Driving Innovation in Crisis Management for ***E***uropean ***R***esilience

D23.21 - Performance and Effectiveness Metrics in Crisis Management Experiments

Document Identification	
Due Date	31/01/2015
Submission Date	01/02/2017
Status	Final
Version	3.0

Related SP / WP	SP2 / WP23	Document Reference	D23.21
Related Deliverable(s)	D21.21, D23.31	Dissemination Level	PU
Lead Participant	ATOS	Lead Author	Raul Sevilla
Contributors	ECORYS, FHG-INT, JRC	Reviewers	Denis Havlik (AIT)
			Marcel Van Berlo (TNO)

Keywords:
DRIVER methodology, Capacity building, performance, metrics, iterative development

This document is issued within the frame and for the purpose of the *DRIVER* project. This project has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 607798

Document Information

List of Contributors	
Name	Partner
Raúl Sevilla	ATOS
Jaime Martín	ATOS
Diego Chanto	ATOS
Dick Mans	ECORYS
Isabelle Frech-Linde	FHG-INT
Merle Missoweit	FHG-INT
Chiara Fonio	JRC
Dagi Geister	DLR
Adam Widera	WWU

Document History			
Version	Date	Change editors	Changes
0.1	2014-12-10	ATOS	Initial version / table of contents
0.2	2015-01-21	ATOS, ECORYS, FHG-INT	Added contributions
0.3	2015-02-19	ATOS	Full draft for external review
1.0	2015-02-27	ATOS	Final version submitted
1.1	2016-01-14	ATOS	Executive Summary, Introduction and Conclusions are revisited. Section 2 is added, together with a scientific background. References to original sources are improved in the whole document. Liaison with other DRIVER activities is enhanced. Examples are added. Section 5 is transformed as an example, to serve as a bridge with D23.41 in forthcoming deliverables
1.2	2016-02-01	ATOS	Scope and way ahead clarified
2.0	2016-02-29	ATOS	Quality check performed on this document

2.1	2016-09-27	ATOS	Proposed new structure
2.2	2016-11-10	ATOS	Reviewed and consolidated structure
2.3	2016-11-24	ATOS, ECORYS, FHG-INT	Added contributions to section 3
2.4	2016-11-30	AIT, ATOS	Reviewing the available structure and text. Improving the readability. Improving the management summary, sections 1 and 2.
2.5	2017-01-30	DLR, WWU	Partial rework of the document following feedback from ATOS, FOI, FHG and JRC
3.0	2017-02-01	Atos	Final version for submission

Table of Contents

Project Description 7

Executive Summary 8

1 Introduction 9

1.1 Foundations of DRIVER Methodology 9

1.2 Document Overview 10

2 Measuring Performance in Crisis Management Experiments 11

2.1 Performance Measurement Scope in DRIVER Experiments 11

2.2 Iterative Development of Experiments 13

2.3 Objectives and Aimed Capabilities 14

2.4 Identification of Performance Drivers 17

2.5 Parameters and Scoring 21

2.5.1 The SMART Rule 22

2.5.2 Metrics and Scoring 23

2.5.3 Scoring and Decision Making 24

3 Guidelines and Recommendations 26

3.1 General Guidelines to Establish Metrics 26

3.2 Recommendations and Common Problems 28

3.2.1 Quantitative and Qualitative Methods for Data Collection 28

3.2.2 Creation of Questionnaires 31

3.2.3 Reliability and Validity of Groups Selection 32

3.2.4 Designing and Evaluating Effective Surveys 37

4 Conclusion 40

References 41

List of Tables

Table 1: Summary of anticipated problems and according recommendations in context of performance measurement in DRIVER Experiments 39

List of Figures

Figure 1: Performance Measurement Dimensions in DRIVER Experiments 12
Figure 2: Functional classification of CM tasks. 15
Figure 3: Number of publications per year 17
Figure 4: Subject areas of search results 18
Figure 5: Subject areas of refined search results 19

List of Acronyms

Abbreviation / acronym	Description
ACRIMAS	Aftermath Crisis Management System-of-systems Demonstration
CD&E	Concept Development and Experimentation
CG	Control Group
CM	Crisis Management
DRIVER	DRIVing innovation in crisis management for European Resilience
ICT	Information and Communications Technology
KPI	Key Performance Indicator
NG	New-capability Group
PF	Performance Framework
PI	Performance Indicators
R&D	Research and Development
TRL	Technological Readiness Level

Project Description

DRIVER evaluates solutions in three key areas: civil society resilience, responder coordination as well as training and learning.

These solutions are evaluated using the DRIVER Test-bed. Besides cost-effectiveness, DRIVER also considers the societal impact and related regulatory frameworks and procedures. Evaluation results will be summarised in a roadmap for innovation in Crisis Management and societal resilience.

Finally, looking forward beyond the lifetime of the project, the benefits of DRIVER will materialize in enhanced Crisis Management practices, efficiency and through the DRIVER-promoted connection of existing networks.

DRIVER Step #1: Evaluation Framework

- Developing Test-bed infrastructure and methodology to test and evaluate novel solutions, during the project and beyond. It provides guidelines on how to plan and perform experiments, as well as a framework for evaluation.
- Analysing regulatory frameworks and procedures relevant for the implementation of DRIVER-tested solutions including standardisation.
- Developing methodology for fostering societal values and avoiding negative side-effects to society as a whole from Crisis Management and societal resilience solutions.

DRIVER Step #2: Compiling and evaluating solutions

- Strengthening crisis communication and facilitating community engagement and self-organisation.
- Evaluating solutions for professional responders with a focus on improving the coordination of the response effort.
- Benefiting professionals across borders by sharing learning solutions, lessons learnt and competencies.

DRIVER Step #3: Large scale experiments and demonstration

- Execution of large-scale experiments to integrate and evaluate Crisis Management solutions.
- Demonstrating improvements in enhanced Crisis Management practices and resilience through the DRIVER experiments.

DRIVER is a 54 month duration project co-funded by the European Commission Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 607798.

Executive Summary

The present document, *D23.21 - Performance and Effectiveness Metrics in Crisis Management Experiments*, aims at defining the right measures to be used in capturing experimental data on the one hand and in gathering the views of a heterogeneous community of interest in relation to these measures. The main output is the ***provision of practitioner guidelines and recommendations for selecting measures and indicators of performance, ways of collecting, processing and storing data and ways to operationalise these in a specific experiment.***

The work presented here is part of the DRIVER Test-bed development that intends to enable a structured and efficient capability development process in Crisis Management. The DRIVER Test-bed builds upon the results of the project ACRIMAS, which identified the Concept Development & Experimentation (CD&E) methodology as a suitable approach for Crisis Management capability building. CD&E adapts basic scientific methods to the concept development and validation process in the military and defence domain. DRIVER intends to draw upon such results to build up a consistent and homogeneous pan-European capability building methodology for Crisis Management.

In D23.21 we introduce a common process and key concepts that will ensure coherent performance measurement in experimentation. The common process for identifying and evaluating performance and effectiveness measures contains the identification of generic performance indicators as well as a standardized procedure to identify specific experiment-driven performance indicators.

The framework provided by the DRIVER Test-bed is not specific to any of the DRIVER dimensions (societal resilience, professional response, training and learning) because the actual key performance and effectiveness is different in each of them. Instead, the framework has to be understood as a general procedure that SP3-4-5-6 experiments should adapt to the needs and goals of the involved end-users. The obtained results will be utilized for the guidance methodology developed in the subproject Test-bed, task 202.2.

1 Introduction

1.1 Foundations of DRIVER Methodology

Dealing with innovative solutions in Crisis Management (CM) demands a structured and systematic approach to discover the needs and potential for innovation, to ensure knowledge development and identify lessons learnt. Potential solutions may cover one or more dimensions within CM: human factors, processes, technology, regulations or organisation.

The purpose of developing, conducting and assessing experiments within DRIVER is threefold. First and foremost, experiments are used to demonstrate how the practical value of specific innovative Crisis Management solutions can be assessed in a pragmatic and systematic way. Second, they are used to test the usability of the DRIVER methodology and consequently to improve it to a point where it can be used as a practical guideline by practitioners outside the project. Finally, they are also used to evaluate and improve the DRIVER Test-bed itself, a support environment that simplifies the task of implementing, running and assessing the experiments.

The underlying methodological approach that is followed by the project is known as *Concept Development & Experimentation* (CD&E) and originates from the military domain.[1] CD&E defines a way to develop new concepts, by experiencing the challenges, developing and evaluating the new concept in a realistic setting before expensive resources are being acquired or before organisational changes are being implemented. CD&E is a creative process where a concept is developed through brainstorming, evaluation sessions and analyses combined with input from experiments.

Alberts et al (2002) [3] remark that the focus of the experimentation is twofold: i) all key concepts should be understood by all experimentation members and advisory practitioners involved, and ii) the focus should be set on refining the concept in a learning-by-doing approach. The development of a concept starts with an initial concept idea. It starts from either a need or capability gap, or from a new opportunity or new solution. The lines of development are defined and the concept grows to include all. It is important that the concept matures along all the determined lines of development already in the early phases of the CD&E process until the concept can be demonstrated or trialled in a relevant operational setting. During the development, the concept will be assessed in experiments, including some or all lines of development. These will provide important input for the further development of the concept, or its rejection if it does not provide added value. This is the iterative nature of CD&E. The final evaluation of the steps of a CD&E process results in an evidence-based recommendation with respect to the proposed new concept. The concept has now matured and is ready for implementation [4].

Hence, in the CD&E new solutions and ideas are iteratively tested (multiple scenarios, interoperability, etc.) by a series of controlled experiments addressing different research questions.

[1] According to NATO, CD&E is one of the tools enabling the structured development of creative and innovative ideas into viable solutions for capability development [1]

Results can be then used to further develop the concept until operational capability is reached. Concepts can also be rejected, if it turns out that they do not provide added value or are not cost-efficient.

