used to assess the achievement of the other General Objectives in the GLWQA. However, some challenges and limitations related to applying the framework need to be considered. The report contains 8 recommendations related to the testing of the framework on the selected General Objectives and use of the framework in future triennial progress reporting on the GLWQA.

[1] International Joint Commission 2015. *About the IJC*, http://www.ijc.org/en_/About_the_IJC

INTRODUCTION

Under the Great Lakes Water Quality Agreement (GLWQA), the International Joint Commission (IJC) is charged with evaluating the extent to which government programs and other measures are achieving the objectives of the Agreement.

In 2013, the IJC commissioned a literature review of relevant program evaluation approaches and methodologies used in similar environmental policy contexts. Based on this literature review and evaluation of possible models, the contractors (Hill and Eichinger) developed a proposed framework in their report "A framework for assessing the effectiveness of programs and other measures developed to address the objectives of the Great Lakes Water Quality Agreement".

The proposed framework combines quantitative and qualitative methods in a Quantitative-Qualitative Effectiveness Framework (QQEF), and includes an approach that converts qualitative expert survey responses to a score that attributes the contribution of programs and other measures to progress toward objectives of the Agreement through a Great Lakes Environmental Effectiveness Metric (GLEEM) score.

The Commission decided to test the framework to determine its suitability for supporting the Commission's 2017 Great Lakes Triennial Assessment of Progress Report, which is called for by the Agreement.

The stated objectives of the project were to:

1. Test the proposed framework through its application to an assessment question related to Great Lakes Water Quality Agreement objectives, and;
2. Provide advice to the Commission on the framework's suitability for supporting the Commission's Great Lakes Triennial Assessment of Progress Report.

This report is based on a research design developed to test the proposed Quantitative-Qualitative Effectiveness Framework (QQEF) and the model to generate Great Lakes Environmental Effectiveness (GLEEM) scores proposed by Hill and Eichinger. The approach flows directly from the General and Specific Objectives that are clearly stated in the Great Lakes Water Quality Agreement (GLWQA). Given the scope of this project, and ultimate reporting requirements in Article 7.1(k), particularly part (iii) related to triennial assessment of progress on the extent to which programs and other measures are achieving the General and Specific Objectives of the GLWQA, all 9 of the General Objectives in the GLWQA were the starting point for determining which of the objectives might be candidates for testing the Hill-Eichinger framework and analyzing the potential of using this framework for IJC reporting purposes related to the Agreement.

RESEARCH DESIGN & METHODOLOGY

Hill and Eichinger outline the following 4 essential steps related to the QQEF and subsequent generation of GLEEM scores[2]:

1) choosing indicators;
2) assembling the data collection tool;
3) identifying expert interview/survey groups; and
4) conducting appropriate data analysis.

One critical step prior to the four stages outlined by Hill and Eichinger was selecting which objective(s) in the GLWQA to apply and test the QQEF and GLEEM score framework.

This section of the report is thus structured in five main sections, each with sub-sections, reflecting the phases of the research design and execution, as well as the 4 essential steps outlined by Hill and Eichinger in order to test the framework. These same five sub-sections are then used to analyze the application steps suggested by Hill and Eichinger and discussed further in the methodology analysis section of the report.

Phase I - Selection of Test Cases & Research Design

The project to test the QQEF and GLEEM Framework was designed and executed over several phases (see Appendix I). The first step of Phase I involved a 1-day meeting with IJC staff to discuss the scope of the study and determine which of the GLWQA objectives and related indicators the project should use to test the framework. In consultation with IJC staff, we discussed several different approaches to selecting a test case or cases.

Given that there are 9 General Objectives, one approach considered was to test the framework related to all of the 9 Objectives. Time and budget limitations did not make this option feasible. Another approach discussed was to select just one of the 9 General Objectives to see whether the framework, methodology and generation of GLEEM scores could, in fact, be applied to a real test case. A third approach was to select more than one case for testing the framework, in order to gauge whether the framework applies equally well across different Objectives. In consultation with IJC staff, we selected option 3.

[2] Hill and Eichinger 2013, 33-42.

The starting point for our rationale in recommending the selection of two of the General Objectives relates to the stated objectives of this project (above) and the ultimate use of the findings from testing the framework. If the primary goal is to make inferences or generalize about the applicability of the QQEF framework and GLEEM score to the General Objectives ultimately related to GLWQA reporting requirements in Article 7.1 (k), we felt a focus on at least two cases was required. This would allow us to test the utility of the framework, as well as the related QQEF methods and GLEEM scores, in two different policy and program areas. As part of testing the framework, it is important to understand the degree to which it can be applied across the different General Objectives.

How, then, do we go about selecting cases for comparative analysis that differ in these important respects? Social scientific methodology would suggest we employ a Most Different Systems Design that is theory-driven using key cases. This research approach tries to compare a small number of cases that are different in most respects on all but the variable of interest (dependent variable). The dependent variable in this study is the level of achievement/outcomes associated with each General Objective. While we cannot apply this approach rigorously in this project, our recommendation was to test the framework on two different cases to provide insight into whether certain factors seem to be associated with variation in perceived achievement and outcomes.

We used several criteria for selecting two of the General Objectives to test the model:

a) General Objectives where existing indicators have been used in IJC reporting,[3] or are in use by the Parties[4], selecting one General Objective for which IJC already has indicators in use and one for which it does not;
b) General Objectives for which some foundational indicators work has been completed[5], and that have been considered as Program Effectiveness Indicators (PEIs) in recent work by IJC[6];
c) Selection of one test case with clearly identifiable programs and measures associated with a General Objective, and one with less clearly defined programs and measures associated with a General Objective;

[3] For example, International Joint Commission 2011. *Assessment of Progress Made Toward Restoring and Maintaining Great Lakes Water Quality Since 1987*, Draft Report, October 2011. International Joint Commission 2013. *16th Biennial Report on Great Lakes Water Quality and Accompanying Technical Reports*, April 2013.
[4] US. EPA. National Water Program and Great Lakes Restoration Initiative Indicators; Environment Canada's Canadian Environmental Sustainability Indicators
[5] International Joint Commission, *Internal Draft Report on Program Effectiveness Indicators*, June 2014.
[6] International Joint Commission, *Program Effectiveness Indicators Workshop Report*, March 2014.

d) General Objectives that are clearly associated with programs and other measures in Annexes; in this respect, selection of one General Objective with an associated Annex and one with no associated Annex, to test whether the framework and GLEEM are broadly applicable or only related to General Objectives with clear program priorities and programmatic efforts through Annexes;
e) General Objectives for which the IJC has completed comprehensive program inventories[7];
f) General Objectives that have had a longer vs. shorter history of implementation (i.e., older vs. newer challenges, older vs. newer Annexes);
g) General Objectives that have a wide range of identifiable participants to survey;
h) General Objectives for which the survey team members have particular knowledge and expertise.

We conducted a review of the 9 General Objectives using these criteria (see Appendix II.) It was clear from the review and discussion with IJC staff that the 9 General Objectives differ in important ways, in particular: the extent to which they are clearly defined and targeted; the number of identifiable programs associated with each Objective; the availability of previous indicator work and program documentation; and, the cohesiveness and 'identifiability' of potential survey participants implicated by that Objective.

Based on our discussion with IJC staff and the tabling of several additional criteria during our November 2014 meeting, including consideration of the draft PEI workshop report and recommendations to prioritize certain PEI indicators, we selected two of the General Objectives to test the QQEF and GLEEM framework:

GO (ii) ***that the waters of the Great Lakes should allow for swimming and other recreational use, unrestricted by environmental quality concerns*** **and;**

GO (vii) ***that the Great Lakes should be free from the introduction and spread of aquatic invasive species that adversely impact the quality of the waters of the Great Lakes.***

The choice of GO(ii) and GO(vii) provided variation on several of the criteria above, including the clarity of the stated goals in each Objective, the duration and history of implementation, the number of associated programs and measures, the existence of a related Annex, and the number of indicators currently in use.

[7] For example, Dupre, S. 2013 *An Inventory of Nutrient Management Efforts in the Great Lakes*, prepared for the International Joint Commission's Lake Erie Ecosystem Priority Management Team, March 2013.

The GO(ii) test case is one where the goal is clearly stated, includes a dashboard measure (swimmable waters), has two of the earliest measures of Great Lakes environmental protection efforts already in use in IJC reporting (beach advisories and beach closures), has the recent addition of a third indicator in use (number/percentage of beaches open and safe for swimming), and has a number of programs and measures (direct and indirect) related to achieving the Objective. Yet there is no associated Annex with detailed programs and other measures.

GO (vii), by comparison, has a set of goals that are more broadly framed, involves more recent policy efforts; has a larger number of programs and measures associated with it (direct and indirect); and has a designated Annex on aquatic invasive species (Annex 6 in the GLWQA). Some indicators are identified in the PEI workshop report but no performance indicators have been used related to this GO in IJC reporting.

Both GO (ii) and GO (vii) involve subject matter with which the researchers were familiar, both are the subject of foundational indicators work by the IJC, and for both General Objectives a wide range of participants could be identified for survey purposes.

Testing the QQEF and GLEEM score framework on these two cases allows us to compare the utility of the framework across General Objectives in the GLWQA. We can develop and test several hypotheses related to these differences. For example, we might expect to test some of the following:

i. given the longer-standing goal and longer-term focus on GO (ii), that it would garner higher levels of achievement scores from expert assessments;
ii. given the more recent and more diverse set of programs and measures associated with General Objective (vii), one would expect lower levels of achievement assessments by experts;
iii. given the smaller number and longer-standing use of key indicators related to GO (ii), expert respondents would express a higher level of support for the indicators;
iv. given the association of an Annex with General Objective (vii), one would expect some consensus on the suite of indicators used to measure progress
v. given the longer-standing goal associated with General Objective (ii) one might expect a dedicated group of experts engaged in programs and measures related to this objective, with a range of organizational affiliations and with a higher number of dedicated hours to the achievement of this objective, compared to General Objective (vii);

vi. given the association of an Annex with General Objective (vii) one would expect an identifiable and dedicated group of experts engaged in programs and measures related to this objective.

While testing comparative hypotheses is not required for the QQEF or GLEEM scores proposed by Hill and Eichinger, this allows for the possibility of further testing the authors` assertion that the QQEF promotes valuable comparisons of effectiveness across GLWQA General Objectives.

Finally, in addition to discussion and selection of the test cases in Phase I, the outcomes of the meeting with IJC staff included: i) identification of the indicators for which all available information would be collected; iii) a preliminary QQEF research design, including an online survey and qualitative questions; iv) preliminary sampling criteria to be used as the basis for generating the purposive sample of participants; and v) foundations for research and development related to the indicator backgrounder suggested by Hill and Eichinger so that all study participants would be presented with an indicator summary for the Objective they are being asked to evaluate.

Phase 2 – Survey Design & Implementation

On-line Survey

Hill and Eichinger recommended the use of a survey to collect data related to achievement of GLWQA Objectives, combining both quantitative and qualitative questions. In their proposed framework, Hill and Eichinger recommended that 3 questions[8] form the foundation of the survey and generate the necessary data to test the model, based on the Oslo-Potsdam model. Q1 relates to achievement of the General Objective, Q2 the contribution of existing programs and measures to that achievement and Q3 the degree to which the indicators accurately demonstrate the achievement of the given Objective.

Q1 asks participants to assess the current state of the Objective using an ordinal scale.[9] It is intended to provide data measuring the perceived current condition of the Objective (actual performance) = AP in the GLEEM model. A coefficient of

[8] The exact number, set and sequencing of questions was not clearly outlined and needed to be determined. For example, Q2 and Q3 on p.37 of the consultant's report contained multiple questions that need to be disaggregated.

[9] the scale is not clearly outlined in the Hill and Eichinger report; we assume that 0 is assigned a value of very low perceived level of accomplishment; 10 very high level of perceived accomplishment

variation would then be generated for the model application stage. Given the survey response ranges and differences using a 10 point scale, we adapted this scale to 0-5 and then re-weighted to arrive at a value out of 10 in order to apply the GLEEM formula.

Q2 asks participants to identify and explain the perceived contribution of programs and other measures to the Objective. Hill and Eichinger recommended this be a qualitative question, specifically an open-ended survey question to be coded by a small, independent team using a scale adopted from the evaluation literature. The coded, open-ended questions would then be given a quantitative weight using the scale from 0 (no contribution) to 1.0 (full/complete contribution). Instead, we recommended use of a quantitative question and a 7-point scale, followed by open-ended questions, rather than assigning values to qualitatively coded responses for calculation of the GLEEM score.

Q3 is designed to collect quantitative and qualitative data on 'the indicators' provided to participants in the information package, and solicit their perspectives on the degree to which the indicator(s) accurately demonstrate the condition of the given Objective.

Questions 1 and 2 of the survey are used as the basis for calculating the GLEEM score for that Objective, using the formula adopted from Hill and Eichinger.

In addition to the required questions to test the framework, we also included several respondent-specific questions and several open-ended questions throughout the survey to collect additional context and information from respondents.

Indicator Backgrounders

A significant part of Phase 2 was generating the survey baseline backgrounders on the overall condition related to the Objective(s). Hill and Eichinger recommended the inclusion of this background research to ensure that participants had baseline information related to the evaluation, and possessed similar levels of knowledge on the state and nature of the Objective under consideration by the survey.
IJC, SOLEC, government documents and scholarly publications related to Objectives (ii) and (vii) were collected. The backgrounders included identification of the programs and other measures currently in place to contribute to the Objective. Drafts of these backgrounders were reviewed by IJC staff and finalized as pages 1 & 2 of each survey (see Appendix IV). One challenge we encountered here was to provide the key background and baseline information without making the opening of the survey onerous and negatively affecting response rates.

Study Participants

A non-random, purposive sample was recommended by Hill and Eichinger, including experts with 'insider knowledge' and 'outsider perspectives' to assess the achievement of a given Objective, the contribution of existing programs and measures, and the current indicators in use related to that Objective. We adopted this approach using a purposive elite sample for this study.

The team, in consultation with IJC staff, generated the lists of survey participants using the following criteria:

- Including representatives from all levels of government (federal, state/provincial, local) with mandates and responsibilities for implementing the GLWQA, including Indigenous communities and local bodies and authorities;
- Including representatives from umbrella organizations in the Basin (such as Great Lakes-St. Lawrence Cities Initiative, Council Great Lakes Governors, Conservation Ontario);
- Including participants from specialized watershed agencies (Conservation Authorities, Sea Grant organizations, extension programs, etc.);
- Including participants/experts from relevant private sector and non-governmental organizations; and
- Including academics with some expertise related to the selected Objective.

