Improving disaster response evaluations

Supporting advances in disaster risk management through the enhancement of response evaluation usefulness

Beerens, Ralf Josef Johanna

2021

Document Version:
Publisher's PDF, also known as Version of record

Improving disaster response evaluations

Supporting advances in disaster risk management through the enhancement of response evaluation usefulness

RALF JOSEF JOHANNA BEERENS
FACULTY OF ENGINEERING | LUND UNIVERSITY

Improving disaster response evaluations

Supporting advances in disaster risk management through the enhancement of response evaluation usefulness

Ralf Josef Johanna Beerens

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Engineering,
Division of Risk Management and Societal Safety, Lund University, Sweden.
To be defended at lecture hall V:B (V-building), John Ericssons väg 1, Lund,
Friday 3rd of September 2021, at 10.15 am.

Faculty opponent
Associate Professor Bjørn Ivar Kruke, University of Stavanger, Norway.

Organisation: LUND UNIVERSITY Faculty of Engineering Division of Risk and Societal Safety	**Document name:** DOCTORAL DISSERTATION **Date of issue:** 3 September 2021 **Sponsoring organization:** Instituut Fysiek Veiligheid (IFV) (ENG: Institute for Safety of the Netherlands)
Author: Ralf Josef Johanna Beerens	

Title and subtitle:
Improving disaster response evaluations: Supporting advances in disaster risk management through the enhancement of response evaluation usefulness

Abstract
Future disasters or crises are difficult to predict and therefore hard to prepare for. However, while a specific event might not have happened, it can be simulated in an exercise. The evaluation of performance during such an exercise can provide important information regarding the current state of preparedness, and used to improve the response to future events. For this to happen, evaluation products must be perceived as useful by the end user. Unfortunately, it appears that this is not the case. Both evaluations and their products are rarely used to their full extent or, in extreme cases, are regarded as paper-pushing exercises.

The first part of this research characterises current evaluation practice, both in the scientific literature and in Dutch practice, based on a scoping study, document and content analyses, and expert judgements. The findings highlight that despite a recent increase in research attention, few studies focus on disaster management exercise evaluation. It is unclear whether current evaluations achieve their purpose, or how they contribute to disaster preparedness. Both theory and practice tend to view, and present evaluations in isolation. This limited focus creates a fragmented field that lacks coherence and depth. Furthermore, most evaluation documentation fails to justify or discuss the rational underlying the selected methods, and their link to the overall purpose or context of the exercise. The process of collecting and analysing contextual, evidence-based data, and using it to reach conclusions and make recommendations lacks methodological transparency and rigour. Consequently, professionals lack reliable guidance when designing evaluations.

Therefore, the second part of this research aimed to gain an insights into what make evaluations useful, and suggest improvements. In particular, it highlights the values associated with the methodology used to record and present evaluation outcomes to end users. The notion of an 'evaluation description' is introduced to support the identification of four components that are assumed to influence the usefulness of an evaluation: its purpose, object description, analysis and conclusion. Survey experiments identified that how these elements – notably, the analysis and/ or conclusions – are documented significantly influences the usefulness of the product. Furthermore, different components are more useful depending on the purpose of the report (for learning or accountability). Crisis management professionals expect the analysis to go beyond the object of the evaluation, and focus on the broader context. They expect a rigorous evaluation to provide them with evidence-based judgements that deliver actionable conclusions and support future learning.

Overall, this research shows that the design and execution of evaluations should provide systematic, rigorous, evidence-based and actionable outcomes. It suggests some ways to manage both the process and the products of an evaluation to improve its usefulness. Finally, it underlines that it is not the evaluation itself that leads to improvement, but its use. Evaluation should, therefore, be seen as a means to an end.

Keywords:
Crisis; Disaster; Emergency; Disaster Risk Management; Preparedness; Exercise; Simulation; Response; Performance; Evaluation; Usefulness; Design; The Netherlands.