Within DRIVER we adapt and adjust the CD&E process to the CM domain in general and the Test-bed in particular. More specifically, the CD&E approach is used as a method that will support the evaluation of new solutions. Starting with small cases, the solution requirements increase through a higher complexity of the test cases, e.g. by adding more CM organizations, extending the period of relief operations or considering cascading effects. In a context where science and research meet the "real world" the term "experimentation" might be misunderstood as a classical experiment in natural sciences focusing on quantitative research methods (like setting up controlled laboratory experiment environments or the evaluation of specific hypotheses). Looking at the main objectives of DRIVER – improving the capability development in CM, identification of promising solitions and creating a more shared understanding of CM across Europe – the nature of the experiments must be interpretivist. I.e. the performance of a particular solution has to be reflected from the practitioners point of view, hence in a qualitative manner. CD&E explicitly suggests several qualitative data collection techniques in order to identify the impact of a particular concept to a given problem or need, e.g. the execution and analysis of interviews during observations. This in turn allows DRIVER to explore and discover "real" performance and effectiveness in CM as perceived and experienced by those who are actually doing the work in the field. Of course, a couple of existing generic performance measures can and are already applied in DRIVER experiments. It will be mainly the analysis of a context-dependent (user or system) experience of solutions being able to assess its relevance, usefulness, and maturity. At a first glance such a mixed approach of quantitative and qualitative research appears to be contentious because of being eclectic (i.e. building upon contrary philosophical assumptions and epistemologies). However, in context of a demonstration project like DRIVER it is the required method of choice in order to meet the practitioner's needs in learning, experiencing and understanding the performance of new CM solutions.

1.2 Document Overview

Besides the introductory chapter, the deliverable contains the following two sections:

Section 2: *Measuring Performance in Crisis Management Experiments* introduces the performance measurement scope in DRIVER experiments and provides several methodologies how coherent performance measurement in Crisis Management must be identified. This section clarifies the used terminology and gives an overview of applied methods. Based on elaborated Crisis Management functions, an iterative approach to experimentation is given as well as considerations for establishing meaningful performance indicators.

Section 3: *Guidelines and recommendations* is divided in two parts. The first one devoted to provide general guidelines for establishing effective metrics. Examples of performance and effectiveness metrics in CM are given. The general guidelines are followed by recommendations in specific topics such as data collection, creation of questionnaires and selection of testing groups.

2 Measuring Performance in Crisis Management Experiments

When following the CD&E based approach for collaboratively creating a valuable capability for Crisis Management, several general activities have to be conducted. First, the performance measurement scope including all perspectives and relations has to be clarified. Then, a basic concept and corresponding objectives and research questions need to be elaborated. These may be of preliminary nature and must be refined to the practitioners' and experiments' needs during the experimentation planning process. Based on the identified objectives, different measures of performance and effectiveness can be derived. With these measures at hand, a set of specific indicators and target values needs to be defined in order to enable an evaluation of the tested solutions. The development of appropriate measures and metrics is a non-trivial task in a Crisis Management experiments because effectiveness has a different meaning for the involved relief organizations (e.g. fast rescue by fire fighters vs. adequate transportation of wounded people by paramedics) and it varies over time (e.g. consideration of costs in the immediate response vs. the reconstruction phase). Besides, the solutions involved in a CM experiment require an appropriate consideration and differentiation concerning the contribution to mission objectives. In order to meet these needs, the performance measurement approach of DRIVER experiments must provide (i) a general set of rules and guidelines regarding measurement requirements and (ii) appropriate procedures describing how specific performance and effectiveness measures can be found.

2.1 Performance Measurement Scope in DRIVER Experiments

Before presenting the guidelines how relevant performance indicators will be identified, it is important to define the scope of what should be measured during DRIVER experiments. Because of the functional complexity of specific measurement "objects", the first step is to categorize them according to the DRIVER logic before specific Key Performance Indicators (KPIs) can be identified. The following figure illustrates the architecture of the DRIVER performance measurement dimensions.

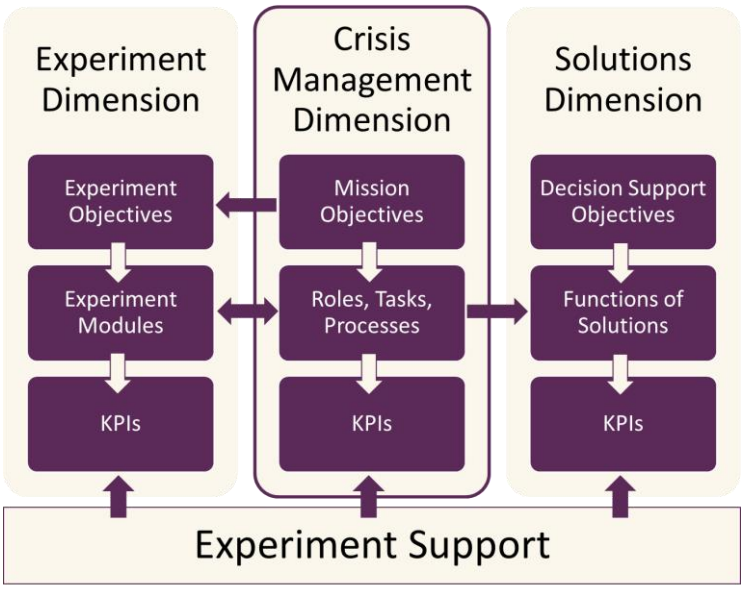

Figure 1: Performance Measurement Dimensions in DRIVER Experiments

The three dimensions include the experiment dimension, the solution dimension and – as the core DRIVER dimension – the Crisis Management dimension. All three performance measurement dimensions are served by an overall performance measurement experiment support, where all potentially relevant guidelines, recommendations, and experiment data is collected, stored and processed (e.g. the actual KPI definition guidelines, generic KPIs, domain-specific KPIs or data storage policies).

The experiment dimension covers the perspective of the experiment owner and measures all relevant data which related to what the predefined experiment objectives. One example in case of volunteer management could be the question how many voluntary (unpaid) participants could be motivated to join an experiment or in case of logistics if all relevant tasks could be executed within the planned experiment time frame. The experiment objectives are defined by the experiment owner, but the main source are the CM end-user needs and, hence, the objectives of the missions being simulated in an experiment. In order to "operationalise" the experiment objectives, experiment modules are derived (e.g. communication and coordination of volunteers taking part in the experiment). Within this module, the experiment owner is able to define which processes are required to fulfil the objectives and assign specific weighting. This step contains an estimation of the effectiveness of each process (with relation to the experiment objectives). Once this task is done, the experiment owner can apply the PM guidelines to deduce specific and relevant KPIs.

The CM dimension is, however, the key performance measurement area. The identification of CM objectives, described as mission objectives, is the foremost place to indicate whether a change of a process, the application of a new technology or a training module has an impact on the CM performance. Besides, the CM objectives need to be understood as the determining element of experiment objectives and the decision support objectives. Due to the different relief situations, stakeholders and time horizons the measurement objects vary in terms of specific roles, tasks, and processes. The question if a particular performance is effective or not can only be evaluated once the

involved actors and their responsibilities and practices are defined. These definitions have to be used to identify and configure the appropriate KPIs.

Finally yet importantly, the solutions dimension must be measured in order to learn whether a particular piece of technology or a new process has the potential to drive innovation in CM. The solutions objectives have always a relation to ease or support one particular decision making process, even if this is only defined as a new standard operational procedure. Hence, the decision support objectives build the first starting point for evaluating the performance of a particular solution. These objectives need to be derived or at least have a direct relation to the CM objectives, in terms of a practical impact. The identified objectives can be used to extract specific solution functions which in turn can be used to derive appropriate KPIs. One important aspect here is that the KPIs need to have a relation to the CM KPIs. To give an example, a high usability of software might be absolutely irrelevant because the software itself has no contribution to the relevant CM performance (which does not mean, usability does not have to be measured, but its CM impact is key for the overall evaluation).

Having the three dimensions and its interrelations in mind, this document (and the work in D23.21) provides a guidance to identify relevant KPIs. This process is supported with generic rules of performance measurement approaches (including a dedicated SotA analysis with links to existing performance measurement approaches), guidelines and recommendations.

2.2 Iterative Development of Experiments

Experimentation for capacity building requires a continuous development due to the different interest of practitioners and the complex interaction between CM solutions. In general, the monolithic approach of designing completely the experiment followed-up by a one-shot execution is too simplistic to achieve successful results.

As indicated in chapter 1, DRIVER leverages from the CD&E approach to establish an iterative development through a series of frequent and rapid periods. A practitioner may choose among different tools and methods in a single iteration depending on the kind of experiment, such as brainstorming, focus groups, evaluation sessions, simulations, analyses in operational settings, etc.

Regardless of the option, the following key rules should always be satisfied:

- The iteration begins with an assessment of the current status followed by the planning of the future work. The whole experiment team will agree on the concrete goals for the next increment, prioritizing the objectives. The team should focus on a manageable (small) number of specific areas of improvement for each iteration.
- All personnel involved (experiment owner and staff, volunteers, external evaluators, observers, etc.) has to be informed of the objectives.
- Speed is critical to enable the team to schedule a realistic amount of work. In case that the estimated workload for achieving the primary goals is too large, the objectives shall be broken down into smaller ones and only the less important ones will be delayed to a future iteration. As a general rule, no iteration should last more than 3-4 months.

- Frequent meetings are organized where all team members review their progress and new short-term tasks are assigned. This meeting should not exceed 15-30 minutes.
- The results are reviewed at the end of the iteration. The team assesses the work done in the iteration and the progress preformed for the whole experiment.

Take for example an experiment to enhance closer civil-military cooperation in a humanitarian crisis. A typical first iteration can be the identification of relevant organizations that support humanitarian operations at different levels, ranging from the field to policy makers to operational personnel, and to have a focus group to gather lessons learnt on past experiences. The objective would be to organize a workshop were current practices are analysed, and the iteration outcome should be a list of obstacles and bottlenecks identified in the workshop. At the end of the iteration, the team should agree on the main 3 obstacles, in agreement with the external stakeholders involved, and draft a plan on how to address them. At this point the first iteration ends, setting an initial guidance on how to focus a meaningful experiment for this topic. Needless to say, these objectives have to be further defined or may even be replaced by others during the experiment development process.

2.3 Objectives and Aimed Capabilities

In order to identify objectives in CM experiments, the first step is to identify and structure practitioner realities involved in CM, e.g. specific tasks, processes or workflows. The ACRIMAS project [8] has developed a functional model that categorizes three main types of CM tasks, also called *functions*, each of them divided into sub-functions (see Figure 2). The three main groups of functions are the following:

- **Preparatory functions, which** aim, prior to crisis events, to improve the capabilities to carry out operational and supporting tasks.
- **Supporting functions,** which support one, more or all operational tasks during a crisis;
- **Operational functions,** which are directly involved in minimising the effects of crisis events (e.g. damage reduction and avoiding cascading crisis events).