The first stage of generating a sample was to use the Great Lakes Policy Research Network (GLRPN) database of 900 policy actors in the Great Lakes derived from several participant lists collected from Great Lakes events and meetings from 2010-2014. The list includes a comprehensive representation of individuals and organizations with Great Lakes policy and program implementation mandates and responsibilities, particularly related to the GLWQA. It includes members of the Great Lakes Executive Committee, all Annex leads, a comprehensive list of state and provincial policy and agency actors, leaders from other public authorities, Indigenous leaders, leaders from a variety of non-governmental organizations and other stakeholders. The database is an important and valuable asset that was reviewed, supplemented and updated to generate a sample for this analysis. The database also includes email addresses, which made an online survey our preferred data collection method.

A Research Assistant was then hired to conduct a search of other relevant lists and publications related to GO(ii) and GO(vii) to enhance the participant lists. IJC staff also provided lists related to the two selected Objectives.

Draft lists were then provided to IJC staff for review. At the request of U.S. EPA, we removed Great Lakes regional staff from Region V from both participant lists. This may have implications that are discussed further in the findings sections.

The resulting participant lists (see Appendix IV) included:

- 104 Participants with expertise related to General Objective (ii)
- 98 Participants with expertise related to General Objective (vii)

Research Ethics

Once the research design survey instrument had been finalized, and the sample of participants determined, the team developed a draft recruitment package and draft survey instrument in consultation with IJC staff. This documentation was also submitted as part of the research ethics process at Ryerson University. The review of the recruitment package and draft survey instruments took approximately 4 weeks.

During the ethics review process, the survey instruments were also pre-tested by IJC staff. The surveys were approved by the Ryerson Research Ethics Board in March 2015, and administered on May 6, 2015 using *Fluid Surveys*.

Phase 3 - Data Aggregation and Analysis

In the four-week period that the survey was fielded in May 2015, the Research Assistant prepared SPSS and a codebook for the quantitative survey data (see Appendix V). In the month during which the survey was in the field, three reminders were sent in order to boost response rates. Fluid Surveys only sends reminders to those who have not yet completed the surveys. The reminders were effective and raised the response rates after each reminder was sent.

After three reminders the following response rates resulted for the two surveys:

General Objective (ii) = 33/104 = 32% response rate
General Objective (vii) = 41/98 = 42% response rate

Generally, the response rate was above-average for an on-line survey conducted for social scientific research and use of software tools like Survey Monkey and Fluid Surveys.[10] However, given the expertise and engagement of the potential

[10] Social science online survey response rates are typically in the 25-30% range, depending on the

participants in GLWQA-related work, we had expected to receive higher response rates - in the 50% range. This is an issue we will discuss further in the Analysis of Methodology section below.

The survey was closed on June 5, 2015. Data aggregation and analysis began on June 8. On June 10, a workshop was held with the three members of the research team and PhD students from the Policy Studies program at Ryerson University related to coding the open-ended questions. The workshop focused on coding Q1.1. The primary purpose of the workshop was to help students learn about the coding process, but a secondary goal was to get the team started on the coding of the open-ended questions and test for inter-coder reliability.

Workshop participants were broken into two groups, one focused on GO (ii) and the other focused on GO (vii). All workshop participants were given Q1.1 to code using a three step process: 1) initial review and clustering of responses, 2) identification of key categories and themes in coding key and 3) coding of all of the responses. Then all the groups shared their categories and findings.

While there was some inter-coder reliability, for several of the categories, one of the groups had 3-4 main coding categories and the other 7-8 categories. To some degree this is a function of how many open-ended responses were provided. There were 36 open-ended responses to GO(ii)2, Q1.1 (see Appendix VI) and the coding workshop resulted in testing inter-coder reliability. While different groups came up with different keywords to tag and categorize the comments, there was consensus on 4-6 key findings from Q1.1 from the GO (ii) and GO (vii) surveys.

Following the workshop, two team members then independently coded and wrote the summaries for the remaining open-ended questions. Given the small number of open-ended responses (see Appendices VI and VII), the use of a qualitative textual analysis software package such as NVivo was not required.

---

scholarly source cited related to survey response rates. Fluid surveys states 25% based on the general public surveys administered online through their site, See Penwarden 2014 from Fluid Surveys http://fluidsurveys.com/university/response-rate-statistics-online-surveys-aiming/

FINDINGS

In this section, a summary of the findings is presented. This section is divided into two main parts: the first presenting findings related to General Objective (ii) and the second presenting findings related to General Objective (vii). Each section includes three subsections. First, the summary data are presented, interpreted and analyzed for the evaluation and indicator questions. Second, a summary of the respondent data are presented and interpreted. Finally, each section ends with the cumulative GLEEM score as well as the GLEEM scores for each of the programs and measures associated with the given General Objective. Some preliminary interpretation and analysis in the context of the given Objective is included in this section. A comparative analysis of findings across the two General Objectives and two surveys is then provided.

General Objective (ii)

For GO (ii), 37 participants started the survey and consented to participation. However, only 33 participants went on to complete the survey and provide valid responses (19 from Canada and 14 from the U.S.). There is also some variation in the number of valid responses by question as some participants chose not to respond to some questions.

Achievement of General Objective

Question 1:

> On a scale of 0 to 5, where do you think we are in terms of achieving General Objective (ii) of the Great Lakes Water Quality Agreement - *that the waters of the Great Lakes should allow for swimming and other recreational use, unrestricted by environmental quality concerns?*

Response Scale: not achieved at all (0); very little achieved (1); some achievement (2); partially achieved (3); mostly achieved (4) and fully achieved (5).

Measures of Central Tendency and Standard Deviation for GO (ii)

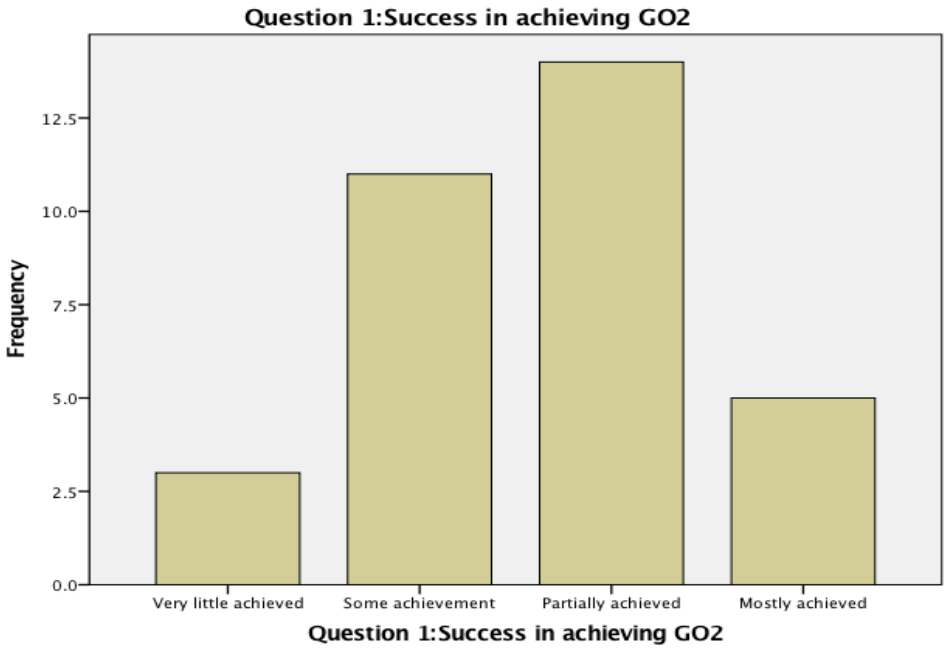
N	Valid	33
	Missing	4
Mean		2.64
Median		3.00
Mode		3
Std. Deviation		.859
Variance		.739

The mean response for Q1 was 2.64 meaning that, on average, respondents assessed the achievement of GO (ii) between some achievement and partial achievement. In their proposed framework, Hill and Eichinger recommended including a coefficient of variation as a measure of dispersion of the responses received. Standard deviation and variance are statistical measures of dispersion used to assess how far away individual data points are from the mean, or average, within the data set. Here we use the standard deviation. In the case of GO(ii), the standard deviation is .859, meaning that 85.9% of responses fall within the range of plus or minus one standard deviation around the mean of 2.64 (within 1 response below [1.64] or 1 response above [3.64]). In this case, there is a high degree of agreement among respondents that this Objective has been some or partially achieved. This is also evident in the frequency table.

Question 1: Achievement of GO (ii)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not achieved at all	0	0	0	0
	Very little achieved	3	8.1	9.1	9.1
	Some achievement	11	29.7	33.3	42.4
	Partially achieved	14	37.8	42.4	84.8
	Mostly achieved	5	13.5	15.2	100.0
	Fully achieved	0	0	0	100.0
	Total	33	89.2	100.0	
Missing		4	10.8		
Total		37	100.0		

The most frequent rating was 'partially achieved' – 37.8% of respondents indicated this Objective had been partially achieved. The next most common response was some achievement with 29.7% of respondents making this assessment. However, it is notable that only 13.5% indicated this Objective as being mostly achieved. None of the respondents indicated the Objective was not achieved at all, and none of the respondents indicated that this Objective has been fully achieved. Also notable is that 4 respondents did not provide a response to this question.



The Figure above clearly summarizes that most participants were of the opinion that General Objective (ii) had been partially achieved or some achievement has been attained.

Question 1.1 asked respondents to indicate whether they had any explanation or justification for their assessment.

This open-ended question generated a range of comments. Of the total of 37 respondents to the GO(ii) survey, 20 respondents provided comments related to their assessments of achievement of General Objective (ii).

In coding the 20 responses to Q1.1, the most common categories of responses related to reasons for achievement or lack of achievement including: need more/better measures and data (indicators good/bad/data quality and quantity; indirect causes) (8 respondents); system/indicator variation (5 respondents); new challenges and sources (4 respondents); and the need for more effort (3 respondents), in that order.

The most common issue identified related to current measures of achievement for this this Objective including: "We are using an inaccurate measure. … we do not know to what extent environmental quality concerns actually restrict recreation. We need a better indicator." Another commented that, "Beaches are still posted with

advisories or closed too frequently. And the data we use to post beaches with advisories, or close them, should be real-time data. This technology is available and should be used to test all Great Lakes beaches." In this vein, respondents noted that, "The tools are available"; and "identification of the watershed with all point and non-point source pollution is aiding in water quality/predictive modelling studies".

Others based their assessment of achievement on other measures: "There are quite a few 'hits' on the health unit beach website" and "GLRI provided significant funding to remediate beaches that reported the most closures/advisories" indicating that their assessment was based on those areas and beaches that received targeted Great Lakes Restoration Initiative funding. Another respondent commented that achievement needs to be better measured in terms of public awareness noting "the success of this GLWQA Objective can be measured by rising level of awareness within the general public".

Variation related to geography, seasons and timing of data collection were also noted by several respondents in comments such as: "It depends on the lake and locations greatly"; "I don't see a general decrease in beach closures and advisories, only inter-annual variation. Successes tend to be local"; and "recreational water quality varies between the different points we sample".

There were also several who commented on new sources and challenges: "Many of the reasons that beaches were closed/posted in the early stages of the Agreement have been adequately dealt with (poorly treated wastewaters, combined-sewer overflows etc….beaches may be closed due to other reasons, fecal sources (birds)…climate change…more frequent weather effects"). Others noted new challenges, "in Ontario by-passes of sewage/partially treated sewage after heavy rainfalls; private sewage disposal systems have no on-going monitoring program". Another respondent noted, "the Agreement acknowledges support for work on existing threats, namely blue-green algae blooms" but noted new challenges and threats, "land use changes, population increases are adding more load to the system".

Some comments related to outstanding effort required to achieve this Objective, i.e., "There is still a lot of work that needs to be done", "more work needs to be done", and "we have some work to do to reach an all around quality of water that is acceptable for swimming".

Finally, some commented on the need to focus more broadly on what other non-governmental organizations are doing related to achievement of this General

Objective. One such comment was, "The Swim Guide[11] also publishes reports on year to year changes. I know the data well and understand how poorly we are doing trying to meet the goal of swimming the Great Lakes." This particular respondent went on to provide the actual data that underpinned his/her assessment of Q1.

In summary, the most common comments related to reasons respondents arrived at their general assessment of whether General Objective (ii) has been achieved were:

- existing measures are insufficient/need different measures
- there is a need for more/better data collection & dissemination
- current variation in indicators and data collection is an issue
- there are new threats/problems/challenges (new sources, climate change)
- there has been increased non-government participation and need for inclusion of public/NGOs in measures
- more needs to be done

Finally, there is some recognition that measuring achievement is challenging. As noted by one respondent: "The issue is very complex. I don't think we will ever truly achieve this objective." The fact that none of the respondents indicated full achievement in response to Question 1 reflects recognition that both achievement and measuring achievement are challenging.

Contribution of Various Programs and Measures to General Objective (ii)

Question 2 in the survey asked respondents to indicate the contribution of various programs and measures to achieving General Objective (ii) using a 7-point scale from no contribution to complete contribution: no contribution (0); low contribution (0.15); low-medium contribution (.35); medium contribution (.50); medium-high contribution (.65); high contribution (.85) and complete contribution (1.0).
This is in keeping with the suggested weighting of responses by Hill and Eichinger [12]..

Eight programs and measures were included in this question for General Objective (ii) (labeled 2a through 2h in the tables below). The list of programs and measures covers all the major government and non-government programs and measures related to recreational and beach water quality.

[11] The Swim Guide is produced by Waterkeeper organization based on data from government and non-government organizations available in a web-based format and via an mobile app. For more information see https://www.theswimguide.org/guide/about/

[12] see weighting recommended by Hill and Eichinger p. 39

This question generated 30 valid responses. Appendix VIII provides frequency tables for each of the programs and measures. Summary statistics are provided and reviewed here.

	Program/Measure
2a	source water protection programs
2b	waste water treatment programs
2c	nutrient management programs
2d	performance based watershed plans/programs
2e	local water quality monitoring programs
2f	public health monitoring programs
2g	NGO/community monitoring and reporting programs
2h	NGO certification/flag designation programs

Question 2: Summary Statistics									
		2a	**2b**	**2c**	**2d**	**2e**	**2f**	**2g**	**2h**
N	Valid	30	30	30	30	30	30	30	30
	Missing	7	7	7	7	7	7	7	7
Mean		2.80	3.80	3.47	3.33	3.43	3.77	2.47	2.73
Median		3.00	4.00	3.50	3.50	3.00	4.00	3.00	3.00
Mode		3	5	5	4	3	3	3	3
Std. Deviation		1.606	1.349	1.432	1.269	1.073	1.135	1.074	1.230
Variance		2.579	1.821	2.051	1.609	1.151	1.289	1.154	1.513

It is evident that wastewater treatment programs received the highest contribution ratings from respondents followed by public health monitoring programs. The average assessment for wastewater treatment programs was 3.8 /7 which overall is medium (.54) and 3.77 for public health monitoring programs which is also medium (.538) and just slightly below the mean assessment for wastewater treatment programs. As the summary data below and in Appendix VII outline, wastewater treatment programs, public health monitoring programs and nutrient management programs also receive several high contribution assessments from a number of respondents.