ISBN 978-91-7895-923-5 (Print) 978-91-7895-922-8 (PDF)	**Language:** English

Improving disaster response evaluations

Supporting advances in disaster risk management through the enhancement of response evaluation usefulness

Ralf Josef Johanna Beerens

LUND
UNIVERSITY

Supervisor
Professor Henrik Tehler, Lund University

Co-supervisor
Professor Nils Rosmuller, Institute for Safety of the Netherlands (IFV)

Faculty opponent
Associate Professor Bjørn Ivar Kruke, University of Stavanger

Examining committee
Professor Erna Danielsson, Mid Sweden University
Associate Professor Carl-Oscar Jonsson, Katastrofmedicinskt centrum, Region Östergötland
Associate Professor Stefan Svensson, Myndigheten för Samhällsskydd och Beredskap (MSB)

Sponsoring organisation
Institute for Safety of the Netherlands (IFV)

Cover and illustrations
Institute for Safety of the Netherlands (IFV), Broeksteeg Graphic Design

Division of Risk Management and Societal Safety, Lund University
P.O. Box 118, SE-22100 Lund, Sweden

ISBN 978-91-7895-923-5 (Print)
ISBN 978-91-7895-922-8 (PDF)

Printed in Sweden by Media-Tryck, Lund University
Lund 2021

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

To my beloved family

Table of Contents

Summary

Although the future is difficult to predict, it is possible and necessary to learn from the past. Recent emergencies, disasters and crises show us that emergency response organisations must continuously review, adjust or develop their skills, procedures and systems as this will maximise their preparedness to respond effectively and efficiently to future events.

Evaluation is a tool that supports this cyclic process. It can provide answers to stakeholders' questions, and help responders and their organisations to review, develop or even improve their preparedness. Experience from both simulated and actual events can be used to enhance future activities—but only if the product and the process are perceived as useful by the end user. Unfortunately, it appears that this may not be the case. Evaluations and their products are rarely used to their full extent or, in extreme cases, are seen as a paper-pushing exercise. In order to transform this perception, it is critical to identify what end users consider as important or useful. Therefore, the question that underlies this research is – how can evaluations (or their perception) be improved? The answer will help crisis management professionals to improve their response preparedness.

The study evolved over time as new questions were guided by findings from earlier investigations. The first part characterised the state-of-the-art, both in theory and practice, while the second part sought to gain insights into ways to enhance the usefulness of evaluations. The final strategy combined carefully-selected quantitative (survey experiments), and qualitative (document analyses and expert judgement) methods. The Dutch crisis management system was used as the basis for a case study of current practice, as this provided a coherent context. The findings are expected to be useful for individuals, teams, organisations or systems (crisis management professionals or first-responders) who are seeking to be better-prepared for future disasters.

The initial findings indicated that, despite increased academic attention, few studies have examined the topic of disaster management exercise evaluation. The current literature is limited to a specific discipline and/ or evaluation type. Both theory and practice tend to view evaluations on a case-by-case basis, creating a fragmented field that lacks coherence and depth. The lack of scientific rigour, in particular, means that professionals do not have reliable, valid guidance when designing exercise evaluations. In addition, most documentation does not justify or discuss the applicability of the selected methods, or link the overall purpose with the specific context. Furthermore, there is a lack of transparency regarding how evidence-based data is analysed, and used to reach conclusions and make recommendations. This first stage established that it is

difficult to know whether current evaluations are effective or useful, and how they contribute to disaster preparedness.

The next stage built upon the initial findings. Crisis management professionals were asked to evaluate real-world examples. Here, the aim was to investigate what aspects of evaluations influence their usefulness. The notion of an evaluation description was introduced to support the identification of four components: purpose, object description, analysis and conclusion. The results indicated that how the analysis and/or conclusions are documented influences perceived usefulness. Furthermore, different components are more-or-less useful depending on the purpose (learning or accountability). Crisis management professionals highlighted that a rigorous analysis should go beyond the object of the evaluation and take into account its context. Furthermore, they felt that it should provide them with evidence-based, actionable conclusions.