From a thematic perspective, the needs and gaps that are addressed by DRIVER have been detailed in the state of the art analysis at the beginning of the project [9]. DRIVER SP3 mainly contributes to the tasks “Inform and involve the public” (operational) and “Community awareness raising” (preparatory). SP4 provides tools for enhancing different supporting tasks as shown in Figure 2; details on needs and gaps covered by SP4 can be found in Annex 4 of D42.1 – Final report on architecture design [10]. SP5 supports the tasks “Training and Exercise” and “Evaluation” (preparatory) and “Inform and involve the public” (operational).

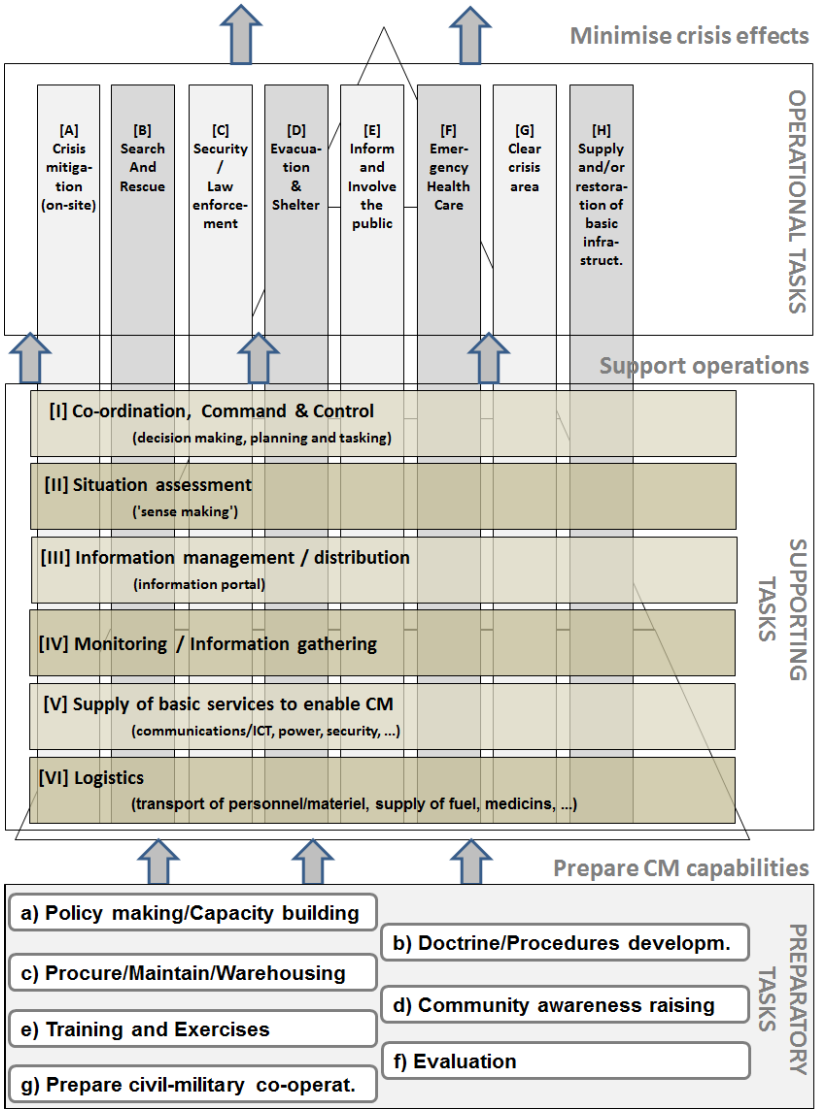

Figure 2: Functional classification of CM tasks.

These CM functions offer a common ground for all members of the CM community to discuss and agree on the key areas an experiment might focus on. For instance, the experimentation team usually starts with a rough idea of the issues and possible solutions that need to be studied. However, the participants might not have a clear understanding of the way an experiment should be designed and of the key indicators that need to be measured in the experiment. In such cases, focus groups or expert interviews can help to broaden the vision on the concept and subject under study. A simple demonstration of the solutions, followed by a structured discussion can help to design an improved experiment that addresses the specific objectives of the CM practitioners. Following the iterative approach described in section 2.1, the experiment team starts with gathering the basic issues and needed capabilities with different stakeholders. Based on the outcomes of this first iteration, a first set of research questions and measures can be elaborated. The example below describes this process in DRIVER applied to the sub-function “volunteer management”.

Example: Objectives and research questions volunteer management

The sub-function "volunteer management" was identified as a gap or useful addition to the European Crisis Management functions (see [8]). It was described as part of the operational task "Inform or involve the public", which concerns Early Warning of threatened people and Communication Management to the public and the (social) media providing information on a (threatening) crisis event, including realistic guidelines on:

a) Methods and solutions that allow individualized informing of the citizens based on the "need to know" and "able to understand" considerations;
b) Methods and solutions that allow efficient use of citizens as auxiliary resources that are activated and managed as a part of the overall Crisis Management system.

In order to assess the added value of the function or operational task "Inform or involve the public", one consequently has to assess the following:

1. To what extent does Crisis Management become more effective, when the general public is informed on behavioural requirements before, during and after an incident? Does the added value outweigh the risks and the costs (taking into account the solutions at our disposal)?
2. To what extent does Crisis Management become more effective, when volunteers are better prepared? Does the added value outweigh the costs?
3. To what extent does Crisis Management become more effective, when spontaneous volunteers are pre-organised? Does the added value outweigh the costs?

This in turn can be tested during DRIVER experiments. Potential key research questions, and objectives could be:

<u>Better informed public:</u>
Research questions: (how) can the offered methods and solutions contribute to this goal? Does the improved informing change, e.g., the amount of people that are able to help themselves? Does it change the amount of people that are able to help others or lower the total costs of the aftermath phase?
Objective: Investigate the benefits of informing the public during a crisis event and examine the effects on total costs.

<u>Better prepared volunteers:</u>
Research question: (how) can the offered methods and solutions contribute to this goal? Does the improved preparation, e.g., reduce the occurrence of psychological and physical harm for volunteers? Are volunteers able to perform their tasks more efficiently?
Objective: Investigate the effects of improved preparation of volunteers on physical/psychological harm and effectiveness of volunteer operations.

<u>Affiliated volunteers:</u>
Research question: Does pre-organisation simplify the management, e.g. by reducing the amount of time that professionals (or professional volunteers) need to handle spontaneous volunteers?
Objective: Investigate the effects on time and cost for management processes when comparing pre-organized versus spontaneous volunteer acquisition.

2.4 Identification of Performance Drivers

A literature analysis of publications dealing with performance measurement in Crisis Management[2] within Scopus, the largest abstract and citation database for peer-reviewed literature[3], results in more than 500 hits. Comparing the number of relevant sources per year, it becomes obvious that there is an ongoing increase of results starting in the early 2000s with a peak in 2006, which is two years after the major south east Asian tsunami disaster. Since then, publications have increased even more with a maximum of more than 50 sources in 2013 (see Figure 3).

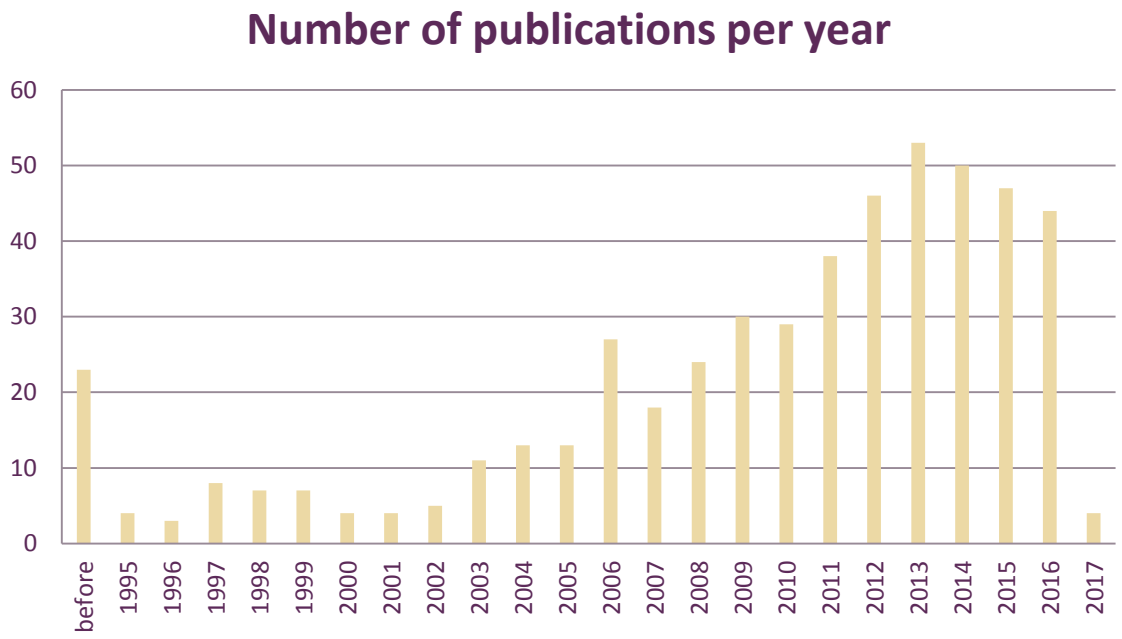

Figure 3: Number of publications per year

Results stem from a wide variety of research areas, including Social Sciences, Business and Management, Engineering and Computer Science among the top ones (see Figure 4).