Question 2: Summary of Contributions of Programs and Measures

The table below summarizes the general findings related to the assessment of various programs and measures. In some instances, the frequency tables indicate

considerable agreement. For some programs the assessments are mixed, in that there is no consensus and the distribution of responses ranges from low to high. For some of the programs the most common assessments are split between two main responses. For General Objective (ii), source water protection programs, wastewater treatment programs and nutrient management programs were assessed, on average, as making medium-medium/high contributions to achieving this Objective. Wastewater treatment programs received the highest number of high assessments (29.7%). Many of the programs were assessed as making medium contributions.

When combining assessments for medium, medium/high and high contributions, 67.5% indicated both local water quality monitoring programs and public health monitoring programs were in this range; 64.8% assessed wastewater treatment programs in this range; 59.4% assessed performance-based watershed plans/programs in this range and 45.9% assessed source water protection programs in this range.

Respondents, on average, assessed non-government, community monitoring and certification/designation programs as making low/medium contributions. There seem to be some notable differences between established government programs in 2a, 2b and 2c compared to more bottom-up, often non-government programs in 2d, 2e, 2g and 2h. Public health programs are also interesting in this regard as they are typically government programs in Canada and the US but are locally administered.

The frequency tables in Appendix VII also highlight that a very small number of respondents indicated that some programs and measures (2a, 2b, 2c) made no

	Program/Measure	**General Assessment**
2a	source water protection programs	medium-medium/high (35%)
2b	waste water treatment programs	medium-medium/high (35%) and high (29.7%)
2c	nutrient management programs	medium-medium/high (35%) and high (27%)
2d	performance based watershed plans/programs	medium/high (59.4% when combined)
2e	local water quality monitoring programs	32% medium; medium/high (67.5% when combined)
2f	public health monitoring programs	medium (27%); high (24%); medium/high 67.5% when combined
2g	NGO/community monitoring and reporting programs	medium (32.4%); low/medium (70% when combined)
2h	NGO certification/flag designation programs	Mixed – low-medium (18.9%); medium (21.6%) and medium-high (18.9%)

contribution at all, and some respondents indicated for these same programs that 2a and 2b made complete contributions to achieving General Objective (ii). None of the other programs and measures received: 'no' or 'complete' contribution assessments.

Other Programs and Measures Identified by Respondents

Question 2.1 asked respondents to identify any other programs and measures that contributed to the accomplishment of GO (ii) but were not included in the list provided (2a-2h), and to use the scale in Q2 to indicate the contribution of that program or measure.

Of the 33 valid respondents, only 12 provided comments related to this question and none used the scale as part of their response.

Several respondents commented that they felt stormwater management programs and combined sewer overflows (decoupling, presence of and interception, capture by CSOs) should have been included as a separate program (distinct from waste water treatment programs, a program which was included in the program list). In this respect, respondents commented that the following stormwater measures and programs should be separated out:

- Interception and treatment of direct stormwater discharges to beach/nearshore areas
- Interception and treatment of flows from rural sources
- enforcement of sanitation codes

In addition to stormwater programs, other programs identified included; other beach programs such as beach grooming, bird control programs; septic tank reassessment programs and community infrastructure improvement programs. One respondent noted that incentive programs to individuals/companies should also be included.

Two respondents were critical of the programs included. One respondent noted that monitoring programs "only provide already known information, does not really solve the problems". Another noted that, "I believe the above scale is too vague to be able to adequately assess the contributing *factors* to Great Lakes restoration. It's particularly problematic since it doesn't align to the Annexes, making it difficult to tie any of the progress into the GLWQA. Furthermore, it is isn't clear which programming is being referenced, and which actor for each component of programming."

Other NGO initiatives such as www.theswimguide.org were identified and other public sector initiatives such as the U.S. Great Lakes Restoration Initiative (GLRI), Conservation Authorities and Remedial Action Plans were also noted as being important but no specific programs or measures used by these organizations were identified related to beaches, swimming or recreational water quality. This indicates some distinctions between government and non-government and programs at various scales may need to be included if a more comprehensive list of programs and other measures were to be developed for this General Objective.

Question 2.2 asked respondents if they had any comments on specific programs that are particularly effective or to identify any gaps that exist.

Nine comments were received for this question (two additional responses indicated "No" and "None"). The most common comments focused on gaps, i.e., "Blue-green algae blooms need to be monitored in addition to bacteria" and more attention needs to be given to "the relative risk of eColi from animal discharges vs. those from human sewage" and "large scale polluters such as industry and municipal WWTPs". Enforcement and financial assistance programs were also noted as gaps. One respondent noted very specific programs in the US such as the NPDES program and the Illicit Discharge Elimination Program (IDEP) and municipal separate storm sewer system (MS4) programs as part of NPDES. Here, respondents were highlighting the need to move beyond the human sewage focus of current programs and measures.

One respondent focused on the need for a combination of water quality program measures and public health programs and measures without noting specifics. Another called for monitoring and modeling programs, explaining that "the other missing link is the epidemiology of illness outbreaks" though this kind of "epi study is very hard to link to beach use" and water quality.

Some of the comments in this section did not respond to the question but pointed to what might be regarded more broadly as governance issues. For example, one respondent commented, "there is no on-going trilateral process for both levels of government to engage with non-government organizations and First Nations governments on activities or policy on a regular basis". This may be a gap in the sense that the respondent is asking for some specific programmatic measure of engagement to be included related to the achievement of GO (ii). Similarly, another respondent noted that Niagara has a Water Strategy and "programs that benefit the achievement of this General Objective" but did not identify the programs, or indicate

if these were particularly effective or addressed a gap in the list of programs and measures included in Q2.

Suitability and Accuracy of Indicators for General Objective (ii)

Hill and Eichinger recommended the inclusion of a third critical question related to indicator appropriateness to serve as a confidence check on the indicators that were selected to represent the Objective. Question 3 asked whether the indicators in the backgrounder and listed in the question accurately demonstrate the condition of G02 today. There were three main indicators listed and respondents were presented with yes, no, and no opinion response options.

For this question, 28 valid responses were received (including 2 indicating no opinion related to the indicators) and 9 responses were missing. Of the 28 who did indicate their opinions on the three indicators, 40.5% indicated that the number/percentage of beach advisories per season was a good indicator related to achievement of General Objective (ii). However, 29.7% felt this was not a good indicator.

	Indicators	**General Assessments**
3a	Number/percentage of beach advisories per season	Mixed – 15 respondents (40.5%) felt this was a good indicator; 11 (29.7%) did not
3b	Number/percentage of beach closures per season	Mixed – 15 respondents (40.5%) felt this was a good indicator; 11 (29.7%) did not
3c	Number/percentage of days beaches are open and safe for swimming per season	Yes, the majority of respondents felt this was a good indicator

The opinions about the use of this indicator are therefore mixed. This also may have something to do with whether the respondents were from Canada or the U.S. as U.S. respondents would be more familiar with Indicator 3c and its use in reporting.

Similarly, the opinion about the number/percent of beach closures per season as a measure was mixed. The percentage of respondents who felt this was a useful and valid indicator was 40.5% and those who did not were 29.7%

Question 3a: Number/percentage of beach advisories per season

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	15	40.5	53.6	53.6
	No	11	29.7	39.3	92.9
	No Opinion	2	5.4	7.1	100.0
	Total	28	75.7	100.0	
Missing		9	24.3		
Total		37	100.0		

Question 3b: Number/percentage of beach closures per season

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	15	40.5	53.6	53.6
	No	11	29.7	39.3	92.9
	No Opinion	2	5.4	7.1	100.0
	Total	28	75.7	100.0	
Missing		9	24.3		
Total		37	100.0		

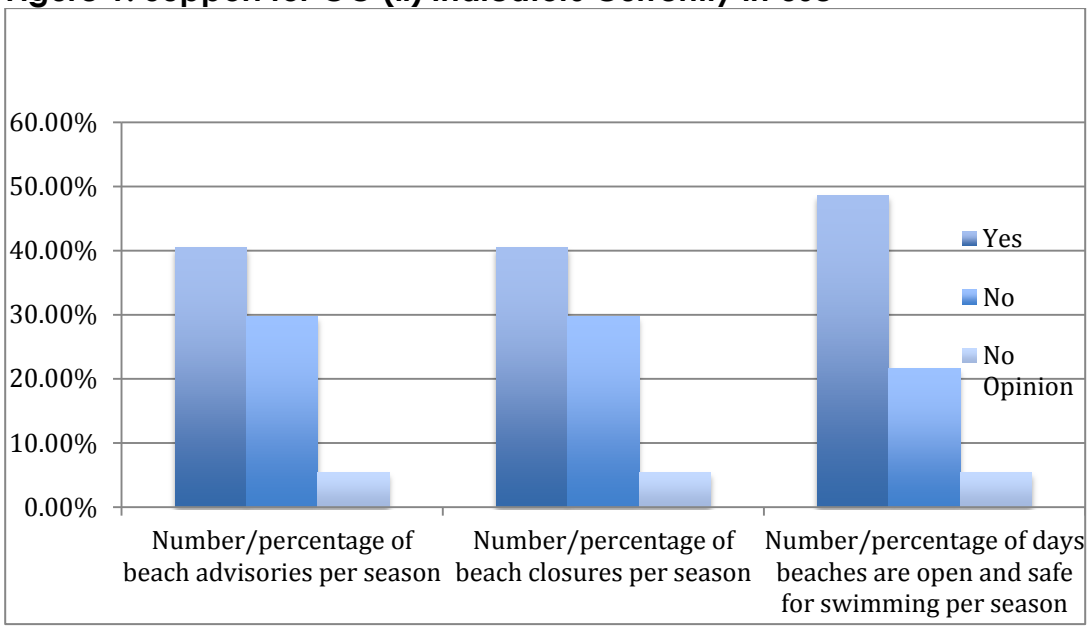
When comparing these two indicators with the more recent indicator adopted by the U.S. - namely, the number/percentage of beach days that are open for swimming per season - more respondents felt this was a good indicator related to General Objective (ii).

Question 3c: Number/percentage of days beaches are open and safe for swimming per season

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	18	48.6	64.3	64.3
	No	8	21.6	28.6	92.9
	No Opinion	2	5.4	7.1	100.0
	Total	28	75.7	100.0	
Missing		9	24.3		
Total		37	100.0		

A comparison of preferences for the three indicators currently in use [see Figure 1 below] indicates that almost half of most respondents (48.6%) feel the number/percentage of days beaches are open and safe for swimming per season is a good indicator related to General Objective (ii).

Figure 1: Support for GO (ii) Indicators Currently in Use



As summarized in the Figure above, it is clear that most of the respondents felt these three measures were satisfactory for measuring the achievement of General Objective (ii). Interestingly, most felt the more recent addition of number/percentage of days beaches are open and safe for swimming per season was a good indicator with almost 50% indicating this is the case. One might hypothesize that the positive response to this indicator might be a reflection of the U.S. respondents' opinions as this is the current indicator used in the US, yet more Canadian than U.S. participants responded to this survey.

The other notable finding is that 20-30% felt these are not satisfactory or good indicators related to achieving General Objective (ii). The qualitative comments in the next section are very valuable in indicating that there is no consensus, and that there are concerns about the data used, data collection and reporting related to these indicators.

As indicated by the following comments in response to Q3.1, standardization of indicators, data collection protocols and perhaps an index of these three indicators would strengthen the basin-wide reporting related to GO (ii).

Qualitative Comments Related to Indicators

Question 3.1 asked if respondents wanted to provide any other comments on the indicators listed in the survey related to GO (ii).

This open-ended question generated the second highest number of comments (compared with Q1.1) with 17 of 37 respondents providing comments on the indicators. In fact, some of the comments in response to this question generated similar coding categories related to problems with existing measures, lack of data and variation.

Expressing some similarity to comments received in response to Q 1.1, the most common category of comments about the indicators focused on problems with the existing indicators and their measurement (data collection; only snap shot data; time lags between data collection and action; need for real time data; poor longitudinal data, inaccurate data etc.). General comments related to the indicators being only as good as the data and measurements associated with them. As one respondent noted the indicators are good "if these advisories are based on the measurements'. Another noted, "it's only a snap shot in time and not showing real time data". These contributions clearly indicate a concern with the measures and data collection methods associated with the indicators. Others commented that, "this indicator [without specifying which of the three] does not directly measure health risk. Therefore we are overstating the risk. At this point it is our best option available but we need to develop an indicator that can be widely used that actually assesses health risk, not just E Coli or fecal coliform presence".

In addition, similar to the coded responses in Q 1.1, jurisdictional variation in the indicators was the second most common comment. For example, "Given that recreational water quality objectives can vary between jurisdictions, this information does not reflect the overall picture of water quality in the Great Lakes. A safe and open beach in one jurisdiction may have unacceptable water quality in another and would warrant an advisory. Nor does it provide a way to determine if water quality is improving or deteriorating over the years". Another commented that, "Additionally, many jurisdictions use a risk assessment approach to beach advisories, rather than a hard threshold value of E.coli presence. As such, some jurisdictions may have less frequent advisories than others".

Another respondent, while noting that "These indicators provide good information on the relative condition of beaches with some indication of year over year successes or long term trends", agreed that jurisdictional variation is an issue: "They should not be used to compare U.S. vs Canada given the difference in standards used to determine beach postings".

Related to comments about variation by jurisdiction, there were responses focusing on the frequency and lag time issue for the data that underpins the indicators. One respondent noted that there is a "need to add the number of days or frequency a beach is monitored, and the number of beaches monitored on each lake" and "Ontario data is misleading as many beaches are only monitored by public health on a weekly basis" causing lags in reporting and posting. This lag time between data/monitoring and decision-making/reporting was noted by two other respondents. One respondent commented on the need for monitoring on a daily basis. The prevalence of comments on variation in measures across jurisdictions, as well as data collection methods and frequency, indicate these are important limitations of existing indicators.

A third category of comments related to better indicators that could be used. One respondent commented that local factors such as high bird populations and frequent rain should be included in the indicators. One respondent noted "a better approach would be to list the beaches that report exceedances and review the data over the monitoring season. If the data shows spikes, then stormwater is more likely. If the data show consistently high levels, then chronic discharges are more likely. Better progress seems to happen when specific information is gathered and targeted".

Despite the fact that many respondents view the indicator of 'number of open beach days' as a good indicator (48.6%), one respondent commented: "I think it is more compelling to track advisories and closures than it is to track open beach days". Another explained that "open and safe implies a comprehensive assessment" and this is not the case. Yet another noted that, "a more accurate indicator would be an index of the three indicators and a risk estimate related to full exposure". A different respondent agreed, noting that "a more ideal indicator would be the three indicators above but based on a true pathogen measurement…and a risk assessment to full body exposure/full body immersion to human health. We are a substantial way from that ideal". Another respondent was very specific about the kind of indicator and data collection methods that should be used: "microbial source tracking (MST) is needed to identify the current sources of fecal contamination".