Together, these findings underline the importance of systematic, rigorous and evidence-based evaluations. They identify various issues and provide some insight into how to manage the product(s) of an evaluation in order to make it more useful. Overall, this research identified approaches that will help to ensure that the evaluation product meets its intended purpose from a user perspective. This, in turn, is likely to have a positive influence on preparedness and response. It underlines that it is not the evaluation itself that leads to improvement, but its use. Evaluation should therefore not be seen as an end in itself, but as a means to an end.

Translations of this summary in Swedish and Dutch can be found in Annex A and a visual summary is included in Figure 1.

Visual summary

Introduction

Recent emergencies, disasters and crises show us that emergency response organisations must continuously review, adjust or develop their skills, procedures and systems as this will maximise their preparedness to respond effectively and efficiently to future events.

Evaluation

Evaluation is a tool that supports this cyclic process. It can provide answers to stakeholders' questions, and help responders and their organisations to review, develop or even improve their preparedness. Experience from both simulated and actual events can be used to enhance future activities – but only if the product and the process are perceived as useful by the end user.

This Research

It is critical to identify what end users consider as important or useful. Therefore, the question that underlies this research is – how can evaluations (or their perception) be improved? The answer will help crisis management professionals to improve their response preparedness.

Research Aim: *To enhance our understanding of the role of evaluation in disaster risk management (DRM).*

Research question: *How can the usefulness of disaster response evaluations be improved with respect to their contribution to disaster risk management?*

Methodology

The study evolved over time as new questions were guided by findings from earlier investigations. The first part characterised the state-of-the-art, both in theory and practice, while the second part sought to gain insights into ways to enhance the usefulness of evaluations. The final strategy combined carefully-selected quantitative (survey experiments), and qualitative (document analyses and expert judgement) methods. The Dutch crisis management system was used as the basis for a case study of current practice, as this provided a coherent context.

Visual summary

Conclusions

Conclusion I
There is a lack of coherent, cohesive and systematic scientific attention given to building a solid knowledge base that could support best practice in the design of effective and useful evaluations for both simulated and real disaster events. This has a direct impact on the structured improvement of the DRM process.

Conclusion II
This research showed that the way evaluations of (simulated) disaster responses are documented and presented to users influences their perceived usefulness. This perception can be enhanced by the use of a user-focused, clear and rigorous approach to documentation, the presentation of the analysis, and/ or the actionability of the conclusions.

Conclusion III
There is a need to develop models, frameworks or even generic standards that contain clear components, or minimum requirements, that support evaluation designers. These tools should form the basis for the collection of evidence-based feedback on the outcomes of the operational response. This would support a cyclic connection between evaluation outcomes, preparedness training and the optimisation of resources.

Overall conclusion

Evaluation should be seen as a tool with great potential for the DRM community. As a means to an end, both theory and practice should work together to improve perceptions of usefulness. This important step forward would help to deliver and support evidence-based learning that, in turn, would help responders to learn from the past, and better-prepare for the future.

Figure 1: Visual summary of the research
This figure provides an overview of the research by illustrating its key components.

Acknowledgements

Undertaking PhD research can be compared with setting out on a new adventure; you make a plan that includes some interesting places that you would like to visit, make the necessary preparations, buy a ticket, pack your bags and off you go. As with life in general, you never really know exactly what will happen on the way, or precisely how long your journey will take, and you need to deal with the uncertainty by being flexible, resilient and persistent. Luckily you are not travelling alone, you will meet new people who are interested in what you are doing, and want to know more about your trip and past experiences; often they can also provide support as you continue your trip and reach your destination(s). Here I would like to express my gratitude to all of the people involved in my PhD journey.