[2] The search term („performance measurement" OR effectiveness) AND („crisis management" OR „disaster relief") has been used.
[3] https://www.elsevier.com/solutions/scopus

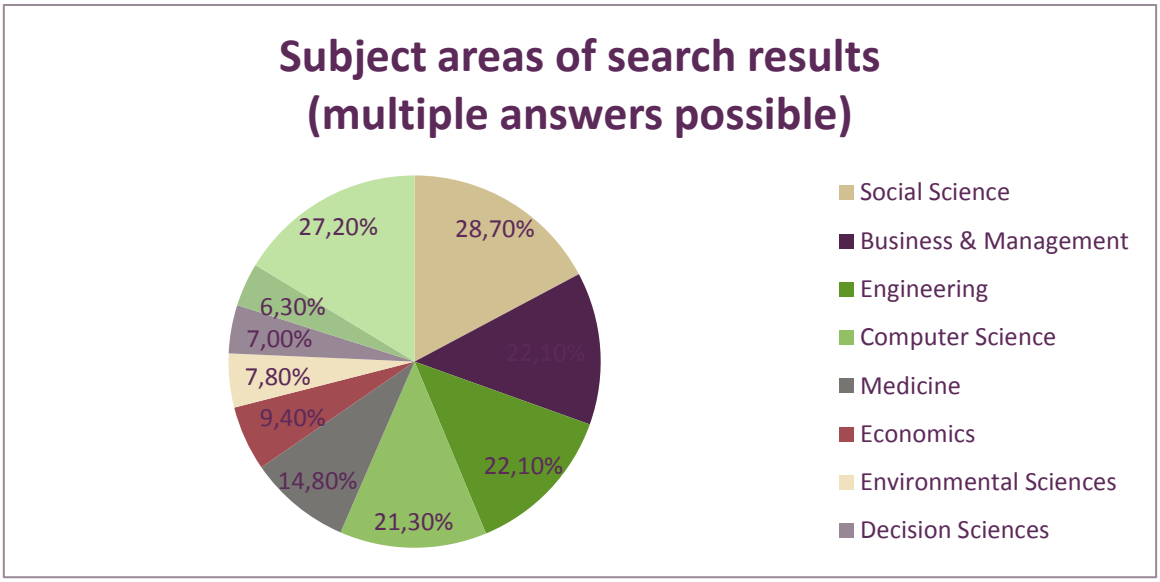

Figure 4: Subject areas of search results

However, when reviewing the first sources, it becomes apparent, that especially by including "effectiveness" in the search term, the search delivers many results which only measure the effectiveness of e.g. the model or algorithm that has been developed. Having excluded the term "effectiveness", the search only leads to 14 results. The excluded sources are not necessarily concerned with the performance measurement of Crisis Management itself but rather try to provide a solution to a problem and aim at measuring the effectiveness of this specific solution. Nonetheless, they are of interest, as even when giving statements about the effectiveness of one single model, algorithm etc. certain methods for doing so have to be developed or applied.

Considering only sources which deal with experiment-related performance measurement by extending the search term, i.e. adding "AND experiment" to it, leads to a significantly less amount of sources. Only 34 out of the over 500 publication satisfy the refined search term. While the yearly amount of publications shows similar behavior as the first search, the results of the refined search mostly stem from more technical research areas such as computer science or engineering (see Figure 5).

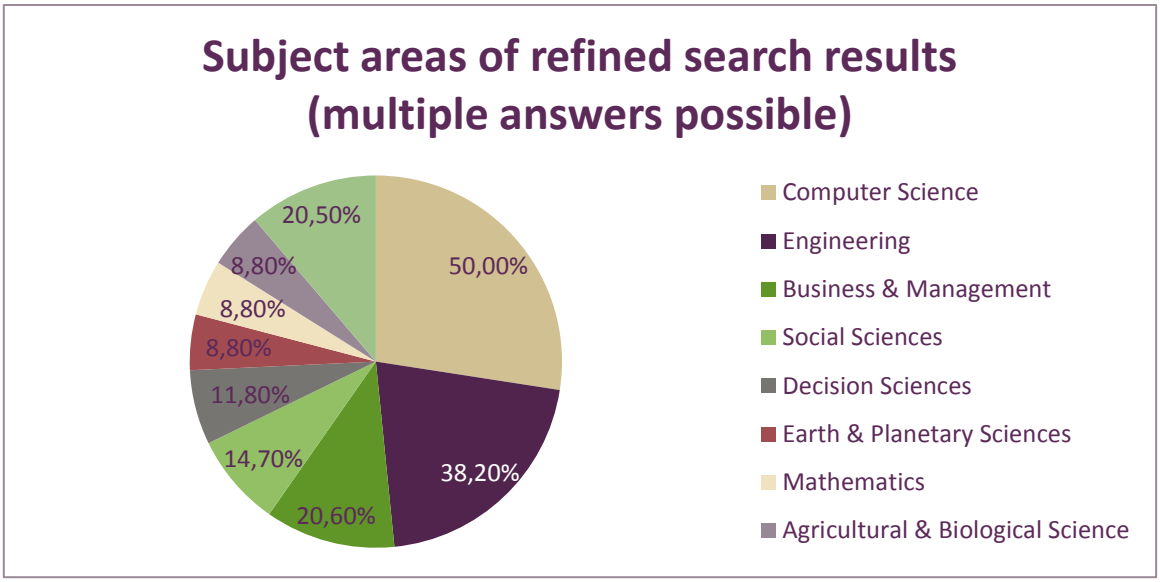

Figure 5: Subject areas of refined search results

One major insight of the sources is that the performance measurement of disaster relief operations is a difficult task due to various challenges which can be classified in four categories [40]:

1. *Evaluation based on value judgement:* Any evaluation needs to be based on certain values, i.e. in order to be capable of assessing how successful an operation has been there need to be values which define what is supposed to be successful. At least implicitly, these values, which can also be understood or formulated as objectives of an operation, will always be based on subjective opinion and personal beliefs.
2. *Complexity of crisis situations:* The high complexity of crisis situations significantly affects the way how a relief operation can be analyzed, understood and evaluated. High dependencies and complicated relationships between actors as well as causal ones lead to great difficulties when trying to understand what happened as well as why it happened.
3. *Questionable validity of information:* When evaluating an operation, this evaluation has to be based on information about how the course of events during the operation. Often such information is gained by conducting interviews etc. and rarely based on e.g. ongoing data collection. Consequently, there is always the question how reliable humans are as a source of information and therefore how valid the information is on which an evaluation is based upon.
4. *Limiting operation conditions:* Every disaster relief operation can have negative effects or outcomes, which simply could not have been prevented, independently of how successful the operation has been. Any immediate and unavoidable casualties caused by the crisis should not be included into the evaluation of an operation. For example, the number of injured people is not relevant for the evaluation while the time until they receive help is. Overall, it can be difficult to distinguish between the evaluation of the operation's performance itself and the analysis of what might have happened under different circumstances.

In general, these challenges make it difficult to establish a generic performance measurement approach for Crisis Management. Hence, the SotA analysis confirms the iterative approach to develop relevant KPIs in a participative way. Owen et al. (2016) also concludes that before suggesting measures to assess Crisis Management, it is necessary to understand the complex and often intertwined challenges and relationships of Crisis Management and its actors. Only after making sense of the interaction between the underlying value base and the events and actions of an operation, methods to measure and evaluate the performance of it can be developed [41]. Overall, already the short literature search shows that the high variety in scenarios, tasks, stakeholders etc. related to disasters, results in a lack of generic performance indicators – especially when evaluating Crisis Management experiments. The results of the SotA analysis are stored as they might be reconsidered for specific measurement approaches (e.g. an open KPI set for humanitarian logistics scenarios, see [41]).

To sum it up, the identification of value adding concepts or solutions in CM can only be achieved through a context-dependent or specific measurement, analysis and adjustment of its exemplary application in artificial environments, experiments, trials, serious games or exercises. For performance measurement a set of indicators and its explicit relation to tasks, processes and organization- or mission-specific targets is necessary. Each indicator might have a weighted importance for the overall performance, and that is the reason why the identification of KPIs is needed. KPIs represent a set of measures focusing on those aspects of organizational performance that are most critical for the current and future success of the organization [11]. A systematization and categorization of potential KPIs prevents an isolated view and possible misinterpretation. Thus, the indicators can be related to each other and weighted by targeting specific objectives [39]. However, because of the different actors involved in CM operations, different processes with specific objectives and relations need to be considered for the identification of relevant KPIs for an experiment. The findings from the SotA described above offer a huge source of potentially appropriate indicators. They can be considered as an open set of CM KPIs. However, in order to ensure the relevance of the KPIs in particular DRIVER experiments, a specific set of KPIs needs to be developed for each experiment. The achieved results (i.e. the identified KPIs and its application during experiments) need to be stored and documented in order to become part of a DRIVER specific KPI set in the DRIVER Test-bed. Thus, instead of providing a compilation of existing KPIs (and metrics which probably are not describing the effectiveness of tasks or solution) the present document suggests procedures and rules how to identify and develop appropriate measures. A recourse to the SotA will always be one of the first steps during the definition of specific KPI.

Below an example for the sub-function “volunteer management” is elaborated.

Example: Defining measures for volunteer management

In order to assess the added value of the sub-function “volunteer management”, one consequently has to define meaningful measures for performance and effectiveness. Exemplary measures applied in project management are:

- safety,
- time,
- cost,

- resources,
- scope,
- quality, and
- actions.

These measures need to be adapted to the specific context and objectives in Crisis Management. Staring from these general measures, an iterative approach is applied in DRIVER to continuously redefine and detail those measures to reflect specific aspects of a central sub-function or solution to be assessed.

Potential measures for volunteer management derived from the above categories could be:

Better informed public:
Measures: e.g., resources assigned to inform the public, costs of information/data preparation and provision

Better prepared volunteers:
Measures: efficiency of volunteer work, quality of volunteer work, occurrence of physical/psychological harm

Affiliated volunteers:
Measures: e.g., time needed to manage volunteer work, resources needed to manage volunteer work

Different parameters can then be applied to assign each measure with appropriate metrics. The parameters represent those capabilities or characteristics so significant that failure to meet the threshold value of performance can cause for the concept or system selected to be re-evaluated or terminated. The next section explains the meaning and usage of parameters.

2.5 Parameters and Scoring

In order to monitor the performance, the experiment team has to define a number of relevant KPIs. These KPIs provide a way to quantify the key outcomes of the experiment and assess the performance of individual parts of the experiments. To determine clear KPIs is an essential part of the experiment planning.

The number of KPIs that are defined for an experiment should be kept low in order to assure that the experiment design can be focused and to simplify the task of interpreting the results. Typically **3-5 parameters** are a reasonable amount for handling small/medium-sized experiments. For complex experiments, a manageable number of parameters can increase up to approximately **20 variables** [12]. On top of this, a number of KPIs must be defined at project level as a way to monitor the formal development of the experiments, asses the risks and trigger the risk mitigation procedures (if needed) in a timely and appropriate manner. Thus, the performance measurement approach might

consist of function or task specific KPIs, which are connected to the solutions as well as experiment specific KPIs, which are related to the overall CM mission.

As observed in the identified SotA and experienced in projects of several DRIVER consortium partners, very few of these can be defined upfront. This means that defining the appropriate KPIs must be an iterative task following a predefined process and considering dedicated rules. In the following three rules to be considered during such a process are described.