From the coding of responses there may also be two general groups of comments related to the background and expertise of the respondents. Clearly, some comments stem from a public health perspective and the human health goals associated with General Objective (ii), while others are more focused on water quality more broadly. This is supported by the fact that 18 of the 33 respondents were affiliated with public health organizations and 5 indicated their educational background was public health.

Other Indicators?

Question 3.2 asked respondents to list other indicators or data that should be considered to better evaluate and report on the state of GO (vii)?

Building on comments provided in 3.1 above, 14 respondents provided comments about other indicators or data that could be used to better evaluate General Objective (ii). Some responses to this question overlapped with comments provided for 3.1 above.

"We need ongoing research to develop a new indicator" and "recent research out of Niagara has shed some light on just how difficult predictive modeling can be. In a nutshell, each beach is different".

Some respondents offered straightforward suggestions to improve the current indicators: 'correlations with rainfall'; 'frequency of beach monitoring'; 'number of beaches per lake'; and 'changes in number of advisories", were all suggested as additions to existing indicators and their associated data. Others offered quite broad suggestions, "general health of the animals and flora of the lake". Another commented that, "levels of phosphate/nutrients/dissolved oxygen may be a viable indicator related specifically to recreational water concerns such as blue-green algae events". "A sanitary survey could identify which beaches are more vulnerable to pollution and which ones are not. Then target the beaches that are more vulnerable …identify the sources of pollution…then address the sources of pollution". Reiterating the response provided in 3.1, one respondent again suggested microbial source tracking.

An additional suggestion was that, "A better approach may be to list beaches that report exceedances and review the data over the monitoring season. Better progress seems to happen when specific information is gathered and targeted". This comment calls for some measure of problem beaches or areas over time, perhaps a measure of change in number of advisories over time. Indeed, the number of beaches

monitored and the frequency of beach water quality monitoring were mentioned by several respondents in Q3.1 and Q3.2. In the opinion of one respondent, "The bottom line is that water sampling for fecal coliform is perhaps the best we have currently" but the comments provided in response to Q3.1 indicate that more than half of the respondents feel that the indicators and measures could be improved and standardized.[13]

[13] Some of these comments support findings and recommendations from previous IJC reports including recommendations: to develop binational, standardized basin-wide surveillance and monitoring protocols in conjunction with preventive risk management strategies, that binational standardized criteria for beach postings be adopted and that a binational, systematic, centralized and timely way to evaluate and report waterborne illness in the Great Lakes and track what is happening on the local, regional, state, provincial and federal levels be developed (IJC 2009). Similar recommendations were made again in 2011. Another IJC working group made the recommendation to integrate Canadian and U.S. beach monitoring databases and develop approaches to effectively communicate relevant information about recreational water quality risks to the public (IJC 2011). While it is clear progress has been made in terms of communicating what data does exist to the public, the issue of standardized indicators, data collection methods, databases and data reporting endure according to the respondents to this survey.

Findings Related to Respondents

The remainder of the questions related to data collection on the background of respondents. Of the 37 total respondents, 27 opted to answer the person-centered, demographic questions, 10 did not. It is difficult to determine why 10 did not respond to these questions. The survey was not long and it was probably not for reasons of survey fatigue.

Question 4: Years of Experience related to GO (ii) and GLWQA

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than 12 months	2	5.4	7.4	7.4
	1-3 years	3	8.1	11.1	18.5
	4-6 years	2	5.4	7.4	25.9
	7-10 years	6	16.2	22.2	48.1
	11-15 years	4	10.8	14.8	63.0
	16-20 years	3	8.1	11.1	74.1
	21-25 years	1	2.7	3.7	77.8
	more than 25 years	6	16.2	22.2	100.0
	Total	27	73.0	100.0	
Missing		10	27.0		
Total		37	100.0		

First, the survey respondents for General Objective (ii) represented a broad distribution in terms of years of experience. The most common response was 7-10 years experience or more than 25 years experience (16%), followed by those with 11-15 years experience, when combined; some 37.8% had more than 10 years experience and 54% had more than 7 years experience. This is the vast majority of respondents when the fact that 10 respondents chose not to respond to this question.

In terms of the percentage of time that respondents spent engaged in work related to this General Objective of the GLWQA, the majority (54%) spent between 1 and 20% of their time on work related to this objective. Three respondents (8%) spent between 61-80% of their time engaged in work related to this objective. This information may be valuable in terms of follow up research, advisory roles, or key informant interviews related to this General Objective.

Question 5: Time Engaged in G0 (ii) Work related to GLWQA

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1-20%	20	54.1	74.1	74.1
	21-40%	3	8.1	11.1	85.2
	41-60%	1	2.7	3.7	88.9
	61-80%	3	8.1	11.1	100.0
	Total	27	73.0	100.0	
Missing		10	27.0		
Total		37	100.0		

Some 21.6% of respondents were answering this survey based on a basin-wide perspective. However, when combined, most were answering based on a lake specific perspective. The responses covered all of the lakes and 4 respondents selected 'other, please specify' reporting: Lake St. Clair; Georgian Bay; both basin wide and Lake Ontario; and inland lakes and Great Lakes beaches as the basis of their responses.

Question 6: Perspective and Scale

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Basin Wide	8	21.6	29.6	29.6
	Lake Erie	3	8.1	11.1	40.7
	Lake Ontario	5	13.5	18.5	59.3
	Lake Michigan	1	2.7	3.7	63.0
	Lake Huron	3	8.1	11.1	74.1
	Lake Superior	3	8.1	11.1	85.2
	Other	4	10.8	14.8	100.0
	Total	27	73.0	100.0	
Missing		10	27.0		
Total		37	100.0		

Question 7 asked respondents to identify the field or discipline that best describes their educational/training background. The most common response was science (32%), the second most common response was 'other' with 9 respondents (24%) listing fields other than those provided in the survey. Of those who responded 'other, please specify', 5 reported public health, and others reported environment and physics, environmental health, policy, and water quality microbiology/engineering.

Question 7: Educational Background of Respondents

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Science	12	32.4	44.4	44.4
	Engineering	2	5.4	7.4	51.9
	Medical	1	2.7	3.7	55.6
	Law	1	2.7	3.7	59.3
	Social Science	2	5.4	7.4	66.7
	Other	9	24.3	33.3	100.0
	Total	27	73.0	100.0	
Missing		10	27.0		
Total		37	100.0		

Question 8 asked respondents to identify the type of organization they worked for.

Question 8: Organizational Affiliation of Respondents

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	University/Research	2	5.4	7.4	7.4
	Local/Municipal government	8	21.6	29.6	37.0
	State/Provincial government	6	16.2	22.2	59.3
	Federal government	2	5.4	7.4	66.7
	First Nations/Metis/ Tribal/Indigenous	2	5.4	7.4	74.1
	Environmental NGO	3	8.1	11.1	85.2
	Other	4	10.8	14.8	100.0
	Total	27	73.0	100.0	
Missing		10	27.0		
Total		37	100.0		

The most common response was local/municipal government (21.6%). However, the respondents identified affiliations with a wide range of organizations.

The survey findings for GO(ii) are thus based on a fairly broad spectrum of experts from a range of organizations, government, non-government, Indigenous and universities/research institutions. This is positive in the sense that it does not provide an evaluation of the General Objective and effectiveness from the perspective of one sector or organizational perspective.

Of the 4 respondents who indicated 'other' for this question, 2 identified public health units, one health unit and one local health unit, so basically another 4 from a health organization perspective. Interestingly, respondents do not associate this with local/municipal government. The comparatively low number of federal government respondents may be a function of the numbers in the potential participant list who received the survey, the U.S. EPA request not to survey their Region V staff and the fact that recreational water quality and beaches are primarily a subnational responsibility related to the programs and measures. Also notable is that none of the respondents indicated industry/private sector or watershed/regional authority as their organizational affiliation.

Question 9 asked respondents to indicate whether they were affiliated with any of the Annexes in the GLWQA. Only two respondents indicated Annex affiliations: one of the respondents was affiliated with two of the Annexes 3 (Chemicals of Mutual Concern) and Annex 4 (Nutrients) and 1 respondent was affiliated with Annex 3 (Chemicals of Mutual Concern).

Question 10 asked respondents to identify, in no particular order, the top five experts related to this GLWQA General Objective. Some 29 experts were identified with some respondents identifying as low as 1 or 2 experts and others identifying 5. Three of the individuals listed by respondents are involved with the study, two are former IJC Commissioners, two are former staff members of the IJC Regional Office in Windsor and two are current IJC staff members. When cross-referenced with the participant list for General Objective (ii) in Appendix II, all of the top eight experts identified by more than one respondent were on the participant list and received an invitation to complete the survey - with only two exceptions. One is a member of GLPRN and thus a possible conflict might have existed with sending this person a survey. Also, two individuals were on the potential participant list generated by the researchers but were removed as U.S. EPA Region V, requested its staff not be surveyed.

Taking these exclusions into account, an additional12 people listed by respondents were not on the potential participant list. This is important information and the

identified experts should be added to the participant list for future research. In order to determine whether representation of some experts or organizations was an issue, we would need to determine the organizational affiliations of the missing experts. In some cases, one or two individuals from a given organization were sent the survey, and thus participants from some organizations were limited in number to ensure broad and diverse representation in the participant list.

Two respondents listed organizations instead of individuals with expertise (Surfrider Foundation; Department of Fisheries and Oceans; Ontario Ministry of Environment and Climate Change; Environment Canada; Conservation Authorities; Health Units - Ministry of Health and Long-Term Care) and one respondent listed areas of expertise sedimentation; hydraulic modeling; erosion).

GLEEM Scores for GO (ii)

As outlined by Hill and Eichinger, the effectiveness evaluation requires two pieces of information: 1) a quantitative measure of the overall condition of the Objective [achievement outcome]; and 2) an assessment of the role programs and other measures have played in contributing to that condition [combined and individually]. From the survey, responses to Questions 1 & 2 provide this information.

Q1 is designed to ask participants to assess the current state of the Objective using an ordinal scale. It generates data that measures the perceived current condition of the Objective (actual performance) = AP in the GLEEM model.

Q2 asks participants to identify and explain the perceived contribution of programs and other measures to the Objective. This question generates data that the model designers then suggest be coded (using a coding framework) and weighted (using an ordinal scale). Ultimately, the responses are assigned a quantitative weight, averaged and then a coefficient of variation is calculated for this question. This, in turn, is used to calculate the no-regime counterfactual (the estimated state of the Objective in the absence of any of the identified interventions, programs or measures).

The first two questions are then used as the basis of calculating the GLEEM score for the Objective overall, as well as for each of the programs and measures included in Q2.

The following steps are derived from Hill and Eichinger (2013, pp. 40 and 41) and revisiting the Dombrowsky formulas (2008) for calculating the GLEEM score overall, as well as the GLEEM scores for each program and measure.

1. The Actual Performance (AP) value is calculated by dividing the mean value of responses to Question 1 by 6 (number of possible responses) and multiplying by 10 to arrive at a value out of 10 (as Hill and Eichinger used a scale of 1-10 instead of our 0-5).

$$AP = (\bar{x}\ Q1\ \div 6) * \ 10$$

AP = (Mean Response Question 1 [2.64/6]* 10 (converted to value out of 10[14]) = **4.39**

[14] the scale recommended by Hill and Eichinger is not clearly outlined; they state it is a 0-10 scale with 0 indicating no level of accomplishment to 10 being complete accomplishment but no values for a survey question were provided as they recommended this be a qualitative question. We adapted this scale to 0-5 and then re-weighted to get value out of 10 in order to apply the GLEEM formula.

2. The No-Regime Counterfactual (NR) value for the overall GLEEM score is calculated by multiplying the Actual Performance (AP) by 1 – Q2 (perceived contributions of programs and measures).

$$NR = AP * (1 - Q2)$$

Given there are multiple programs and measures the calculation requires averaging the responses to questions 2a-2h[15] for each respondent (*y*), then determining the mean of all responses. This value is then divided by 7 (the number of potential responses) to provide a value from 0 (no contribution) – 1 (complete contribution) to get the value for Q2. This value is then subtracted from 1 (which represents a perfect score on the value), and finally multiplying that value by the AP value.

$$NR = AP * [1 - \left(\frac{\bar{x}\,[\bar{x}Q2(y)]}{7}\right)]$$

NR = AP [4.39] * (1 - Mean of Responses to Question 2a-h [3.25]/ 7=[.46])
NR = 4.39 (.54)
NR = **2.37**

3. Collective Optimum (CO) always equals 10.

4. The GLEEM score then is calculated using (AP-NR)/(CO-NR).
(*Perceived* Actual Performance – No Regime counterfactual) / (Collective Optimum - No Regime counterfactual)

GLEEM = (4.39 - 2.37) / (10 – 2.37)
= 2.02/7.63
= .265

The GLEEM score of .27 indicates that the perceived effectiveness of the listed programs and measures combined in achieving General Objective (ii) is 'low-medium' using the scale of: no contribution (0); low contribution (0.15), low to

[15] Hill and Eichinger recommended creating this value by coding qualitative responses and assigning a value from 0-1. We used a quantitative question (Survey Question 2) and use the mean of responses to all those questions divided by number of possible responses (7) to arrive at value from 0 to 1.

medium contribution (0.35), medium contribution (0.50), medium to high contribution (.65), high contribution (.85) and complete contribution (1.0).
When combined with the findings from Q1, which is an assessment of the achievement of the General Objective independent of the programs and measures, the GLEEM score does add some value to interpreting the degree to which the identified/current programs contribute to the achievement of the Objective.

GLEEM Scores by Program and Measure for General Objective (ii)

Hill and Eichinger also indicate that GLEEM scores and effectiveness evaluations can then be completed for each of the programs and measures within Objectives as well. To test this, separate calculations for each program and measure using a modified version of the formula is required

$$NR = AP * [1 - \left(\frac{\bar{x}\ Q2(y)}{7}\right)]$$

Here the average response for each program and measure is used (where y = 2a through 2h) divided by 7 (number of potential responses to Q2) then subtracted from 1 and multiplied by the overall actual performance value.

	Program/Measure	**GLEEM Score**
2a	source water protection programs	0.24
2b	waste water treatment programs	0.30
2c	nutrient management programs	0.28
2d	performance based watershed plans/programs	0.27
2e	local water quality monitoring programs	0.28
2f	public health monitoring programs	0.30
2g	NGO/community monitoring and reporting programs	0.22
2h	NGO certification/flag designation programs	0.23

GLEEM scores for each of the programs and measures are all close to the overall GLEEM score (.27). Although program-level GLEEM scores all indicate that various programs and measures contribute, somewhere between low and low-medium contributions to the achievement of the GO(ii), wastewater treatment programs and public health monitoring programs are view as contributing slightly more than other programs and measures and NGO programs contributing less.