I started this journey in Lund together with Jos, Kurt and Henrik. You helped me pack my bags and set off. I am very thankful that you provided me with this opportunity and supported. In particular I would like to thank Henrik for his continuous support, encouragement and valuable feedback. Whatever happened, you stood by my side and were there to help. I guess sometimes it must have been hard for you as my main supervisor, but you were always there to support me and encourage me in order to reach my destination. I also would like to thank Nils as co-supervisor from the IFV. In particular, your pragmatic support and innovative feedback were very helpful. Then there is David, not a formal supervisor, but a friendly and helpful person who I greatly respect as he is a real expert in disaster management. I am very thankful that we met during a conference in Davos some time ago and remained in touch, I was greatly encouraged by your continuous emphasis on the importance of this research.

At the beginning of my PhD journey, I was engaged in EU civil protection exercise evaluations and EU projects. I met many people from different backgrounds and interesting stories at various events. You all inspired me to focus my PhD research on evaluation. Some occasions that made a big impression on me, and my work, were the larger European exercises such as EU FloodEx and EU Taranis. From this perspective I would like to thank Piet, Peter, and all the organisations and evaluators that I have worked with, and all the responders who let me evaluate them. In particular I would like to thank Erie and Phil who, from then on, were involved in my journey.

There were also some locations that could be seen as ‘a home away from home’. First there is the IFV, my basecamp. Although I was supported by many colleagues from different departments, I should highlight the Research Department and the Crisis Management Academy. You acted as a listening ear or an expert, but when needed also as a valued, critical friend. In particular I would like to mention Saskia, Jolanda, Ron,

Clemon, Martina, Marije, Anne, Edith, Marian, Coen, Karin, Veerle, Annemieke, Wim, and Herman. But of course, there are many others.

Then there is the Division of Risk Management and Societal Safety, another basecamp, where I met fellow travellers but also guides. I would like to thank all of my colleagues in the Division for exchanging traditions such as the Swedish 'Fika' and the Dutch 'stroopwafel', that match surprisingly well. But moreover, I would like to thank you for the interesting discussions that we had. In particular I would like to mention Hanna, Peter, Tove, Björn, Linn, Christian, Per, Alex, Henrik, Magnus, Jenny, Olof, Johan, Johanna, Ann, Misse, Mo, and Marcus. It was a pity that due to Covid-19, we could not physically meet in the final stages.

In addition to meeting people at various places or occasions I also met many others on the way. A random few I would like to mention are Ben, Bert, Teun, Joost, Tijs, Roy, Rob, Rixt, Paul, Jan, Jochen, Michel, Marcel, Arjen, Marlous, Bas, Ruben, Ira, Jonas, Ellen, and Elaine. All these people supported me in various ways. I acknowledge that many others will go unmentioned, for example, all the participants in the consultation exercises or the survey experiment, my students and fellow lecturers of the Master's degree in Crisis and Public Order Management and the experts from various safety regions who participated in the informal platform on crisis management evaluation. However, I would like you to know that you have not been forgotten as you had an impact, large or small, known or unknown, during this journey.

But this journey was impossible without the support of my parents, family and friends 'back home'. Basically, one person made this all possible and that is my wife Lotte who always had my back, not only as a wife, but also as my best friend. Without your love, humour, and support it would never have been possible, and without mentioning all the details, there is one relevant thing that I would like to point out in this 'travel context' and that is your support in overcoming my fear of flying.

As this journey came to an end and worldwide travel became impossible due to Covid-19, a new adventurer was born, our son Reinier. I hope that one day we can travel together so that I can show you the world, and tell you all the stories about this, and other journeys when you grow up, and teach you the valuable lessons that I have learned through performing my own evaluations.

Apeldoorn, June 2021

Ralf

Thank you very much - Tack så mycket - Dankjewel!

Appended papers

I. **Beerens, R.J.J.**, & Tehler, H., (2016). Scoping the field of disaster exercise evaluation - A literature overview and analysis. *International Journal of Disaster Risk Reduction,* 19, 413–446. https://doi.org/10.1016/j.ijdrr.2016.09.001.