2.5.1 The SMART Rule

Peter Drucker's management by objectives [15] set a basic criteria for the identification of KPIs. As a mnemonic, Drucker refers to the SMART rule that summarizes conditions that every well-designed indicator must meet. The five conditions that lie behind the abbreviation **SMART** are:

- **S**PECIFIC: it has to be clearly stated what the parameter is measuring. That is, the parameter to be measured must be defined in a way that does not leave place for interpretations by different observers. So in here a series of specific "W" questions must be answered:
  - What is to be done?
  - Who is involved in?
  - Where would it be done?
  - Which are the requirements and constraints?
  - Why is this needed to be done?
- **M**EASURABLE: the parameter must quantify the progress of a capability. This condition answers the question "how do you know it would meet expectations?" and also helps defining the objectives using a series of assessable terms such as quantity, frequency, costs or deadlines.
- **A**CHIEVABLE: the resources to measure a parameter must be realistic and negligible compared with the overall effort that is required to define, organise, execute and assess the experiment. In other words, it could answer the following question: "Is the person/team able to achieve the measurable objectives?
- **R**ELEVANT: the parameter must be an indicator that is of strategic interest for the experiment owners and isn't already quantified by another indicator. Parameters that do not provide new and important (relevant) insights that may influence decision-making are useless. So it will answer if the experiment should be done, why and also what the impact will be.
- **T**IME-BOUND: the indicator must represent the state of the experiment at a certain moment in time when the (iteration of) the experiment is executed.

2.5.2 Metrics and Scoring

All KPIs are meant to provide measurable quantities even though in some cases they may represent qualitative metrics. Below we provide an overview of common metrics.

Binary

Some aspects or services can only be assessed in a binary way, indicating whether a desired state is present or not, with a **yes/no** metric. These aspects have nothing to be gained by improving them beyond the level of adequacy. An example would be compliance with standards; no additional work is needed to improve a solution in this respect if this requirement is already satisfied.

Numerical

The level of performance can be directly measured numerically as they can be counted in a simple mathematical way. It is easy to establish a KPI for numerical aspects because quality is expressed with,

- Absolute numerical value: for example, the total number of volunteers taking part in an experiment or total number of factors that has to be considered by a decision-maker.
- Proportion in relation to a baseline: for example, the percentage of network bandwidth used or the relative humidity of air.

Subjective

Subjective measures are linked to a subjective or qualitative assessment. They are usually expressed by integer numbers according to ranges of the level of performance. For instance, in customer satisfaction rating surveys ranging from 1 to 5 (1=very poor, 2=poor, 3=acceptable, 4=good, 5=very good).

An important aspect in evaluating the collected data of an experiment is to carefully consider target values for each metric. The outcome of an experiment strongly relies on defining what values of a parameter indicate success and what parameters result in the failure/nonfulfillment of an experiment's objective.

Example: Setting parameters: Applying the SMART rule

Following the example of "volunteer management" (section 1.1), different parameters can be set up according to the SMART rule.

In terms of planning:

- Minimum/maximum number of volunteers needed for the experiment (numerical value)
- In case of hosting an experimentation campaign in different countries, it may be needed to translate the documentation to the native languages.
  Cost of the translation (numerical value in euros)

In terms of organization and personnel:

- Are volunteers capable of performing their duties without supervision? (binary answer: yes/no)
- If they need supervision, ratio of volunteers per supervisor (percentage)

In terms of equipment and systems:

- Number of mobile devices that volunteers have to carry for an optimal operational efficiency (numerical value)[4]

In terms of training:

- Do volunteers require of any training previous to the experiment execution? (binary answer: yes/no)
- Percentage of volunteers that require training (numerical value)
- Total cost of training per volunteer (numerical value in euros)

In terms of evaluation, let us assume that volunteers will be interviewed to provide their feedback after the experiment:

- Total number of interviewers (numerical value)
- Ratio of volunteers interviewed (percentage)

2.5.3 Scoring and Decision Making

In the preceding sections KPIs as direct measurements of key aspects in the experiment. One important topic within CM experimentation is the need to involve stakeholders with different interest to enrich the experiment and test several aspects of capacity building. This implies the necessity to define common parameters in order to assess whether objectives of an experiment have been met.

Necessarily a collaborative methodology must be put in place to reach trade-off interests between all involved parties. All stakeholders must have the opportunity to be involved in the experiment from the beginning, including during the objectives and parameters setting. The most effective way is to organize regular meetings with end-users and practitioners to assess the progress. Below, some options are introduced to foster a collaborative parameter development.

- Parameters exposition: with this technique an initial list of parameters is presented in a workshop. The aim is to offer an idea of what the experiments' measures are and foster the refinements.
- Parameters wall clustering: is an exercise to group parameters based on objectives or complementary activities. Discussing the parameters associations and deciding the appropriate clustering ensures the end-user involvement. Furthermore, it helps to analyse the experiment from different points of view. Taking a DRIVER example, an experiment

[4] Notice that the *effective operational efficiency* can be seen as subjective parameter itself, measured from 1 (very poor) to 5 (very good) by the evaluators.

involving SP3 and SP5 solutions should be analysed from both point of views, civil societal resilience and training and learning.

The overall assessment of the success of a concrete experiment or aspect of that experiment can be very complex and time-consuming. This is especially the case, when considering a set of heterogeneous and possibly conflicting parameters. A popular example is the balance between cost, time and resources. Reducing e.g. transportation time of relief goods to the crisis location is most often associated with an increase in cost and assigned resources. To be able to decide on the performance and effectiveness of a proposed CM solution, the experiment team or measurement framework applied has to be able to evaluate multiple, partly conflicting parameters against an overall objective. The underlying problem is commonly known as Multiple-criteria decision-making (MCDM) or multiple-criteria decision analysis (MCDA). An introduction to this approach can be found e.g. in [18]. One good applicable method for decision making for these kinds of problem statements is the analytic hierarchy process (AHP).

The AHP is a structured technique for organizing and analysing complex decisions, based on mathematics and psychology. It has particular application in group decision making and is used around the world in a wide variety of fields of application, from government, business, industry, to healthcare and education. It is an interesting approach to be integrated into DRIVER, because it builds upon collaborative decision making enabling the consideration of different interests and points of view (e.g., practitioners, system developers, crisis managers and experiment leaders etc.).

The procedure for using the AHP can be summarized as:

1. Model the problem as a hierarchy containing the experiment's objectives, the alternatives and measures for reaching it, and the parameters for evaluating the alternatives.
2. Establish priorities among the elements of the hierarchy by making a series of judgments based on pairwise comparisons of the elements. For example, when comparing pre-organized and spontaneous volunteer assignment, the crisis managers might say they prefer timing before costs and costs before resources.
3. Synthesize these judgments to yield a set of overall priorities for the hierarchy. This would combine the stakeholders' judgments about costs, price, quality and timing for each assessed capability/solution.
4. Check the consistency of the judgments.
5. Come to a final success evaluation based on the results of this process

This procedure can well be reflected in the methodology of DRIVER and the CD&E approach (see Figure 1). Starting with the initial concept and refining associated parameters continuously in close collaboration with practitioners, experiment leaders of SP3-4-5, solution and platform providers and crisis managers until an aimed capability is achieved or all objectives are met. This might also include the nonfulfillment of several aspects and the overall decision against implementing a proposed solution.

3 Guidelines and Recommendations

3.1 General Guidelines to Establish Metrics

As explained in chapters 1 and 2, the CD&E approach is of iterative nature. Namely, solutions and ideas are iteratively tested in experimental activities. This ensures that the development of both concepts and solutions is refined and validated from the beginning of the process until the implementation phase. This approach applies also to the establishment of metrics and measurement. From the initial experiment formulation until the detailed experiment plan, refinement may be needed with regards to the validity and applicability of methods. With this premise in mind, the guidelines provided below are general and refer only to:

- The experiment design phase (e.g. the formulation of the objectives; the definition of criteria of success etc.)
- Technologically driven experiments, namely experiments in which technical solutions are assessed for different purposes.

In order to establish relevant and adequate metrics, the following general guidelines can be used:

1. Clearly formulate the objective of an experiment and the research questions
2. An overall methodology must be decided to gather evidence to address the objective and research questions.
3. Ethical and privacy considerations should be taken into account. For instance: registering and storing personal data, combining data-sets.
4. A clear statement of the expected outcomes should be elaborated. For instance: less time needed, less errors made, more considerations taken into account during the decision-making process, less experienced workload and stress, effective task performance with less personnel, increased job satisfaction, increased number of volunteers, etcetera.
5. Clearly definite the criteria for success of the experiment. For instance, 10% less time needed, 20% less errors, 15% increased job satisfaction, 25% less costs.

Some typical objectives and research questions of an experiment may include:

- Test functionality and features of a single technology: Can a task be performed? Does the tool contribute to the function it is supposed to contribute to?
- Test a particular configuration of technologies (interoperability, benchmarking): are technologies working seamlessly with other tools to provide a given function or in conjunction with other functions (and tools therein) at system of systems level?
- Test effectiveness of (configuration of) technology in a given setting (for a particular user group or in a given cooperation scenario): are tasks performed faster and/or better?
- Test functioning and features of a single concept or functionality (part of an existing technical solution): can a task be performed faster and/or better?

- Test effectiveness of an organizational / procedural approach: are tasks performed faster and/or better?
- Evaluate cost-benefit of solutions / approaches: are certain technologies / approaches a good investment option for an organization (operational benefit in relation to life-cycle costs)?

Secondly, experiments will be designed differently depending on which level of Crisis Management is addressed. Experiment and mission objectives must address expected outcomes, and tasks and metrics must be designed accordingly. The following levels can be identified:

- Technical: test device or software
- Operational: improve operations in the field
- Tactical: improve situation awareness, command and control; improve decision making
- Strategic: guide investments in innovation; improve preparedness, capabilities, etc.
- Systemic: influence Civil Protection system in a MS and in the EU.

Example: Setting objectives and metrics for different levels of CM

Following the levels of Crisis Management, different objectives and metrics might be applicable to evaluate an experiment's success. Below are some common examples of these entities:

Technical: Possible objectives could be to assess, if a specific task can be performed with the support of the solution. Or if the solution would be perceived by practitioners as valuable support during a crisis event. The metrics applicable in this case could be the usability (w.r.t. system usability scale[21]) in the categories acceptance, the overall confidence level in the solution or the mean required training times.

Operational: Possible objectives at the operational level could be to assess, if the applied solution is able to contribute to distributing humanitarian goods more efficiently. This can be for instance evaluated by the mean transportation time or the average number of resources needed to accomplish a specific task.