These scores are not surprisingly in keeping with the descriptive statistics generated from Survey Question 2. Question 2 asked respondents directly about the contributions of various programs and measures to the achievement of the General Objective. Like responses to Q2, GLEEM scores at the program level allow for analysis and comparison of the various programs and measures to the achievement of GO(ii).

General Objective (vii)

Achievement of General Objective

Question 1:

> On a scale of 0 to 5, where do you think we are in terms of achieving General Objective (vii) of the Great Lakes Water Quality Agreement - *that the Great Lakes should be free from the introduction and spread of aquatic invasive species that adversely impact the quality of the waters of the Great Lakes*?

Response Scale: not achieved at all (0); very little achieved (1); some achievement (2); partially achieved (3); mostly achieved (4) and fully achieved (5).
For GO (vii), 24 of the responses received were from U.S. participants and 17 responses were received from Canadian participants for a total of 41 valid responses.

Q1: Measures of Central Tendency and Standard Deviation

N	Valid	41
	Missing	2
Mean		1.93
Median		2.00
Mode		2.00
Std. Deviation		.848
Variance		.720

The mean assessment of achievement for Q1 was 1.93. On average, the respondents rank the achievement of G0 (vii) closest to 'some achievement', supported by the other measures of central tendency. In the case of G0 (vii), the standard deviation is .848, meaning that 84.8% of responses fall within the range of plus or minus one standard deviation around the mean of 1.93 (within 1 response below .939 very little achieved) or 1 response above the mean at 2.93 (partially achieved). In this case, there is a high degree of agreement among respondents that this Objective has some achievement but less than partial achievement.

Question 1: Achievement of GO (vii)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not achieved at all	3	7.0	7.3	7.3
	Very little achieved	6	14.0	14.6	22.0
	Some achievement	24	55.8	58.5	80.5
	Partially achieved	7	16.3	17.1	97.6
	Mostly achieved	1	2.3	2.4	100.0
	Fully achieved	0	0	0	100.0
	Total	41	95.3	100.0	
Missing		2	4.7		
Total		43	100.0		

As indicated by the frequency table for this question, 55.8% of respondents indicated some achievement; 16% indicated partial achievement and 14% indicated very little achieved. Three of the respondents (75) indicated the Objective was not achieved at all, and none of the respondents indicated that this Objective has been fully achieved. Also notable is that 2 respondents did not provide a response to this question.

Question 1.1 asked respondents whether there was any explanation or justification they would like to provide for Question 1.

Of the 43 respondents to the GO(vii) survey, 30 provided explanation or justification for their ranking on Question 1. Interestingly, of these 30, seven respondents took issue with GO7 itself as being an unattainable objective. As one respondent noted, "complete eradication or prevention ('should be free') is a very high bar and some would argue unattainable." Another respondent provided an alternative way of conceptualizing the goal, asking whether the real measure of success is "that rates of introduction decrease to minimal levels."

In coding the responses, the most common categories included: the varying success across prevention vs. intra-basin spread of AIS; the need for stronger regulatory and other measures for both introduction and spread (17 respondents); coordination measures in place or needed (5 respondents); and individual initiatives that had been particularly promising or required greater attention (3 respondents), in that order.

This is perhaps not surprising, given that responses to Question 1 focus overwhelmingly on some achievement (see Table above).There were many concerns expressed about the adequacy of current programs and measures. As one respondent summarized, "[w]e are making progress, but we have a long way to go if our ultimate objective is freedom from 'introduction and spread'." In their evaluations, in almost all

cases, respondents agreed that programs and measures aimed at preventing AIS introductions had achieved a higher level of success than those aimed at minimizing spread. Many respondents noted that the rate of invasion has 'slowed' or 'been reduced', and for the most part this was attributed to ballast water regulations. Respondents' perspectives on ballast water programming were quite nuanced, however, as existing regulations were considered to be too weak, uneven across borders and not fully implemented, and within-basin ballast controls on domestic shippers inadequate. One respondent declared that, "The three main pathways (canals and waterways, ballast, and the trade of live organisms are being addressed but laws and regulations are still not sufficient to ensure that new AIS do not enter the system via those pathways." Further, considerable concern was expressed about non-ballast introduction pathways, particularly importation and trade (e.g., fresh fish, aquarium). Another respondent noted that, "[w]hile some measures have been taken to reduce risk of invasion with some vectors, those measures are not commensurate with the extraordinarily high damage AIS can inflict."

Half of those commenting on this question expressed concerns about the lack of progress in addressing AIS spread within the Great Lakes basin. Comments were uniformly negative; in the words of one respondent, for example: "There is little achievement towards reducing spread." The exception to the perceived lack of progress in stemming the spread of AIS around the Basin is the decades-long effort to control the sea lamprey; three respondents pointed to this programming as "successful" and having "achieved a lot". By contrast, two respondents noted a clear lack of success in dealing with dreissinids, an area where we "still have a ways to go." One went so far as to declare "we have lost the battle with mussels."

The need for a higher level of more effective coordination among jurisdictions was highlighted by five respondents as a barrier to achieving GO(vii). Two noted the Mutual Aid Agreement among Great Lakes states and Ontario as beneficial in terms of aligning efforts across jurisdictions and others made reference to more informal mechanisms. Relatedly, Annex 6 was singled out by four respondents as being "one with significant progress" and as "part of the ongoing effort" to achieve coordination of existing and planned activities. However, it was also noted that Annex 6 has "not yet been utilized to full potential", and "needs to be better leveraged" (i.e., by two federal governments). Overall, the comments indicated that coordination mechanisms needed to be stronger to meet the challenge of variations in programming and effort across jurisdictions and ecosystems.

Additionally, three respondents expressed concerns about asymmetrical resources and capacity across jurisdictions. As one explained: "Efforts, resources and regulations vary by jurisdiction and prevention efforts are only as strong as the

weakest link." On the other hand, rising levels of awareness, were mentioned by several respondents as a successful outcome of recent program activities.

Contribution of Various Programs and Measures

This question asked respondents to indicate the contribution of various programs and measures to achieving General Objective (vii) using a 7-point scale: no contribution (0); low contribution (0.15), low to medium contribution (0.35), medium contribution (0.50), medium to high contribution (0.65), high contribution (0.85) and complete contribution (1.0). Ten programs and measures were included in this question.

	Program/Measure
2a	Programs/regulations blocking dispersal pathways
2b	Risk assessment programs for new introductions
2c	Programs/regulations for aquaculture, aquarium and bait industries
2d	Programs/regulations for recreational activities
2e	Community education, awareness programs
2f	Border control/inspection programs
2g	Monitoring/surveillance programs
2h	Information-sharing protocols
2i	Rapid response protocols
2j	Programs for preapproving eradication/containment technologies

Question 2: Summary of Contributions of Programs and Measures

		2a	2b	2c	2d	2e	2f	2g	2h	2i	2j
N	Valid	39	39	39	39	39	39	39	39	39	39
	Missing	4	4	4	4	4	4	4	4	4	4
Mean		3.77	3.31	3.10	2.95	3.21	3.56	3.49	3.13	3.26	2.56
Median		4.00	3.00	3.00	3.00	3.00	4.00	3.00	3.00	3.00	2.00
Mode		5	3	1[a]	3[a]	3	4	3	3	3	1
Std. Deviation		1.530	1.472	1.569	1.297	1.281	1.392	1.335	1.218	1.482	1.586
Variance		2.340	2.166	2.463	1.682	1.641	1.937	1.783	1.483	2.196	2.516

a. Multiple modes exist. The smallest value is shown

Almost all of the programs and measures received average ratings in the low-medium range. Programs/regulations blocking dispersal pathways received the highest average contribution assessments, followed by border control and inspection programs, both however with average just above low-medium

	Program/Measure	General Assessment
2a	Programs/regulations blocking dispersal pathways	Mixed- medium/medium-high (28%); high (25.6%); low (21%)
2b	Risk assessment programs for new introductions	Medium (25.6%); medium-high (16%); high (18.6%)
2c	Programs/regulations for aquaculture, aquarium and bait industries	Mixed – no consensus
2d	Programs/regulations for recreational activities	Medium (23%) to medium high (23%); low-medium (21%)
2e	Community education, awareness programs	Medium (37%)
2f	Border control/inspection programs	Medium (23%) to medium high (25%); high (16%)
2g	Monitoring/surveillance programs	Mixed - Medium (30%)
2h	Information-sharing protocols	Mixed – low medium to medium high (72%)
2i	Rapid response protocols	Mixed - medium (25.6%)
2j	Programs for eradication/containment technologies	Low (25.6%) – low medium (18.6%)

As outlined in the table above, most respondents felt that programs/regulations blocking dispersal pathways made a medium-high contribution to achieving the General Objective but there was not consensus on this. Combined, 60.5% felt risk assessment programs for new introductions made a medium-high contribution. Almost half of respondents (46.6%) felt programs and regulations for recreational activities made a medium or medium-high contribution. About one-third felt that community education and awareness programs made a medium contribution, with the remainder of respondents expressing mixed opinions. Monitoring programs and information sharing protocols received mixed assessments. Rapid response protocols and programs for pre-approving eradication/containment technologies received mixed and low contribution assessments respectively.

The detailed data summaries in Appendix VII provide more insight into respondent's opinions about the contributions of various programs and measures to achieving General Objective (vii).

Other Programs and Measures Identified by Respondents

Question 2.1 asked respondents to identify any other programs and measures that contributed to the accomplishment of GO (vii) but were not included in the list provided (2a-2h), and they were asked to use the scale in Q2 to indicate the contribution of that program or measure.

In examining the few open-ended responses here (only nine), it was difficult to distill any categories or themes. The ballast water regulations and sea lamprey control program were singled out as providing "high contribution". One respondent noted that coordination on the U.S. side had been critical (state-state and federal-state). Several comments adopted a forward-looking lens, highlighting programming and measures that may/are likely to play a greater role in the future in addressing some of the current gaps, including risk assessment methodologies, early detection surveys/new monitoring techniques, new management and control technologies, and the Annex 6 committee (as a coordinating body).

Question 2.2 asked respondents if they had any comments on specific programs that are particularly effective or to identify any gaps that exist.

Responses to this question focused mainly on what has been working well, with emphasis placed on prevention programming; as one respondent noted, "[a]ny measures to disrupt pathways of arrival and prevent new introductions are by far the most effective method of dealing with invasive species." Another pointed out that, "[w]hile rapid response is a worthy activity, too often it diminishes or detracts from what should be the primary focus, i.e., prevention."

In this vein, U.S. ballast water regulations and work on treatment technology, along with public awareness programs, received the most mentions in terms of "effective" programs. Other specific programs and measures singled out as effective include: GLRI project funding, sea lamprey control program, border inspections (re. Asian Carp control), monitoring by DFO and MNR, and recreational boat hull cleaning. USACOE's analyses of pathways and points of weakness was characterized as "excellent".

Program gaps were noted to be: the lack of regulations (across various aspects of AIS prevention and control), and regulatory variations across jurisdictions (e.g., which AIS are covered). For example, one respondent noted 'Ontario as yet has no rapid response protocol' and another respondent noted inadequate regulations related to ballast water on Canadian side.

It is worth noting that several respondents deemed it too early to discuss the contribution of most programs and measures, given that implementation had just begun.

Suitability and Accuracy of Indicators

Hill and Eichinger recommended the inclusion of a third question related to indicator appropriateness to serve as a confidence check on the indicators that were selected to represent the Objective. Question 3 asked whether the indicators in the backgrounder and listed in the question accurately demonstrate the condition of GO (vii) today. There were eight main indicators listed and respondents were presented with yes, no, and no opinion response options.

		3a	3b	3c	3d	3e	3f	3g	3h
N	Valid	39	39	39	39	39	39	39	39
	Missing	4	4	4	4	4	4	4	4
Mean		1.51	1.33	1.49	1.77	1.82	1.85	1.74	1.74
Median		1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00
Mode		1	1	1	1	1	1a	1	1
Std. Deviation		.756	.662	.756	.777	.790	.779	.751	.785
Variance		.572	.439	.572	.603	.625	.607	.564	.617

a. Multiple modes exist. The smallest value is shown.

	Indicators	General Assessments
3a	Programs/regulations blocking dispersal pathways	Majority felt this was a good indicator (58% Yes, 18% No)
3b	Number of new introductions	Majority felt this was a good indicator (69.8% Yes, 11.6% No)
3c	Size of existing AIS populations	Majority felt this is a good indicator (60% Yes, 7 No)
3d	Acres (or tributary miles) controlled for invasive species	Mixed – respondents split on whether this is a useful indicators (42.6% Yes to 35.9% No)
3e	Number of monitoring activities conducted	Mixed – split on whether this is a useful indicator
3f	Number of rapid responses or exercises conducted	Mixed – evenly split on value of this indicator
3g	Number of control projects undertaken	Mixed – respondents split on usefulness of this indicator
3h	Number of control technologies and methods field-tested	Mixed – respondents clearly split on the usefulness of this indicator

In assessing the indicators, it is notable that clear majorities felt that those indicators aimed at AIS prevention were good indicators vis-à-vis the achievement of General Objective (vii):

3a) programs/regulations blocking dispersal pathways
3b) number of new introductions
3c) size of existing AIS populations

However, respondents were split on the value of the other indicators, which are measures of government action as opposed to outcome measures.

Other Indicators?

Question 3.1 asked if respondents wanted to provide any other comments on the indicators listed in the survey related to GO (vii).

This question elicited the second highest number of comments (after Q.1.1). The clear theme that emerges from coding the responses to this open-ended question supports the quantitative findings discussed above – namely, there is a generalized concern about the inadequacy of output-focused measures (which measure program *activities*) and the need for more outcome-oriented measures (which measure *impact* and *effectiveness*). Fully half of the comments here questioned the utility of having numbers relating to outputs; as one respondent noted, "[c]ounting the number of activities (monitoring, responses, control projects, etc.) is useful and good information that we should track; however, I don't think that really gets at the objective we are trying to achieve." Another noted that "(g)enerally, 'counting' how often an activity occurs has very little to do with effectiveness."

There were a series of related concerns about the indicators listed in the survey, though these were less generalized. In declining order of frequency, there were doubts about the appropriateness of numeric indicators as providing meaningful information about conditions; cautions about viewing any one indicator in isolation; and acknowledgements that we need to have a full suite of indicators relating to the AIS policy hierarchy (prevention, detection and eradication, control) that can be viewed together.

Question 3.2 asked respondents to list other indicators or data that should be considered to better evaluate and report on the state of GO (vii)?

The following suggestions were made. The reader will note, as per the discussion in the previous section, that most suggestions lean toward an outcome-focused approach to indicator design.