II. **Beerens, R.J.J.** (2019). Does the means achieve an end? A document analysis providing an overview of emergency and crisis management evaluation practice in the Netherlands. *International Journal of Emergency Management,* 15 (3), 221–254. https://doi.org/10.1504/IJEM.2019.102310.

III. **Beerens, R.J.J.**, Tehler, H., & Pelzer, B. (2020). How can we make crisis management evaluations more useful? An empirical study of Dutch evaluation descriptions. *International Journal of Disaster Risk Science,* 11, 578–591. https://doi.org/10.1007/s13753-020-00286-7.

IV. **Beerens, R.J.J.**, & Haverhoek-Mieremet, K. (2021). What do practitioners expect from an evaluation report? A qualitative analysis of Dutch crisis management professionals' expectations. *International Journal of Emergency Services,* 10 (1), 1–25. https://doi.org/10.1108/IJES-12-2019-0063.

The above-mentioned publications can be found in Annex E.

Related publications

Verheul, M.L.M.I., Dückers, M.L.A., Visser, B.B., **Beerens, R.J.J.**, & Bierens, J.J.L.M. (2018). Disaster exercises to prepare hospitals for Mass-Casualty Incidents: Does it contribute to preparedness or is it ritualism? *Prehospital and Disaster Medicine*, 33 (4), 387–393. https://doi.org/10.1017/S1049023X18000584.

Beerens, R.J.J., Abraham, P.J., Glerum, P. & Kolen, B. (2014). Flood Preparedness Training and Exercises. In: J.L.M. Bierens (Ed.), Handbook on Drowning (2nd edition) (pp. 1009-1016). Springer-Verlag. https://doi.org/10.1007/978-3-642-04253-9_154.

Beerens, R.J.J., Abraham P. & Braakhekke, E. (2012). Maximise your returns in Crisis Management preparedness: A Cyclic Approach to training and exercises. Davos. International Disaster Risk Conference (IDRC). [Conference Proceedings]. https://doi.org/10.13140/2.1.1138.0482.

Jongejan, R.B., Helsloot, I., **Beerens, R.J.J.** & Vrijling, J.K. (2011). How prepared is prepared enough? The Journal of Disaster Studies, Policy and Management, 35 (1), 130–142. https://doi.org/10.1111/j.1467-7717.2010.01196.x.

Beerens, R.J.J., Kolen, B., & Helsloot, I. (2010). 'EU FloodEx 2009: An analysis of testing international assistance during a worst credible flood scenario in the North Sea Area' in FRIAR 2010 Conference Proceedings, United Kingdom: University of Wessex. https://doi.org/10.2495/FRIAR100211.

Abstracts of the above-mentioned publications can be found in Annex B.

Various grey literature publications, such as European exercise evaluation reports were produced in the early stages of this research, for example EU FloodEx (2009), EU ModEx (2010–2011), EU WaterSave (2012), EU Taranis (2013) and MIRG-EX (2016). Furthermore, a range of presentations and workshops related to the topic were delivered.

Contribution statement

Conducting research is a mix of individual and collaborative efforts. With the exception of Paper II, all of the papers included in this thesis were co-authored with other academics and professionals. The author's individual contributions are described below.

Paper I: *Scoping the field of disaster exercise evaluation – A literature overview and analysis.*

This paper mapped the disaster exercise evaluation literature in order to identify key concepts, gaps in the research, and types and sources of evidence to inform further practice and research.

Contributions: First of two authors with overall responsibility for the research effort. Both authors formulated the aims of the research and developed the paper's structure. I designed and conducted the research and analysed all of the empirical data. The co-author (Prof. Henrik Tehler) cross-checked the data and we jointly wrote the paper. During this writing process I was mainly responsible for the background, method, and results and analysis sections. Both authors reviewed the paper prior to publication.

Paper II: *Does the means achieve an end? A document analysis providing an overview of emergency and crisis management evaluation practice in the Netherlands.*

This paper investigated how crisis management evaluations are performed and reported in practice in the Netherlands.