Tactical: Possible objectives at the tactical level could be to assess, if the situational awareness and/or workload of a practitioner could be either improved or remains unaffected when applying the new solution. This can most easily assessed by SAGAT [22] or NASA TLX [23][24].

Strategic: Possible objectives at the strategic level could be to assess, if the proposed solution, process or concept improves the strategic decision-making effectiveness. Measuring the effectiveness of Crisis Management decision-making and processes could be performed by applying well-proven and well-known methods for integrated process improvement, e.g. Business Process Improvement (BPI) [25].

Systemic: Possible objectives at the strategic level could be to assess, if the proposed solution, process or concept has an effect on the Civil protection system. Possible metrics could be the number of standards for CM elaborated and proposed.

A third element to consider is the level of complexity and realism needed in an experiment, and the extent to which a controlled setting can and should be created. Some examples of different levels of complexity and realism include:

- Single device.
- Single technology in controlled environment (e.g. comparison of mobile devices).
- Range of connected technologies in controlled environment (e.g. information exchange between field and HQ).
- Human-computer interaction in laboratory.
- Table-top or serious gaming exercise testing tactical procedures
- Human-computer interaction, combined with technology testing, in realistic adverse conditions.

Within DRIVER, solutions or capabilities addressing different Crisis Management functions have to be experimented in various configurations reflecting the operational reality of EU Crisis Management cross-border operations.

A final important general recommendation refers to *triangulation,* seeking convergence and corroborations of results from different methods focusing on the same phenomenon [19]. Generally, mixed-methods research combining both qualitative and quantitative approaches should be used in all DRIVER experiments. The main reason is the above-mentioned performance measurement dimensions. For instance, a benchmark of different solutions can be executed with a quantitative approach (e.g. by runtime comparisons), but the analysis of its contribution to the CM dimension requires a qualitative approach (e.g. focus groups with crisis managers in order to discuss benefits and drawbacks of a solution). The same rationale can be applied to table-top exercises, if followed by in-depth interviews. A single method is unlikely to serve the objectives of complex experiments.

3.2 Recommendations and Common Problems

3.2.1 Quantitative and Qualitative Methods for Data Collection

Capturing relevant data during experiments is a crucial aspect of the experiment plan. Specifications in the process of data collections, methods and analysis of the evidence are essential part of the overall experiment design and execution. There is a great variety of quantitative, qualitative and mixed-methods that can be used in DRIVER experiments to collect data. For instance, a benchmarking experiment in which different solutions are compared with each other or with a reference solution, can rely both on a quantitative approach (e.g. by defining, for instance, a set of indicators to measure technological capabilities) and on qualitative methods (e.g. the organization of focus groups with crisis managers in order to discuss benefits and drawbacks of a solution). The same rationale can be applied to table-top exercises, if followed by in-depth interviews. A single method is in most cases unlikely to provide reliable results of complex experiments. Data can be collected using several methods. For instance:

Interviews

The spectrum of the interview format is broad, however in this context three main approaches should be considered:

- ✓ Structured: also referred to as standardized, are interviews in which the questions and the answer categories have been pre-determined and put in an interview schedule (more often close questions categorized to facilitate analysis). This method is used if the goals of the research are to produce statistical data. A structured interview is often defined as a questionnaire. Given the high importance of questionnaires as a tool for collection of structured data, their usage is discussed in a dedicated section (3.2.2).
- ✓ Semi-structured: the interview schedule is designed and used by the interviewer but flexibility is key here (e.g. the wording of questions, the possibility to ask additional questions not included in the interview schedule, etc.).
- ✓ Unstructured: as Dunn puts it, "the conversation in these interviews is actually directed by the informant rather than by the set of questions" [27].

The main obstacle of the interview method for collecting research data is that an interview is not always neutral and unbiased. There are certain factors which will influence an interview situation for the interviewer and an interviewee. Thus, it reflects subjective viewpoints within a specific situation. Details on common problems and recommendations on how to produce meaningful results with questionnaires – as an example of a structured interview – are given in section 3.2.2.

Focus groups

Focus groups are moderated group discussions among a group of selected individuals. This method can be traced back to the late 1940s thanks to studies carried out on the social and psychological effects of mass communication [28]. They usually involve a small group of people who gather together to discuss one or more specific issues with the help of a moderator. According to Stayaert and Lisoir, a focus group is a planned discussion among a small group (**4-12**) of stakeholders facilitated by a skilled moderator [29]. The aim is encouraging "a range of responses which provide a greater understanding of the attitudes, behaviour, opinions or perceptions of the participants on the research issues" [30]. The following are some of the most important features of a focus group [31]:

- It involves a small number of people in order to enable in-depth discussions on a specific area of interest.
- It is non-directive and needs an open climate.
- Interaction is a unique feature of the focus group which distinguishes the latter from in-depth interviews. Group interactions are in fact treated as research data.
- While the group process assists people to explore their point of view, a (or more than one) moderator introduces the topic (focus groups need a stimulus), guides the conversation and makes sure to obtain good and accurate information from the discussion.

- The sample (participants of the focus group) is well-balanced (age, gender, socio-cultural background, etc.)

The purpose of a focus group (and moreover participants in an experiment) is to get reliable and valid results representing the viewpoint of different stakeholders. The representative and well-balanced selection of participants is therefore of high importance, because it directly influences the outcome of an experiment. In section 3.2.3 some recommendations for selecting participants for focus and study groups are presented.

Surveys

Surveys are designed to produce statistics about a target population. This kind of research is used to generate data where the objective is to explicitly test hypotheses or investigate propositions about, for instance, attitudes or perceptions. When designing a survey there are some components which are of particular relevance such as sampling and designing questions. As claimed by Gerring, "in constructing a sample one should aim to be representative of a broader population, to include sufficient observations to assure precision and leverage in the analysis, and to use cases that lie at the same level of analysis as the primary inference" [32]. If the representativeness of a chosen sample is not considered relevant for the chosen area of analysis, it should be stated in the research design.

According to Fowler "designing and implementing a survey is a systematic process of gathering information on a specific topic by asking questions of individuals and then generalizing the results to the groups represented by the respondents [26]". Following this definition in DRIVER, surveys can be used to provide a representative overview of different stakeholder groups and their general perception of a proposed new concept, process or solution. This especially helps in early project and experiment phases to adapt the design process to better reflect stakeholders' viewpoints and objectives. Further on, survey results can help comparing and understanding divergent results between different stakeholder groups. Nevertheless, there are various aspects to consider when designing and planning a survey. An overview of common problems and recommendations for conducting adequate surveys for a specific problem statement are provided in section 3.2.4.

The choice of a research approach (quantitative, qualitative or mixed) and of specific methods depends on the objectives of the experiment and on the parameters chosen to measure their satisfaction. While providing a comprehensive list of methods is not the aim of this document, in the next iterations of this deliverable a common methodological framework will be provided. The next sections (3.2.2 - 3.2.4) provide more detailed recommendations with regard to the selection and design of data collection methods relevant for DRIVER.

3.2.2 Creation of Questionnaires

Recommendations are given with respect to using questionnaires, taking into account lessons learnt especially during the SP3-4-5 experiments.

Questionnaires are used as data collection methodology when carrying out a stakeholder consultation. Different methodologies are possible, like surveys, interviews, focus groups, workshops. Also, a combination of methods can be used. Surveys are often used to approach large groups of stakeholders. Interviews can be used to collect more detailed information from specific key stakeholders. Focus groups are often used to discuss about specific topics with a group of experts. Workshops can be used to have an interactive working session with a representative group of stakeholders. For a survey often different questionnaires are made for different stakeholder groups. Each questionnaire aims to collect specific information from each stakeholder group. Questionnaires can also be used as one of the data collection methods during interaction with stakeholders, as part of a focus group or workshop.

When designing a questionnaire it is necessary to identify the main topics that should be addressed. Moreover, open or closed questions should be used. When formulating the questions it is important to choose the type of question (open or closed). Closed questions are questions where you select the answer from a given list of possible answers. For example: “In which country is the head office of your organisation located?” In contrast, open questions give the respondent the opportunity to provide the answer in his/her own words. For example, “What are currently the most relevant trends in CM in Europe?” The number of open questions should be limited to make filling the questionnaire faster and easier to the stakeholder.

Questions should be as much straightforward and clear as possible. The (technical) terms used in the questions should be understandable for most of the respondents, or explained in plain words. It is also important to collect sufficient information about the profile of the respondent, so this can be used at a later stage to carry out the analysis of the answers to the questionnaire. If insufficient profile information is collected, it is in most cases not possible to collect this information afterwards. Missing profile information may lead to an incomplete data analysis. An important issue is to follow given laws and rules on data protection, when sensitive and/or personal data is collected. DRIVER has developed recommendations for ethical research including clear guidance for the partners on how to gather, process, store and delete personal data. This information has been set out in the deliverables D91.3 Ethical Procedures, Risks and Safeguards [33] as well in D95.21 Planning for Ethical Approvals [34] and provides the DRIVER partners with knowledge and practical guidance on how to follow data protection procedures, how and when to conduct ethical approvals, what to take into account for the inclusion of participants as well as how to mitigate risks and safeguard key ethical principles within DRIVER.

The length of a questionnaire is an important concern in any type of interview. One simple reason is that questionnaire length is directly related to completion rate. Long questionnaires tend to cause fatigue and discontinuation. In order to get a good response rate for the questionnaire it should not be too long. Different authors argue about a recommended maximal length (lying between 5 and 30 minutes), which highly depends on several factors including e.g. the media used for contacting respondents, the purpose of the questionnaire or the complexity of questions [26]. When using a

questionnaire that takes longer to complete, this can probably only be used if it is expected that the respondents are interested enough in the topic to complete the whole questionnaire. The respondents should have enough time to fill in the questionnaire; in most cases the respondents should have the questionnaire available for around 2-3 weeks before the deadline. If the questionnaire is filled in during an interactive session with the respondents this is not applicable. Also, as a general rule it is recommended to send a reminder close to the deadline, for example one week in advance. In specific cases, also telephonic reminders can be used, but this requires additional budget and is quite time consuming. It is also important to take into account the holiday periods of the intended respondents when scheduling the questionnaire. It is recommended to test the questionnaire first with a small group of respondents, before sending the questionnaire to the total group of respondents. Together with the questionnaire it is in most cases also useful to provide additional information, for example back ground information about the topic or examples.