- Rewrite the rapid response indicator to focus on the number of responses that have resulted in successful eradication or containment of an invasive species. (The respondent noted that rapid responses are generally only effective in very specific and limited circumstances, and the number of responses attempted would be irrelevant if the majority are unsuccessful.)
- Include the number of high-risk species still in transport pathways.
- Associate levels of funding provided with results (e.g., funding for sea lamprey control, funding for Asian carp work).
- Incorporate emerging data on species IMPACT into indicators (e.g., http://www.glerl.noaa.gov/ftp/publications/tech_reports/glerl-161/tm-161.pdf) needs to be incorporated.
- Include how much money is appropriated annually at the federal levels to stop AIS introductions and controls.
- Add % of Great Lakes jurisdictions with regulations/legislation that would restrict introduction and spread of AIS. Combine this with some measure of enforcement effort for those regulations/legislation.
- Indicate whether a comprehensive regional monitoring program has been designed and whether it is being properly implemented. Numerical factors can be tied to the plan.
- Add social science indicators such as: science literacy, awareness, knowledge, understanding, skills, behaviours, pathways addressed, or impressions generated.
- Include the number of anthropic vectors present in the studied area.
- Include the distribution, frequency, and species focus of monitoring and surveillance activities, along with the effectiveness of monitoring techniques (e.g., rate of species detection).

Findings Related to Respondents

The remainder of the questions related to data collection on the background of respondents. Of the total number of 41 respondents, 38 answered these questions.

First, there was a broad distribution of respondents in terms of years of experience related to General Objective (vii). While the most common response to Question 4 on years of related experience was more than 25 years experience (27.9% of respondents), there was a more even distribution of responses across all the other categories of experience relative to GO (ii), indicating that there were respondents with little experience, medium levels of experience and high levels of experience.

Question 4: Years of Experience related to GO (vii) and GLWQA

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1-3 years	3	7.0	7.9	7.9
	4-6 years	4	9.3	10.5	18.4
	7-10 years	4	9.3	10.5	28.9
	11-15 years	6	14.0	15.8	44.7
	16-20 years	4	9.3	10.5	55.3
	21-25 years	5	11.6	13.2	68.4
	more than 25 years	12	27.9	31.6	100.0
	Total	38	88.4	100.0	
Missing		5	11.6		
Total		43	100.0		

In terms of the percentage of time that respondents spent engaged in work related to this General Objective of the GLWQA, 39.5% spent between 1 and 20% of their time on work related to this objective. Several respondents spent a lot of their time on work related to this Objective: (7%) spent between 61-80% and six respondents (14%) spent 81-100% of their time on work related to this Objective. This information may be valuable in terms of follow up research, advisory roles, or key informant interviews related to this General Objective.

Question 5: Time Engaged in GO (vii) Work related to GLWQA

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1-20%	17	39.5	45.9	45.9
	21-40%	6	14.0	16.2	62.2
	41-60%	5	11.6	13.5	75.7
	61-80%	3	7.0	8.1	83.8
	81-100%	6	14.0	16.2	100.0
	Total	37	86.0	100.0	
Missing		6	14.0		
Total		43	100.0		

Some 65% of respondents were answering this survey based on a basin-wide perspective. Some were answering from a lake-specific perspective from Lakes Erie and Superior. Six of the respondents chose the response 'other, please specify' and reported: all portions of the basin within Ontario; national, but mainly basin-wide; mainly basin-wide but also Lake Erie and Lake Ontario; including inland waters; and St.Lawrence River.

Question 6: Perspective and Scale

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Basin Wide	28	65.1	73.7	73.7
	Lake Erie	3	7.0	7.9	81.6
	Lake Superior	1	2.3	2.6	84.2
	Other	6	14.0	15.8	100.0
	Total	38	88.4	100.0	
Missing		5	11.6		
Total		43	100.0		

Question 7 asked respondents to identify the field or discipline that best describes their educational/training background. The majority (60.5%) indicated science as their background. The second most common response (9%) was 'other' with 4 respondents listing other fields. Two listed policy/government; and the others listing natural and social science (the same respondent also noting the survey instrument should allow for multiple answers) and environmental studies.

Question 7: Educational Background of Respondents

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Science	26	60.5	70.3	70.3
	Engineering	1	2.3	2.7	73.0
	Law	2	4.7	5.4	78.4
	Business	1	2.3	2.7	81.1
	Economics	1	2.3	2.7	83.8
	Social Science	2	4.7	5.4	89.2
	Other	4	9.3	10.8	100.0
	Total	37	86.0	100.0	
Missing		6	14.0		
Total		43	100.0		

Question 8 asked respondents to identify the type of organization they worked for. The most common response was state/provincial government (18.6%) followed by federal government (16.3%). However, respondents indicated that they work for a wide range of organizations and the findings are thus based on a fairly broad spectrum of experts from a range of organizations, government, non-government, Indigenous, universities/research institutions and private sector organizations. The 3 who indicated 'other' noted: binational commission; international commission; retired but spent career in university, federal government and international/transboundary government, all on Great Lakes work.

Question 8: Organizational Affiliation of Respondents

		Frequency	Percent	Valid Percent	Cum. Percent
Valid	University/Research Institution	4	9.3	10.5	10.5
	Local/Municipal government	1	2.3	2.6	13.2
	State/Provincial government	8	18.6	21.1	34.2
	Federal government	7	16.3	18.4	52.6
	International/Transboundary government	2	4.7	5.3	57.9
	First Nations/Metis/Tribal/ Indigenous	4	9.3	10.5	68.4
	Watershed/regional authority	2	4.7	5.3	73.7
	Industry/private sector	3	7.0	7.9	81.6
	Environmental non-government organization	4	9.3	10.5	92.1
	Other	3	7.0	7.9	100.0
	Total	38	88.4	100.0	
Missing		5	11.6		
Total		43	100.0		

It is important to note here that, in response to Question 9, a high proportion of respondents were affiliated in some way with at least one GLWQA Annex Committee.

Of the 43 respondents to the GO (vii) survey, 23 had some connection to an Annex Committee; 14 are/have been affiliated with one Annex and 8 with two or more Annexes (one respondent indicated they are "peripherally involved with an unspecified Annex"). In addition, of the 23 respondents who have an Annex affiliation, fully 17 indicated that the affiliation was with Annex 6 on Aquatic Invasive Species. This finding may indicate the existence of a "tighter" network around GLWQA programs and measures, a proposition that is given some support from the listing of experts provided by respondents in Question 10, below, or it may be linked to the composition of the GO (vii) survey participant list.

Question 10 asked respondents to identify, in no particular order, the top five experts related to this GLWQA General Objective.

In total, 43 experts were identified (almost twice as many as in the GO (ii) survey), with most respondents identifying 3 to 5 experts (most identified the full five).

When cross-referenced with our participant list for General Objective (vii), all experts receiving 3 or more mentions were on our participant list and received an invitation to complete the survey. Of those experts receiving two mentions, all but one were included in the participant list.

However, of those experts receiving one mention, most were not on the participant list (although one was excluded as an IJC staff member), supporting the need for further exploration of the organizational affiliations of the listed experts in order to understand whether representation of some experts or organizations was an issue. This is important information and the identified experts should be added to the participant list for future research. As with the GO(ii) survey, in some cases, one or two individuals from a given organization were sent the survey, and not more than that, to ensure broad and diverse representation in the participant list.

GLEEM Scores for General Objective (vii)

Similar to General Objective (ii) above, the first two questions are used as the basis of calculating the GLEEM score for the Objective overall and for each of the programs and measures included in Q2.

1. The Actual Performance (AP) value is calculated by dividing the mean value of responses to Question 1 by 6 (number of possible responses) and multiplying by 10 to arrive at a value out of 10.

$$AP = (\bar{x}\ Q1\ \div 6) * \ 10$$

AP = (Mean of Question 1 [1.93/6] converted to value out of 10[16]) = **3.21**

2. The No Regime Counterfactual (NR) value is calculated by averaging the responses to questions 2a-2j for each respondent (y), then determining the mean of all those responses and dividing by 7 (the number of potential responses) to provide a value from 0 (no contribution) – 1 (complete contribution), subtracting that number from 1 (which represents a perfect score on the value), and finally multiplying that value by the AP value.

$$NR = AP * [1 - \left(\frac{\bar{x}\ [\bar{x}Q2(y)]}{7}\right)]$$

NR = [3.21] * 1-(Mean Average of Question 2a-j [3.23]/ 7 =[.46])
NR = 3.21 * 1-.46 [.54]
NR = 3.21 x .54
NR = **1.73**

3. Collective Optimum (CO) = 10

4. The GLEEM score is then calculated using GLEEM = (AP – NR)/(CO-NR)

(*Perceived* Actual performance – No Regime counterfactual) / (Collective Optimum-No regime counterfactual)

[16] the scale recommended by Hill and Eichinger is not clearly outlined in their report; they state it is a 0-10 scale with 0 indicating no level of accomplishment to 10 being complete accomplishment but no values for a survey question were provided as they recommended this be a qualitative question. We adapted this scale to 0-5 and then re-weighted to get value out of 10 in order to apply the GLEEM formula.

GLEEM (3.21 - 1.73) / (10 – 1.73)

1.48/8.27 = .179

GLEEM Score = .18

The effectiveness score or GLEEM score of .18 indicates that the perceived effectiveness of the listed programs and measures combined in achieving General Objective (vii) is low using the scale of: no contribution (0); low contribution (0.15), low to medium contribution (0.35), medium contribution (0.50), medium to high contribution (0.65), high contribution (0.85) and complete contribution (1.0).

GLEEM Scores by Program and Measure for General Objective (vii)

Hill and Eichinger also indicate that effectiveness evaluations can then be completed for programs and measures within General Objective (vii) as well. The same calculation was used to calculate GLEEM scores for each of the 10 programs/measures included in Q2 of the survey.

$$NR = AP * [1 - \left(\frac{\bar{x}\ Q2(y)}{7}\right)]$$

Here the average response for each program and measure is used (where y = 2a through 2j), divided by 7 (number of potential responses) then subtracted from 1 and multiplied by the overall AP value.

	Program/Measure	GLEEM Score
2a	Programs/regulations blocking dispersal pathways	.20
2b	Risk assessment programs for new introductions	.18
2c	Programs/regulations for aquaculture, aquarium and bait industries	.17
2d	Programs/regulations for recreational activities	.17
2e	Community education, awareness programs	.18
2f	Border control/inspection programs	.19
2g	Monitoring/surveillance programs	.19
2h	Information-sharing protocols	.17
2i	Rapid response protocols	.18
2j	Programs for preapproving eradication/containment technologies	.15

GLEEM scores for each of the programs and measures are all close to the overall GLEEM score (.18). However, some programs and measures were assessed as contributing more than others. Programs/regulations blocking dispersal pathways was assessed as contributing the most in terms of achieving this General Objective. Border control and inspection and monitoring/surveillance programs are the other two programs and measures that respondents indicated contribute slightly more than of programs and measures to the achievement of this General Objective. However, it should be noted that all of the programs and measures were assessed as only making low contributions to achieving General Objective (vii).

These findings are consistent with the findings from Q2. Q2 asked respondents directly to rate the contribution of various programs and measures. Therefore, GLEEM scores can be calculated at the program and measure level but do not add a lot of additional data about study participant's assessments of the contributions of these various programs and measures to achievement of GO (vii).

Comparative Analysis of Findings

In order to test the QQEF and GLEEM frameworks we recommended that two General Objectives be selected based on both independent and comparative selection criteria.

Based on the selection criteria and comparative research design, we might have expected to find some of the following:

i. given the longer-standing goal and longer-term focus on GO (ii), that it would garner higher levels of achievement scores from expert assessments OR
ii. given the more recent and more diverse set of programs and measures associated with General Objective (vii), one would expect lower levels of achievement assessments by experts

The results from Q1 and the GLEEM scores support the finding that GO (ii) received higher levels of achievement assessments and respondents felt overall that the combined programs and measures associated with GO (ii) were contributing more to the achievement of GO (ii) [medium] than was the case with GO (vii) [low-medium]. However, it is difficult to determine if this has anything to do with the longer term focus on GO (ii) under the GLWQA.

One other comparative note related to this is that none of the respondents to the GO (ii) survey indicated no achievement at all, whereas 3 respondents to the GO(vii) survey indicated no achievement at all.

iii. given the smaller number and longer-standing use of key indicators related to GO (ii), expert respondents would express a high level of support for the indicators .

This is not supported by the survey findings. Respondents to the GO (vii) survey indicated comparatively stronger support for three of the 10 indicators associated with that Objective. Respondents to the GO (ii) survey were more split in terms of their level of support for the three main indicators in use related to this General Objective. The qualitative data collected were useful here in providing insight into these opinions, given that respondents' comments focused more on the limitations of indicators in use, and requirements for improving the indicators, rather than expressing high levels of support for the indicators.

iv. given the association of an Annex with General Objective (vii), one would expect some consensus on the suite of indicators used to measure progress

This is generally supported by the findings. Far more of the respondents to the General Objective (vii) survey had affiliations with GLWQA Annexes, particularly Annex 6. There is also more consensus that three of the indicators are good indicators related to achievement of GO (vii). However, there is less consensus on which programs and measures contribute the most to achieving the GO and overall, the GLEEM score is lower indicating that respondents to the GO (vii) survey feel that combined, and irrespective of having an Annex, the Objective is only low-medium in terms of achievement.

v. given the longer-standing goal associated with General Objective (ii) one might expect a dedicated group of experts engaged in programs and measures related to this objective, with a range of organizational affiliations and with a higher number of dedicated hours to the achievement of this objective, compared to General Objective (vii)

The findings from Q5 in the surveys do not support this as only16% of respondents to GO (ii) indicated having more than 25 years of experience on work related to this Objective, compared to 27.9% of respondents to the GO (vii). Also 54% of GO (ii) respondents indicated they spent less than 20% of their time engaged in work related to GO (ii), compared to almost one-third (32.6%) of respondents to the GO (vii) survey indicated they spent more than 40% of their work time on work related to GO (vii).

Overall, more GO(vii) experts respondents spent the majority of their time on work related to this Objective. This may or may not have implications for achievement but raises some interesting comparative questions such as: do those General Objectives with concerted expertise and effort yield more effective outcomes/more achievement? This analysis indicates that this is not necessarily the case, but longitudinal data would be required to determine this over time.

vi. given the association of an Annex with General Objective (vii) one would expect an identifiable and dedicated group of experts engaged in programs and measures related to this objective.

Responses to Q10 in the survey generated a longer list of experts associated with GO (vii) and a higher degree of consensus on who the key experts were, as measured by number of mentions.

Other Comparative Observations

The above discussion yields only some of the dimensions of comparison that we expected might be interesting results when testing the QQEF and GLEEM score framework. Comparative analysis yields several additional observations and questions.

Responses for GO (ii) were based on a mix of basin-wide perspectives and the perspective of all of the lakes; GO (vii) respondents were overwhelmingly answering from a basin-wide perspective. Does expertise with a basin-wide perspective vs. lake or local perspective make any difference in terms of assessments of achievement?