Contribution: Sole author. This research was solely performed by the author.

Paper III: *How can we make crisis management evaluations more useful? An empirical study of Dutch evaluation descriptions.*

This paper introduced the theoretical concept of the 'evaluation description'. This was used as a basis for survey experiments that identified what makes an evaluation useful for professionals.

Contribution: First of three authors with overall responsibility for the research effort. Together with the first co-author (Prof. Henrik Tehler), I formulated the aim of the research and developed the concept of an evaluation description. I designed and conducted the research, and analysed all of the empirical data. The second co-author (Ben Pelzer) cross-checked quantitative analyses. The first two authors wrote the paper, and the second co-author contributed with constructive reviews. I was mainly responsible for writing the theoretical background, method, measurements and data, results and analysis sections. All authors reviewed the paper prior to publication.

Paper IV: *What do practitioners expect from an evaluation report? A qualitative analysis of Dutch crisis management professionals' expectations.*

This paper builds upon qualitative data that was obtained during the survey experiments that were also used for Paper III. It used qualitative analysis tools and techniques to identify expectations of evaluation reports among two groups of users.

Contribution: First of two authors with overall responsibility for the research effort. In particular, I was responsible for the design and data collection instrument. Together with my co-author (Karin Haverhoek), I analysed the empirical data. I was mainly responsible for writing the theoretical background, method, results and discussion sections. Both authors reviewed the paper prior to publication.

1 Introduction

1.1 Background and rationale

> 'Learn from the past, prepare for the future, and perform in the moment (Van Hoozer, 2008, p. xiii).'

Although we cannot predict the future, it is possible to look back and learn from the past. In the past decade (2010–2019) alone, there were, on average, approximately 343 disasters[1] worldwide, per year (Guha-Sapir, 2020). Hundreds of thousands of people have lost their lives, and millions have been displaced. Material and infrastructure damage has cost society billions of dollars, and specialist first responder organisations have had to be deployed to mitigate their effects.

The decade began with earthquakes[2] (e.g. Haiti 2010, Nepal 2015), a tsunami and subsequent nuclear disaster (Japan 2011), hurricanes and typhoons, for example in the Americas (Sandy 2012), Asia (Haiyan 2013), and the Caribbean (Maria 2017). It ended with fires that swept through countries such as Greece (2018), Brazil/ the Amazon (2019) and Australia (2019/2020). Although in many cases, their effects were limited to regional or national boundaries, some were of a scale or severity that led to them being designated as transboundary or creeping crises (Ansell et al., 2010; Boin et al., 2020). The most recent example of such a crisis comes from the last year of the decade, which saw the emergence of a new respiratory illness (Covid-19). While initially an isolated outbreak, it went on to grow exponentially to become a global pandemic.

In addition to looking back in order to learn, it is also reasonable to anticipate that the consequences of adverse or disastrous events may evolve, and that their shape and dynamics might change (Ansell et al., 2010; Boin, 2009; Boin et al., 2020). As systems

[1] To qualify as a disaster, at least one of the following criteria must be fulfilled: (a) ten or more people reported killed; (b) 100 or more people reported affected; (c) the declaration of a state of emergency; or (d) a call for international assistance.

[2] It should be noted that these are examples of large-scale disasters that gained worldwide attention, and affected entire regions or nations. Fortunately, such events are infrequent. However, there are far more smaller-scale events, such as transport incidents, toxic spills or small fires that also have serious consequences for the people who are affected.

and organisations become more complex and tightly-coupled due to, for example, globalisation, urbanisation and technological advances (Perrow, 1994, 1999) the impact of events on humanity will increase. The skillsets required by emergency response organisations are also likely to change. Good preparation is, therefore, vital. Examples include exercises that simulate conceivable future events. These activities ensure that organisations are able to respond efficiently and effectively to a crisis, and can manage both current and future events at all levels of complexity. In this context, it is important to continuously evaluate performance and, subsequently, adjust the skills and procedures that are used. In cases where there are existing benchmarks or agreements, there is also an opportunity to map performance against targets to review accountability.