For providing the questionnaire to the respondents, this can be done via paper or via Internet. In the past most questionnaires were done via paper, but nowadays questionnaires are more often provided via Internet. Questionnaires to a limited group of respondents can still be provided on paper. When approaching a large group of respondents it is recommended to provide the questionnaire via Internet. Various open-source tools for creating online questionnaires already exist and may be applied in this context (e.g., Google Forms[5], SurveyMonkey[6], TypeForm[7]). To select the right tool it is recommended to contact other DRIVER partners, which have expertise with using these tools. Most research organisations involved in DRIVER have extensive experience with using these questionnaire tools. When processing the responses given to the questionnaires it is important to take into account the existing European privacy regulations. The respondents should give their consent when they are asked to provide personal data. And they should be informed how personal data is stored and processed. If possible the collected responses should be anonymized, so they cannot be traced back to an individual. Also when presenting the results of the analysis of the questionnaires this issue needs to be taken into account.

3.2.3 Reliability and Validity of Groups Selection

To be actually able to not only detect but to isolate the reason for change, it is important to compare the cause-and-effect relationship in an experiment with a setting that runs under normal/ established conditions (not treated by the “cause”) where applicable via a control group.

This chapter focuses on the selection of participants and groups and as such provides recommendations on a) how to mitigate negative effects on validity related to participants and groups and b) how to set up and select a control group among the participants.

In general, as DRIVER experiments usually involve the participation of people, the first rule to follow is to take research ethics into account. DRIVER has developed clear and practical guidance to respect research ethics, customized for its purpose within the project and which are provided in D91.3–

[5] https://www.google.com/forms/about/
[6] https://www.surveymonkey.com/
[7] https://www.typeform.com/

Ethical Procedures, Risks and Safeguards [33]. Related to the topic at hand, i.e. Chapter 7 – "Recruitment of participants in research" of D91.3 should be consulted.

3.2.3.1 *How to Mitigate Negative Effects on Validity*

The following practical recommendations related to participants in experiments are based on [35] and partly adapted/reformulated and simplified to the civil and DRIVER internal use including implementation of own lessons learnt identified from previous DRIVER experiments. The main requirements on validity are:

1. ***Ability to Use the New Capability***

Common problem	Experiment participants can't use or employ the new capability effectively
Recommendation	Provide sufficient practice time for participants to be able to operate and optimally employ the new solution. Not only does the new functionality need to be available ahead of time, but also it should be clear to all participants well ahead of the experiment, how the solution is embedded in the operational context and how it is intended to be used.

2. ***Ability to Detect Change***

Common problem	Variability of participants within an experiment
Recommendation	• Consistency among participants' responses can be improved prior to the experiment by thoroughly training everyone to the same level of performance before the start of the experiment. • When possible, select similar (homogeneous) participants to reduce their variability[8]. • After the experiment, the experiment leader can assess the extent of participants' variability by comparing individual scores across participants (when possible). • When variability is a result of a few outlier cases, the experiment analysis can be performed with and without outliers to determine the impact of outliers on the analysis.

[8] It is important to notice that too much homogeneity is counter-productive since the experiment results could be isolated from actual operations.

3. ***Ability to Isolate the Reason for Change,***

3.a. ***in Single-Group Experiments***

Common problem	Participants change from experiment to experiment
Recommendation	• Monitor for participants' changes over the course of succeeding experiments. Participants may become more experienced and proficient, due to learning effect, or they may become fatigued, bored, or less motivated. Participants' changes over time will produce an increase or decrease in performance in later experiments unrelated to the new capability. • Counterbalance the sequence of experiments (e.g. NG-CG-CG-NG) so a sequential learning effect will affect the new-capability group (NG) and the control group (CG) to the same extent. • Ensure that participants are trained to maximum performance and operate at a steady state prior to experiment start. • Monitor for participant attention which might impact experiment results near the end of an experiment. When possible, compute each experiment's outcome for only those participants who completed all experiments. • After the experiment, analyse the experiment data arranged by time to determine if increases or decreases in performance over time occurred irrespective of the nature of the experiment. If temporal increases or decreases are found, analysis of covariance can be used (with caution) to statistically correct for unrelated temporal changes.

3.b. ***in Multiple-Group Experiments***

Common problem	Participant differences between experiment groups
Recommendation	• With large experiment groups, randomly assign individuals to different groups when possible. • With small treatment groups, use pair-wise matching when individual assignment to different groups is possible and pre-experiment data on all individuals is available for matching purposes. • Use each group as its own control when random assignment is not possible. Each experiment group should use the new capability and the old capability. • Avoid giving the new-capability group "extra preparation" for the experiment which would create artificial group differences (trained group difference). • Monitor for differential participants' dropouts from the different groups over a long experiment to avoid

	evolving artificial differences between groups as the experiment progresses.

3. ***Ability to Relate Results to Actual Operations***

Common problem	Non-Representative Experimental Unit
Recommendation	• Select experiment participants directly from an operational unit that will eventually employ the capability. • Use students, retired CM practitioners/ experts, or government civilians when operational staff is unavailable and the experimental task represents basic human perception or cognition • Avoid the temptation to over train the experiment group to ensure success. An over trained experiment unit is unrepresentative and referred to as a "golden crew." • Explain the importance of the experiment to the participants and their contribution to the effort to ensure the new capability can be thoroughly and fairly evaluated. • Monitor to ensure participants do not "underperform" out of lack of interest or resentment. This may occur when personnel are assigned to participate in the experiment as "an additional" duty and it is perceived to be unrelated to their real mission. • Monitor to ensure players do not "over perform" due to being in the spotlight of an experiment. This is known as the "Hawthorne effect". This effect is more likely to occur in highly visible experiments that have continual high-ranking visitors. In this instance, the participants are motivated to make the capability "look good"

3.2.3.2 Selecting a Control Group among the Participants

As briefly discussed before, control groups are vital for evaluating the cause-and-effect relationship between two variables of an experiment, thus making control groups essential to measure the effectiveness of a new capability/solution. Provided that research ethics [33] as well as the requirements with regard to the competences have been taken into account in the recruitment of participants, the recommendations under section 3.2.3.1 already indicate some guidance on how to select a control group among those participants.

In general, the term "control group" has to be taken with care in the CM context, as the general understanding of the term, at least in the scientific world, suggests that the experiment is done

under controlled conditions, limiting the varying component to one factor, which differentiates the experiment group from the control group. When testing new CM solutions in a most realistic scenario, it is not possible to set up such controlled conditions. The more we control, the less closer to a realistic scenario we are and hence, we may eventually compromise the ability to relate results to actual operations. However, by trying to consider the recommendations to mitigate negative effects, it is possible to optimize the usage of "control groups".

Following the objective of this deliverable to give practical guidance, the scientific background of control groups is not discussed here; instead the focus is on the pragmatic implementation in the DRIVER experiments based on the CD&E approach as used in the comparable military world for capability development.

Selecting participants for "control groups":

Although many different sampling techniques exist, there are basically two main approaches on how to assign individuals to groups (based on Lomann (2003) [36] and White, H., & S. Sabarwal (2014)[37]). The way how to "use" them in the best way to mitigate negative effects on the validity of the results has already been outlined above.

1. *Random assignment*

   The individuals of the "experiment/ treatment group" and the "control group" are randomly assigned chosen from the pool of participants. Each individual has the same probability ending up in one of the groups.

   As indicated in the recommendations above, this approach may be useful when working with large experiment groups; however, it won't work, if the experiment asks for specific units, where different roles and competences must be covered. Thus, random assignment doesn't guarantee equivalent groups in terms of competences, diversity etc. This may tackle research ethics as well.

2. *(Pair-wise) matching*

   When working with smaller groups, individual differences between the two groups play a significantly larger role, thus random assignment is not considered to be useful for setting up the groups. Instead matching may be a reasonable alternative [36] to create similar groups. This demands that a certain amount of information about the individuals in the pool of participants is available before conduction of the experiment, to allow for defining a number of matching criteria. Those criteria should take into account ethical requirements as well as the requirements defined by the tasks to cover in the experiment as well. Criteria might be weighted as well, according to the priorities of the individual experiment. Matching individuals is a mixture of using common sense and theoretical considerations, thus all relevant stakeholders of the CM solutions in the experiment (experiment lead, solution provider, end-user etc.) should be involved in the discussion of setting up the matching criteria. For further details on the process, please see e.g. Lomann (2003) [36].

In any case, the findings of the control groups should be cross-checked with additional questionnaires to be completed by the CM professionals involved to gather their feedback and judgement on the effectiveness of a new CM solutions compared to the established ones.

3.2.4 Designing and Evaluating Effective Surveys

When designing or planning a survey, the most important step is to clearly define the question to be answered by the survey results. A survey is most effective if its purpose can be clearly and distinctively stated. Surveys with vague or overly-broad motivations can become too lengthy or difficult to analyze and interpret. One of the main advantages of surveys in comparison to other data collection methods is that they allow researchers to collect a large amount of data in a relatively short period. A survey needs to be prepared carefully to not only obtain meaningful results with regard to a specific research question, but also to be able to transfer results from a sample to a larger population or stakeholder group.

Different design alternatives with regard to surveys exist and influence quality, cost and timeliness of the results. Main alternatives result from three categories: presentation of questions (e.g. written or interview), method of contacting respondents (e.g. telephone, mail, in-person) and method of recording responses (e.g., paper or electronic). Choosing the appropriate type of survey among these alternatives depends on several factors, including research objectives and timeline, sensitivity and complexity of the research question, the characteristics, abilities and resources of potential respondents and the available budget. With regard to DRIVER, surveys can help to understand different viewpoints of stakeholders when using a new solution. Especially with regard to diverging perceptions that might result from different practitioner groups (e.g., fire fighters, rescue workers), surveys could give insight into the background, origination and motivation of those groups.

Recommendations for an effective survey design comprise:

- Clear statement of objectives: define reasonable expectations that can be accomplished in a single survey
- Avoid too many objectives: limit both the type and amount of information that can be collected using a single survey
- Questionnaire items must be written from the perspective of the respondent, not the perspective of the researcher: avoid using technical terminology
- Determine the appropriate sample size based on desired precision level, confidence level and size of stakeholder group.
- Be explicit about the period of time being referenced by the question
- Save personal and demographic questions for the end of the survey
- Avoid combining multiple response dimensions in the same question
- Try not to give the impression that you are expecting a certain response

Not only the DRIVER experiment methodology is defined as an iterative approach, but also the data collection methods and applied metrics need to be elaborated and specified in an iterative process. Effective survey design requires intensive review, test, and revision of the questions.