In terms of indicators, does more common education/training background of experts have implications for how they assess indicators? Over 60% of GO (vii) respondents indicated science was their background, whereas the background of GO(ii) respondents was more diverse, with only 32% of GO (ii) respondents listing science and the second highest listing 'public health' in the 'other' response category. From the qualitative comments, there is some indication that GO (vii) respondents want indicators to focus more on outcome measures related to the state of environment, while GO(ii) respondents expressed a more mixed desire for utilizing a mix public health indicators and water quality indicators. In both cases, there is clearly a sense that we do not yet have the right (suite of) indicators in place to provide an accurate measure of effectiveness in achieving the General Objective in question.

Another question is whether organizational affiliation has an impact on how the General Objectives, programs/measures and indicators are assessed. Respondents for both surveys worked for a wide range of government and NGO organizations, with more GO (ii) respondents indicating they worked for local/municipal government and more GO (vii) respondents indicating they worked for either state/provincial or the federal governments in Canada and the US. Notable also is that no respondents to the GO (ii) survey indicated industry/private sector or watershed/regional authority as their organizational affiliation but 3 respondents were from the private sector in the GO(vii) survey. This may highlight the key role that recreational and commercial shipping interests play in terms of the prevention and detection of aquatic invasives. Certainly, the various councils and working groups established (generally under the terms of government mandates) have emphasized the involvement of private sector and recreational stakeholders.

The finding that there are lower levels of achievement for GO (vii) when compared to GO (ii) is somewhat expected and there are many factors that can possibly explain lower levels of achievement in GO (vii) compared to GO (ii). Using the QQEF and

GLEEM score framework is valuable but it only provides limited baseline data at one particular point in time. However, the open-ended responses provide additional, useful insights into the GLEEM scores, including why respondents ranked the way they did, where the indicators fall short and how these gaps might be addressed in the future.

This comparison is based on a research design that used the same survey instrument but distributed to experts specific to each General Objective. A research design that used a common set of experts to answer similar questions for all 9 General Objectives would undoubtedly yield different results.

Analysis of Methodology

This section revisits the sections at the beginning of this report, offering observations, comments and reflections on the methodology used to test the QQEF and GLEEM score framework. It begins with some general observations.

Hill and Eichinger recommended that the Quantitative-Qualitative Effectiveness Framework (QQEF) allowed for both descriptive and causal analysis, recognizing that causal analysis is based on inference and judgment from experts and diverse but knowledgeable stakeholders, and that there are some difficulties inherent in evaluating the effectiveness of environmental programs. The QQEF is an outcome assessment evaluation designed to provide an overall judgment on whether and how well the program(s) have met the stated goal(s) or objective (s). They also recognized that meaningful and effective environmental program assessment depends on the degree to which a causal relationship between a program(s) and achievement of the stated outcome can be determined, and this can be challenging. In the case of the application of their suggested framework, they noted the success of the approach depends upon sound data and core indicators, i.e., they must be reliable, longitudinal and linked directly to the Objectives of the GLWQA (Hill and Eichinger 2013, 15).

The model they recommended blended qualitative and quantitative research methods to provide an evaluation of overall effectiveness as well as effectiveness at the program level. The QQEF was recommended as it would make use of structured, systematic scoring to detect and determine the causal link between program activity and observations on the outcome (Hill and Eichinger 2013, 20).

The QQEF clearly produced results that allowed for analysis of both the level of achievement of the General Objectives and of the connection between the selected General Objectives, programs and measures and indicators included. The

inclusion of qualitative components allowed for contextual analysis of the quantitative results, collection of additional information on the contributions of various programs and measures and allowed for the collection of valuable information on the indicators included, the strengths and limitations of those indicators.

Although not recommended or required in the QQEF, the inclusion of questions about the respondents themselves added valuable quantitative and qualitative data that adds to the data set and presents the potential for additional analysis.

However, there are some important methodological observations and reflections worth outlining that have implications for assessing the QQEF and GLEEM score approach.

Selection of Test Cases

The primary purpose of this study was to test the framework in terms of its potential applicability across the 9 General Objectives in the GLWQA. The first methodological step was selection of the test cases. As outlined in Appendix II, the General Objectives vary on some important dimensions. By selecting two of the General Objectives, we were able to test the applicability of the QQEF across more than one of the Objectives.

The selection process itself was valuable as it resulted in theory development and comparative hypotheses related to the QQEF and beyond. The application and standardization across these cases was fairly straightforward, and generated some useful results for both testing the QQEF and doing some additional comparative analysis.

Overall, the QQEF was applicable across the two cases, generated some interesting and valuable findings, and indicates that the QQEF could be applied across the other General Objectives in the GLWQA and GLEEM scores generated for each of the 9 General Objectives (more on GLEEM scores below).

If the QQEF and GLEEM scores were the only product from the qualitative and quantitative data collected, the survey could be a lot shorter (just including the indicator backgrounder, the three questions required for the GLEEM score and the qualitative questions to solicit additional data on the rationale for achievement, programs/measures and indicator assessments).

The identification and assignment of indicators to the two General Objectives was more challenging. Using all available secondary sources such as IJC, SOLEC, GLRI,

and other government sources, combined with academic sources, a suite of indicators was identified. In consultation with IJC staff, the key indicators were then the focus of the backgrounder recommended as by Hill and Eichinger.

A challenge was to condense this background information on indicators related to each General Objective into a 1-2 page backgrounder for each survey. While one or two respondents commented on the backgrounders, the vast majority did not, which may indicate that they found the backgrounders accurate and useful or perhaps that they didn't read them (although prompted to indicate they had read them as part of the online survey design).

Based on the process of developing the backgrounders and developing the indicator lists - which ranged from 3 in the case of GO (ii) to 8 in the case of General Objective (vii) - this is a feasible and useful part of the QQEF. In order to test the value of these backgrounders definitively, future surveys could ask respondents a more specific question about the usefulness of inclusion of this baseline information on indicators. An alternative test might have been to include the indicator backgrounder in one case and not the other, but it would have been very difficult to attribute differences in responses to Q3 to the inclusion or exclusion of this information unless a specific question on this followed Q3.

The research associated with developing the indicator backgrounder was valuable in itself and this application indicates it could be done for all of the General Objectives, those with fewer indicators (like General Objective (ii), and those with more indicators like General Objective (vii).

The other valuable outcome of the QQEF was the generation of the participant lists as the sources of the evaluation data.

Evaluation by Participants with 'insider knowledge' and 'outsider perspectives'

Hill and Eichinger also recommended the source of data for the QQEF and GLEEM scores be generated from individuals and organizations with expertise related to each General Objective – experts with both 'insider' knowledge and 'outsider' perspectives. Essentially, a purposive elite sample was recommended.

The approach used here defined those with 'insider knowledge' as those involved with implementation of the GLWQA, Annexes, and implementation of related programs and measures and 'outsider perspectives' as a wide range of other experts

from government, non-government and the private sector with knowledge and interest related to the General Objective, including academic experts.

As outlined in the methodology section earlier, the GLPRN database of policy actors was a starting point. Then Annex membership lists, other IJC lists and searches of relevant government and non-government websites and documents were used, such as the Great Lakes Beaches Conference program and the Aquatic Nuisance Species Task Force, to generate the final list of participants who would receive the surveys. The process, although somewhat time consuming, was feasible for both of the test cases and would be feasible for each of the 9 General Objectives. In addition, by including a specific question on this, the survey itself yielded valuable data related to experts who can assess overall achievement, programs and measures and indicators in future exercises.

The findings from both surveys indicate that the respondents came from a variety of organizations, had a range of educational/training backgrounds and a range of years of experience related to the General Objective they were assessing. Most participants reported basing their responses based on a basin-wide perspective but also a lake specific or other scale, and this diversity is helpful in terms of understanding the degree to which views on effectiveness are generalized across scales.

The decision to exclude EPA Region V officials is an important point to note. However, the survey groups were broadly representative as indicated by the balanced number of Canadian and American respondents and as indicted by findings from Questions 4 through 10. The removal of these potential participants from the participant lists did not likely affect the findings in any significant way. At the same time, having these participants included would be beneficial in the future as they hold key knowledge, expertise and perspectives related to assessing and evaluating achievements related to the GLWQA.

The lists also did a good job of capturing most of the key experts identified in Q10. The responses to Question 10 are a valuable check on the lists but also indicate that some key experts were missing from the lists. One approach in the future might be to have the lists reviewed by a few identified key experts in addition to IJC staff. A further option is having an online registry whereby a call is made to interested experts to register and agree to participate in evaluation research. There could be an option for them to agree to anonymity or to attribution of their responses. This would also address one of the common concerns in the research ethics process, namely how the responses will be used in results reporting. The registry list might only be accessible by IJC staff and related researchers to avoid survey fatigue and other ethical issues,

or open to the public to address the 'who are the experts' question that is inevitably asked in expert survey research.

Whatever process is used to generate the list of assessors and potential sources of evaluation data, the focus on experts and stakeholders with some knowledge related to the various General Objectives seems to be a valuable approach to collecting evaluation data for reporting purposes. This is also in keeping with the expert and scientific sources of data compiled through SOLEC that IJC uses in progress reporting. Targeting experts with 'insider knowledge' and 'outsider perspectives' to assess the achievement of General Objectives (as recommended by Hill and Eichinger) also gives some legitimacy to the evaluation of findings. Further, having standing lists of experts for evaluation purposes can also be valuable for longitudinal research using the QQEF and GLEEM scores. However, Hill and Eichinger do not define 'outsider perspectives' beyond the example of the participant list they provide and, as noted by some respondents in the open-ended comments, this approach does not include an important focus on public assessment and evaluation.

The focus of the QQEF and GLEEM scores on experts with insider knowledge and those with outsider perspectives (or stakeholders) does not address calls by some academics and non-government organizations for more evaluation frameworks and related research that focus on public awareness and opinion related to achievements of Objectives in the GLWQA. It also does not include a focus on elected leaders and representatives and evaluation frameworks that compare elite and public opinion.

Finally, determining the appropriate number of experts, whether in terms of a survey response rate, or in terms of key informant interviews is always a challenge in evaluation research as the results are not generalizable to a large population.

On-line Survey

Although not explicitly recommended by Hill and Eichinger to test the QQEF and GLEEM score framework, an on-line survey has many benefits. First, survey design, editing and pre-testing are easier (and in this case easily duplicated for two surveys or subsequent surveys on other General Objectives). Second, it is convenient for the experts on the participant list. All participants for purposes of this evaluation research have publicly available email addresses, and many are familiar with online survey formats. Third, online surveys are very easy to administer, track responses and non-responses, and generate reminders. Fourth, the data are easy to aggregate and export for analysis in a statistical software packages like SPSS.

There are also some limitations. First, online surveys may end up in spam folders, as some are screened by spam and junk mail systems. Second, surveys may get buried in email inboxes. Third, given the number of online surveys, there is a risk of survey fatigue contributing to non-responses. Fourth, some may hesitate to respond in digital format as their responses are more 'trackable' by researchers. Fifth, some online survey questions require responses as a condition of proceeding, which may frustrate some respondents causing them to stop completing the survey.

i) Response Rates

The decision to use an on-line survey to collect data from the expert participants was overall a positive decision. Response rates are always an issue with survey research. Although Hill and Eichinger do not indicate ideal response rates for the QQEF and GLEEM scores, we expected the potential participants to be highly engaged and motivated to take part in an evaluation study, and thus anticipated response rates in the 50%+ range. The response rates of 32% for GO (ii) and 42% for GO (vii), while not at the level expected, are higher than typical online surveys[17]. In terms of number of respondents, for GO (ii) most findings are based on data from 33 experts, and in the case of GO (vii), 41 experts. So, in the context of online survey research, the response is satisfactory and within the norm to generate useful findings. Given the purposive nature of the sample of participants and given that the framework does not attempt to generate findings that are generalizable to the total (unknown) population of all experts with insider knowledge and outsiders with valued perspectives, the response rates are satisfactory and sufficient to deploy the QQEF framework and generate GLEEM scores. The findings however need to be interpreted with these limitations in mind.

This raises questions about the reasons for the response rates obtained and how valuable the findings are if the QQEF and GLEEM scores are used for IJC reporting purposes. We might speculate on whether the front-end information (ethics consent pages and key background and baseline information) was onerous and negatively affected response rates. Asking this question of non-respondents is one way to get an answer to this, and to identify other reasons some experts chose not to respond. If response rates could be higher in subsequent applications of the model, this would further support the value of the QQEF.

[17] See Penwarden 2014 in bibliography
http://fluidsurveys.com/university/response-rate-statistics-online-surveys-aiming/

There are also a number of ways that response rates could be increased. The literature on survey research is full of options from prior communication with key experts, various incentives, reducing the time required to complete the survey (by reducing the background material or number of questions), or making it one part of a more iterative research method such as a Delphi Method.

Also, the response rates attained for this survey beg the question of whether key informant interviews might be a better alternative. As Hill and Eichinger note, interviews have three major limitations related to the QQEF and GLEEM score, they are time intensive, resource intensive and place more emphasis on qualitative data in the QQEF. Based on the testing here we agree that the QQEF presents a good balance of quantitative and qualitative evaluation questions.

Another option is designing a survey to collect the data for Questions1-3 and then supplementing with a smaller, representative sample of key informant interviews. Q10 in particular results in the generation of a short list of experts that could be targeted for key informant interviews, future/cyclical surveys, or as reviewers of participant lists should subsequent/longitudinal research be conducted.

Overall, although the survey response rate is lower than anticipated, both surveys yielded some valuable evaluation data and findings and generate sufficient material to generate the GLEEM scores.

ii) Survey Questions

The adaptation of Q2, from a qualitative question as recommended by Hill and Echinger, to a quantitative question supplemented by a qualitative question was a useful way to directly ask experts their opinions on how various programs and measures contributed to the achievement of the General Objective rather than using qualitative coding to determine this. This also generated some useful quantitative results from Question 2. The qualitative supplementary question also generated some useful data. Having quantitative data for Question 2 made calculation of the GLEEM score easier as Question 2 is the basis of the No-regime counterfactual (estimating the state of the Objective absent the intervention of the lists programs and measures). However, in terms of generating and interpreting the GLEEM score at the program level, this does not appear to give us any additional analytical value. It was found that analyzing the statistics from responses to Question 2 produced more useful findings on the perceived contributions of various program and measures to achieving the General Objective.

Hill and Eichinger also suggested Question 3 be included as a check on the assumption that the indicators we are observing are suitable and appropriate related to the Objective being assessed. In general, this was the case, but as the responses to the GO (ii) indicate, there is not a clear consensus on the indicators. Indeed, some qualitative comments indicated this should be a priority. In the case of GO (vii) there was majority consensus on three of the indicators, but mixed assessments of the other five indicators. Qualitative comments received for both General Objectives indicate various opinions about the indicators themselves and their appropriateness as they currently exist related to the achievement of the General Objective. However, many suggestions were made by experts on how to improve the indicators in both cases.