Defining meaningful and valuable KPIs for an experiment is strongly dependant on a series of decisions that need to be made during the planning and design phase of an experiment. The benefit of a new solution, process or capability for Crisis Management can therefore only be assessed after making elaborate decisions about (a) an experiment's objectives, (b) measures to evaluate these objectives, (c) parameters to describe certain characteristics of a defined measures and finally (d) metrics to enable data collection and quantification of these characteristics. Based on the quantifications, a proposed solution, process or capability can be evaluated with regard to its performance and effectiveness within a given experiment scenario.

In this chapter existing methods and best practices when defining objectives, measures and metrics for an experiment have been reviewed and adopted to the DRIVER methodology. The table below summarizes the common problems and recommendations described in this chapter.

Common problems	Recommendations
Different understanding and prioritization of objectives	• Clearly formulate the objective of an experiment and the research questions • Decide on an overall methodology to gather evidence to address the objective and research questions. • Iteratively adapt or specify objective and research questions with stakeholders involved.
KPIs don't reflect the experiment's objective	• Clearly state the overall objective of the experiment with regard to a new solution • Clearly state how the proposed solution is expected to support a target capability/CM function • Redefine measures along the experiment process
Many different (possibly conflicting) metrics resulting in a complex experiment	• Concentrate on a few, but meaningful, measures to reflect an objective • Prioritize different objectives and measures to enable a joint evaluation result • Use methods and techniques (e.g. AHP) for collaborative decision-making in complex problem statements
KPIs not sufficient to characterize a defined function	• Application of the SMART-rule helps defining well-thought parameters
Unclear how an objective can be evaluated	• Objective need to be clearly defined to accomplish a common understanding between all involved parties (e.g., practitioners, experiment team) • An objective needs to be measurable in the context of an experiment • An appropriate data collection method needs to be assigned to each metric

Unclear definition of criteria of success	• Clearly define targeted values to be reached for success or failure
Samples don't represent target group	• Make sure that the experiment participants are representative for target user group; create different experiment groups for different stakeholders • Try to establish a control group to be able to compare results and to differentiate between various group effects
Invalidity or reliability of collected data	• Different data collection methods might be applicable to gather data for a specific parameter: Decide on an adequate method taking into account objectives, timeliness and budget of experiment • Define a minimum sample size needed for that method to be able to provide scientifically sound results • Try to set up experiment groups with regard to possible internal effects (e.g. training, fatigue) • Try to mitigate internal group effects influencing the validity of results

Table 1: Summary of anticipated problems and according recommendations in context of performance measurement in DRIVER Experiments

4 Conclusion

This deliverable provides an initial procedure to support performance measurement in SP3-4-5 experimentation. It gives a general guidance based on the CD&E as well as in the experiences learned from the first DRIVER experiments, and it is expected to further evolve during the project.

An overview of key concepts for establishing effective and homogeneous experimentation in Crisis Management is provided. We set the main foundations for an iterative and collaborative objective setting, that eventually will determine common experiment objectives and the measurements required to achieve a successful capacity building process. The main pillars for this approach are:

- Identify the Crisis Management objectives on study.
- Establish an iterative experiment development, in agreement with all involved stakeholders.
- Agree on a common set of objectives and research questions.
- Assign proper KPIs associated to the performance measurement dimensions starting from the CM objectives.
- Develop meaningful KPIs to the given measures.
- Plan continuous iterations to discuss and adapt all of these elements of a successful experiment to the requirements and needs of the experiment and the stakeholders involved.

To provide an adequate support to DRIVER experimentation, practical guidelines and recommendations are provided. The topics discussed are: quantitative and qualitative methods for data collection, creation of questionnaires and selection of groups. The next steps after this deliverable are to complement the recommendations initiated here with additional content (e.g. the evaluation of the different KPIs identified for the past DRIVER experiment). As the DRIVER Test-bed is meant to evolve in parallel to project experiments, additional support to more complex experiments will be given. At the same time, it is expected that the performance framework will benefit from additional lessons learnt from DRIVER partners.

References

[1] US Department of Homeland Security, (2009), *Target Capability List*, https://info.publicintelligence.net/TargetCapabilitiesUserGuide_17February2009.pdf, last accessed December 22nd, 2016.

[2] MC-0583 Policy for NATO Concept Development and Experimentation (CD&E).

[3] Alberts, David S., Hayes, Richard E, (2002): *Code of Best Practice: Experimentation*, CCRP Publication Series.

[4] Wiel, W.M. van der et al, (ed.), *Concept Maturity Levels Bringing structure to the CD&E process*. Proceedings I/ITSEC 2010. Interservice / Industry Training, Simulation and Education Conference, Orlando, Florida, November 29 - December 2, 2010.

[5] A. Cockburn, Agile Software Development, Addison-Wesley, 2002.

[6] Leffingwell, Dean, (2007), *Mastering the Iteration: An Agile White Paper*. Rally Software Development Corporation.

[7] Sproles, Noel (2001): The difficult problem of establishing measures of effectiveness for command and control: A systems engineering perspective. John Wiley & Sons, Ltd.

[8] Stolk, Dirk et al, (2011): *D5.1 Approaches and Solutions*, ACRIMAS project, deliverable D5.1

[9] Missoweit, M. et al, (ed.): D13.2 Milestone Report Subproject Experiment 2 Design. Deliverable of the DRIVER project, [2017].

[10] Martin, J. et al, (ed.): D42.1 – Final report on architecture design. Deliverable of the DRIVER project, [2016].

[11] Paramenter, David (2010): Key Performance Indicators (KPI): Developing, Implementing, and Using Winning KPIs. John Wiley & Sons.

[12] Kaplan, Robert, and Norton, David, (1996), *The Balanced Scorecard: Translating Strategy into Action*. Harvard Business School Press.

[13] Kovács, G., Spens, K.M. (2007): Humanitarian logistics in disaster relief operations. International Journal of Physical Distribution & Logistics Management 37(2):99-114.

[14] De Leeuw, S. (2010): *Towards a reference mission map for performance measurement in humanitarian supply chains*, IFIP International Federation for Information Processing. pp. 181-188.

[15] Santarelli, G., Abidi, H., Regattieri, A., & Klumpp, M. (2013): A performance measurement system for the evaluation of humanitarian supply chains. In POMS, 24th. Annual Conference of the Production and Operations Management Society.

[16]Drucker, Peter F. *The Practice of Management*. New York: Harper & Row, 1954.

[17]Chang, S.E., Nojima, N. (2001): Measuring post-disaster transportation system performance: the 1995 Kobe earthquake in comparative perspective. Transportation Research Part A: Policy and Practice, (6):475–494.

[18]Triantaphyllou, E., Shu, B., Sanchez, S. N., & Ray, T. (1998): Multi-criteria decision making: an operations research approach. Encyclopedia of electrical and electronics engineering, 15(1998), 175-186.

[19]Feynman, R. P. (1998), *Six easy pieces*, Penguin, London p.24

[20]Patten, Mildred L. (2014): Questionnaire Research - A Practical Guide, 4th Edition, Routledge.

[21]Kirakowski, J and Corbett, M (1988): Measuring User Satisfaction, in D M Jones and R Winder (Eds.) People and Computers IV. Cambridge: Cambridge University Press.

[22]Endsley, M. R. (1995): Measurement of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 65-84.

[23]NASA (1986): Nasa Task Load Index (TLX) v. 1.0 Manual

[24]Hart, Sandra G. (2006): NASA-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 50 (9): 904–908.

[25]Adesola, S. and Baines, T. (2005): Developing and evaluating a methodology for business process improvement, Business Process Management Journal, Vol. 11 Iss: 1, pp.37 – 46

[26]Fowler, FJ. (2002): Survey Research Methods. 3rd ed. Thousand Oaks, CA: Sage Publications.

[27]Dunn, K. 2005. Interviewing in Hay, I. (eds.) *Qualitative research methods in human geography (*79–105) Oxford University Press, Oxford, UK p.105

[28]Merton R.K., Kendall P.L. (1946) 'The Focused Interview', *American Journal of Sociology* (51): 541-557.

[29]Steyaert S., Lisor H., (eds.) 2005. *Participatory Methods Toolkit, a Practitioner's Manual*, Flemish Institute for Science and Technology Assessment, Belgium.

[30]Hennink, M.M. (2007), *International focus group research: A handbook for the health and social sciences*. Cambridge University Press: Cambridge p.6

[31]Liamputtong, P., (2011), *Focus Group Methodology. Principles and Practice.* SAGE, London p.4

[32]Gerring, Jhon, (2012). *Social Science Methodology: A Unified Framework (Strategies for Social Inquiry)*. Cambridge University Press. p.86

[33]Bergersen, S. et al, (ed.): D91.3 – Ethical Procedures, Risks and Safeguards. Deliverable of the DRIVER project, [2015].

[34]Bergersen, S. et al, (ed.): D95.21 – Planning for Ethical Approvals. Deliverable of the DRIVER project, [2017].

[35]*Guide for Understanding and Implementing Defense Experimentation (GUIDEx)*, The Technical Cooperation Program, 2006.

[36]Lomann, Tony, (2003), *Matching Procedures in Field Experiments.* Institute of Applied Research http://capacitybuilding.net/Matching%20Procedures%20in%20Field%20Experiments.pdf

[37]White, H., & S. Sabarwal (2014). Quasi-experimental Design and Methods, *Methodological Briefs: Impact Evaluation 8*, UNICEF Office of Research, Florence

[38]Labbé, P.; Bowley, D.; Comeau, P.; Edwards, R.; Hiniker, P.; Howes, G.; Kass, R.; Morris, C.; Nunes Vaz, R.; Vaughan, J. (2006) *Guide for Understanding and Implementing Defense Experimentation GUIDEx.* The Technical Cooperation Program, Canada.

[39]*Reichmann, T. (1990) Controlling mit Kennzahlen. Grundlagen einer systemgestützten Controlling-Konzeption, München.*

[40] Abrahamsson, M.; Hassel, H.; Tehler, H. (2010) *Towards a System-Oriented Framework for Analysing and Evaluating Emergency Response,* Journal of Contingencies and Crisis Management, 18, 1, pp. 14-25.

[41] Owen, C.; Brooks, B.; Bearman, C.; Curnin, S. (2016) *Values and Complexities in Assessing Strategic-Level Emergency Management Effectiveness,* Journal of Contingencies and Crisis Management, 24, 3, pp. 181-190.

[42]*Widera, A.; Hellingrath, B. (2011) Performance Measurement in Humanitarian Logistics. Proceedings of the NOFOMA conference, Harstad/Norway.*