Inclusion of Questions 4-10 were not required to test the QQEF or generate the GLEEM scores. There were an unlimited number of additional questions we could have asked in the survey including socio-demographic questions related to respondents, knowledge/perspective/opinion questions, etc. In consultation with IJC staff, we had to make some decisions about which ones to include. These questions generated some valuable data that helped contextualize the framework. There are other questions that could be asked such as nationality if the IJC is interested in analysis of variation by Canadian and American respondents.

Finally, one technical finding we discovered while aggregating the data was that *Fluid Surveys* automatically assigns a value of 1 even if the response scale starts at 0. For example, in Question 1 we used a scale of 0-5 for achievement. Fluid Surveys exports the data using values 1-6 for analysis in SPPS. This created a problem when it came to generating summary statistics. It might be beneficial in future surveys to use a scale of 1-6 to avoid having to manually assign the values for statistical analysis.

Overall, using an online survey allowed for the collection of two important data sets:

1) data related to testing the QQEF and GLEEM framework

2) data on respondents and perspectives that are useful for enriching the analysis and can be used for subsequent/longitudinal research

iii) Qualitative Questions

As noted by Hill and Eichinger, the use of qualitative data generally enhances evaluation research. We believe that the inclusion of qualitative questions in both surveys clearly added value to the analysis and interpretation of findings. However, some limitations need to be noted.

Coding open-ended survey data involves interpretation. This is the case whether it is one person coding the text or a small team. Having a small team code the same text and distill the main categories and themes is recommended as a form of inter-coder reliability. However, as the workshop exercise clearly demonstrated for coding Question 1.1 of both surveys, there are varying levels of difficulty associated with identifying categories and generating summary findings. Some of this was a function of the level of expertise of the coders. Those with a background and some knowledge and expertise related to the topic seemed to be able to generate fewer and more precise categories. For GO (vii) for example, given the workshop coders had no background and the responses needed to take into account a broader range of programs and measures, coding was a challenge. This is an important observation. However, ultimately groups of 4-6 came up with similar long lists of categories and general themes. It might be sufficient if small groups of 2-3 with some background related to the Great Lakes engaged in the coding. However, this is also very labour intensive, even with the very small number of responses that had to be coded in the two cases.

Great Lakes Environmental Effectiveness (GLEEM) Scores

The QQEF using a survey did allow for the generation of GLEEM scores overall and at the program level. The overall GLEEM scores did allow for an evaluation of how much achievement could be attributed to the programs and measures included in the survey. Overall, this does also allow for some assessment of perceived effectiveness and comparability across the two selected cases. It provides an evaluation of the *collective* contribution of all the listed programs and measures to the achievement of the General Objective and well as some assessment of contributions of various programs and measures.

The GLEEM framework is based on the Oslo-Potsdam approach to effectiveness measurement which is built on three components that are regularly used by scholars of regime effectiveness and environmental evaluation studies: i) the observed level of problem solving, ii) the no-regime/no-program counterfactual, and iii) some concept or model assumption of what would constitute full problem solving.

The model firstly, defines a point against which actual performance can be compared, and, secondly, provides a common metric that can be applied across a wide range of cases. The merit of the Oslo-Potsdam yardstick is that it explicates in a compact way the types of operation that are necessary if one defines, as most regime and evaluation scholars do, effectiveness in terms of actual performance, a counterfactual no-regime situation and an optimal situation. The GLEEM model has a built-in assumption that programs and measures make some difference (in terms of

improvement on the no-regime counterfactual), and that there is some abstract collective optimum in achieving all that can be achieved (distance to optimum).

The actual performance of a regime can be compared against two points of reference. One is the hypothetical state of affairs that would have come about had the regime not existed. This is clearly the standard we have in mind when arguing that 'regimes matter' or the GLWQA matters, and that programs and measures broadly associated with the GLWQA make a difference in terms of achieving the given Objectives. For comparative research, such a standardised notion of 'relative effectiveness' is particularly attractive in that it helps solve the common metric problem. But any attempt at measuring regime effectiveness involves causal inference, requiring that we separate changes that can be attributed to the existence and operation of the regime itself from those that have been brought about by other factors. This is by no means a trivial exercise (Underdal 2002, Underdal and Young 2004).

Hill and Eichinger adopted the conventional approach defining success in terms of effectiveness. In a common sense understanding, a policy regime - i.e. a set of rules and norms designed to govern a particular system of activities - is effective to the extent that it performs a particular function or solves the problem it was established to solve. The focus is on the extent of goal attainment and the extent to which such goal attainment is caused by the regime of programs and measures included in the model. The fully solved situation is given a score of 1 and the fully unsolved given a score of zero.

Like all survey data, the data generates a GLEEM score at one point in time. It is a snapshot of the perceived effectiveness in relation to the programs and measures identified and included in the survey. This is valuable when interpreted using the scale from 0 (no achievement) to (1.0) full achievement. It would also be more valuable if it could be measured over time. However, like all longitudinal research, there are limitations with this as well.

Mitchell (2002, 71-3) notes that there is potential for measurement error, in that regimes differ with respect to the difficulty of the problem they seek to address. It is difficult to control for this and this can lead analysts to systematically overestimate the effectiveness of regimes that tackle relatively benign problems (Underdal 2002).

Applying the GLEEM scores at the program/measure level basically confirmed the findings from Q2 in the survey but with the benefit of factoring in the no-regime counterfactual (NR). The NR works by selecting a lower bound which is defined as the no-regime or no-program counterfactual, which posits what the measurement for a

given indicator would be in the absence of that program (in this case program or measure0. The challenge at the program level is the NR is estimated on an average, assuming that no other programs or measures are contributing to the achievement of the General Objective.

Interpreting and analyzing the descriptive statistics for Question 2, provides a more useful basis to evaluate the assessed contribution of a given program or measure to achieving the General Objective. This is another rationale for having this as a quantitative question followed by a qualitative question in the survey. However, there is also the risk that survey respondents may quickly make assessments to move through the survey.

Interviews may generate more thoughtful identification and assessment of a more limited number of programs and measures but the qualitative results are more challenging to code and aggregate to generate the GLEEM scores. Using interviews instead of an online survey to generate data for the GLEEM score however would also likely reduce the number of participants and perspectives and this would be an important tradeoff to consider.

RECOMMENDATIONS

The goals of the project were to:

1. Test the proposed framework through its application to an assessment question related to Great Lakes Water Quality Agreement objectives, and
2. Provide advice to the Commission on the framework's suitability for supporting the Commission's Great Lakes Triennial Assessment of Progress Report.

1. The QQEF clearly produced results that allowed for both analysis of the level of achievement of the General Objectives and allowed for the connection between the selected General Objectives, programs and measures and indicators included. Based on the test across the two selected General Objectives, the proposed framework could be applied and used to evaluate other General Objectives in the GLWQA. **We recommend the IJC consider using the framework as a starting point for baseline and longitudinal data collecting and reporting purposes related to all 9 of the General Objectives of the GLWQA.**

2. The value of the GLEEM score and related methodology is essentially comparative across General Objectives and time as the scores are based on survey data that is only one snapshot in time**.** On its own the GLEEM score provides a dashboard metric that can be interpreted using the adopted scale at both the overall and program and measures level. The GLEEM score could also be very useful in reporting if the IJC wishes to report on progress comparatively, according to various expert and stakeholder perspectives, and across time. For example, by testing the GLEEM score across these two cases, it is evident that experts and stakeholders feel the progress and measures have contributed to higher levels of achievement related to General Objective (ii) than to General Objective (vii). This may be useful if the IJC wishes to highlight comparative achievement across Objectives, but also to prioritize additional efforts. If a similar method were used every 2-3 years, the IJC could report on GLEEM score progress over time**. We recommend the GLEEM scores only be used in comparative and longitudinal reporting.**

3. Based on the value of generating expert insider and outsider participant lists to generate evaluation data for the QQEF and GLEEM score frameworks, **we recommend that the IJC develop and maintain a standing roster of expert and stakeholder participant lists related to each of the General Objectives, thereby making assessments of achievements over time a possibility. We also recommend this be considered for Specific Objectives and Annexes in the future.**

4. Response rates using an on-line survey will always be a challenge related to applying the QQEF and generating GLEEM scores. However, given the number of General Objectives and other limitations of evaluation methodologies using interviews and only collecting qualitative data, **we recommend the use of on-line surveys to collect data for the QQEF and GLEEM frameworks** to generate baseline evaluation data and findings that could then be supplemented by a more limited number of key informant interviews, perhaps with those experts identified through Question 10. Future use of the QQEF should also explore all options for improving response rates, including prior communication with potential participants on the importance of their assessments.

5. The backgrounders recommended by Hill and Eichinger required a significant research component; however, this component of the QQEF and GLEEM framework was valuable in itself and related to subsequent design of the survey instrument and Question 2. This application indicates it could be done for all of the General Objectives, including those with fewer indicators (like General Objective (ii), and those with more indicators (like General Objective (vii). **We recommend the 1-2 page backgrounders be included in the survey as contextual and baseline information for respondents, and that these backgrounders be reviewed by IJC Staff and 1-2 key experts related to each of the General Objectives prior to finalizing the survey.** While including these may have some negative implications for response times and response rates, we feel the benefits of baseline information related to the evaluation and indicators outweigh the possible negative implications of including these at the front end of the survey.

6. Given the value-added from the addition of the open-ended questions related to indicators, **we recommend, if the QQEF and GLEEM score framework is applied, that the open-ended questions related to the suitability of the indicators be included,** although the number of those questions could be reduced as there was some redundancy in the open-ended comments related to Questions 1.1 through 3.2

7. Although not recommended or required in the QQEF, **we recommend that Questions 4-10 related to the background and perspective of respondents be included in applications of the QQEF** as they provide valuable findings and important context for interpreting the evaluation data and the inclusion of questions about the respondents themselves added valuable quantitative and qualitative data that also presents the potential for additional analysis. The questions included could be increased or decreased depending on the amount of analysis the IJC would like related to the respondents. However, the high non-response rate to this set of questions is an issue that should be discussed in future applications of the QQEF.

8. Given the usefulness of the open-ended responses in terms of understanding the strengths and deficiencies associated with indicators in our two surveys as well as generating ideas for alternative indicators, **we recommend that open-ended, qualitative questions be included in future surveys and the online survey be supplemented by in-person interviews if improving the indicators and programs associated with various General Objectives is a priority**. The focus here would be on targeting a smaller number of representative interviewees, in order to further probe the survey results and provide specific recommendations for the development of new or improved indicators.

CONCLUSIONS

Evaluating progress and achievement related to complex policy and environmental governance regimes is a challenge facing many jurisdictions and international organizations. Increasing emphasis has been placed on finding useful and broadly applicable instruments for measuring success in achieving objectives. This task is exceedingly complex in a transboundary context where implementation efforts are distributed across boundaries, across scales and across sectors.

Overall, we find that the QQEF can be successfully applied to the two cases chosen (that is, General Objectives (ii) and (vii) of the GLWQA) and, further, we would argue that this framework could be applied across all other General Objectives in the GLWQA. However, GLEEM scores are dashboard metrics, and have limitations when reported for one Objective and at one point in time. If reported comparatively and longitudinally, they are more valuable in understanding the relative success in achieving General Objectives. In the best-case scenario, the results of this quantitative approach can be supplemented and deepened by qualitative evidence and analysis.

Our application of the GLEEM framework generated some interesting and valuable findings. While survey respondents seem to agree that some progress has been made in achieving the two General Objectives under study here, it is clear that there is considerable room for improvement. Moreover, while a consensus seems to exist among survey respondents that the indicators we currently have in place provide some guidance for judging implementation efforts and progress, they also agree that considerable work remains to be done in terms of defining and operationalizing indicators that can effectively show program impact and effectiveness. There are real concerns about the focus of indicators, the data currently being collected and used, and the reporting related to these indicators. The concerns expressed,

especially in the GO(vii) survey responses, about indicators that focus on program activities rather than environmental outcomes, cast some doubt on potential support for Program Effectiveness Indicators in the expert community.

The fact that so few respondents indicated full achievement in response to Question 1 reflects recognition that both achievement and measuring achievement are challenging. It may also indicate a deeper concern about the governance mechanisms in place to coordinate action around the Great Lakes Basin, to overcome differences in approach, and to include and engage the relevant policy actors. Application of the QQEF across the General Objectives in the GLWQA offers some potential to shed additional light on progress and gaps related to achieving the goals of the Agreement.

BIBLIOGRAPHY

Breitmeier, Helmut, Arild Underdal and Oran R. Young 2011. "The Effectiveness of International Environmental Regimes: Comparing and Contrasting Findings from Quantitative Research", *International Studies Review*, 13(4), 579-605.

Dombrowsky, I. 2008. "Institutional Design and Regime Effectiveness in Transboundary River Management: the Elbe Water Quality Regime", *Hydrology and Earth System Sciences*, 223-38.

Dupre, S. 2013. *An Inventory of Nutrient Management Efforts in the Great Lakes*, prepared for the International Joint Commission's Lake Erie Ecosystem Priority Management Team.

Hill, James P. and Daniel Eichinger 2013. "A framework for assessing the effectiveness of programs and other measures developed to address the objectives of the Great Lakes Water Quality Agreement" as report submitted to the International Joint Commission, March, 2013..

International Joint Commission 2009. *Work Group Report, Beaches and Recreational Water Quality*.

International Joint Commission 2011. *Assessment of Progress Made Toward Restoring and Maintaining Great Lakes Water Quality Since 1987*, Draft Report, October 2011.

International Joint Commission 2011b. *2009–2011 Priority Cycle Report on Microbiological Quality of Great Lakes Beaches and Recreational Waters,*

International Joint Commission 2013. *16th Biennial Report on Great Lakes Water Quality and Accompanying Technical Reports*, April 2013.

International Joint Commission 2014, *Program Effectiveness Indicators Workshop Report*.

Miles, E., A.Underdal, S.Andresen, J.Wettestad, J.B.Skjærseth, E. M. Carlin 2002. *Environmental Regime Effectiveness: Confronting Theory with Evidence*, (Cambridge, MA: MIT Press).

Niemeijer, David 2002. 'Developing Indicators for Environmental Policy: Data driven and Theory driven Approaches examined by example", *Environmental Science and Policy*. 91-103.

Penwarden, Rick 2014. *Response Rate Statistics for Online Surveys -What Numbers Should You be Aiming For?*, http://fluidsurveys.com/university/response-rate-statistics-online-surveys-aiming/

Underdal, A. 1992. "The Concept of Regime Effectiveness", *Cooperation and Conflict*, 27(3).

Underdal, A. 2002. *Explaining Regime Effectiveness*, http://www.cas.uio.no/Publications/Jubilee/Explaining_regime_effectiveness.pdf

Underdal, A. and O. Young eds. 2004. *Regime Consequences: Methodological Challenges and Research Strategies*, (Springer Publishers: New York).

Young, Oran, 2011. *Effectiveness of International Environmental Regimes: Existing Knowledge, Cutting-edge Themes and Research Strategies*, National Academy of Sciences, 108(50).