The importance of being prepared is acknowledged in the Sendai Framework for Disaster Risk Reduction 2015–2030 (United Nations, 2015), and the interdependency between preparedness and response is illustrated in the disaster risk management cycle (see section 2.2.1). The implication is that, for example, investments in preparedness can lead to a significant return in the response phase. Preparedness encompasses a wide variety of activities, such as drawing up policies and plans, conducting exercises, developing social networks, and identifying and implementing technological solutions. Ideally, these activities should form a sequential and iterative process that is guided by a needs analysis. The outcomes of this investment, if focused correctly, are systems, organisations, teams and individuals who are better prepared for the next disaster, and able to respond more effectively and efficiently.

Preparedness activities are a very important element in disaster management (Perry, 2004; Sinclair et al., 2012; United Nations, 2010) and exercises are conducted frequently. Nevertheless, few efforts have been made to exploit the potential of exercise-based research, in order to produce generalisable emergency preparedness practices (Hunter et al., 2012). Exercises are the primary experiential means by which both professionals and researchers can train for, or test, a broad range of responses (planning, procedures, skills and knowledge) in a safe, but realistic environment (Alexander, 2002; Borodzicz & Van Haperen, 2002; Gebbie et al., 2006; Payne, 1999; Peterson & Perry, 1999; Savoia et al., 2013; Wybo, 2008).

Despite their importance, exercises are often seen as resource-intensive. Moreover, little is known about their cost, and to what extent they achieve their purpose. Thus, it is not easy to assess their added value (Hsu et al., 2004). Rough calculations suggest that

costs vary from €5,000–10,000[3] for small-scale exercises, to approximately €1,000,000[4] for one large-scale European exercise; it is reasonable to expect that these costs will continue to rise. Simultaneously, the infrequent nature of large-scale emergencies, disasters and crises can be seen as a barrier to preparedness evaluation. Given the lack of real-life events, researchers find it difficult to test hypotheses and identify predictors of effective response outcomes (Hunter et al., 2012). While Wildavsky (1988) argued that there is ample evidence to suggest that we are safer today than we have ever been before, and it can be argued that our response capacity is continuously developing, it is difficult to measure how safe we really are. It is even more difficult to know whether response organisations are really prepared for the next, unforeseen adverse event (Boin, 2009).

In this context, evaluation is an important tool, as it allows both researchers and professionals to gain insights into not only the effectiveness of exercises, but also the effectiveness of preparedness and response activities. It can provide answers to a question such as *does current practice work?* It can help responders at all levels to be better prepared, by answering questions such as *what is needed for it to work?* Heath (1998) notes that the evaluation of an actual or simulated crisis, or training exercise, is probably the most important way to improve disaster management and reduce the loss of lives and resources. In a similar vein, Kirschenbaum (2003) illustrates the role of evaluation:

> When humans were still wanderers, our small communities moved to better hunting or grazing whenever the need arose. With settlement came town development and the oldest types of 'first responders', volunteer firefighters, who were, mainly, just neighbours assisting each other (Kirschenbaum, 2003, p. 2).

Over time, each new disaster has brought with it creative forms of management that were evaluated and, eventually, incorporated into the community. This cyclic process is needed because development is challenging if there is no review or evaluation. An evaluation can identify positive behaviours and processes that should be continued, and less useful practices that should be revised or dropped. Building upon past experience enhances future activities. Exercises, and their evaluation, can also be used to gain

[3] Costs for small-scale exercises are based on personal experience.

[4] Costs for large-scale European exercises are based on financial information provided by the European Union's Civil Protection Mechanism, in its *Technical guidance for UCPM full-scale exercises*, see: https://ec.europa.eu/research/participants/data/ref/other_eu_prog/ucpm/guide/pse/ucpm-guide-practical-exercise_en.pdf [accessed 10 December 2020]. In addition, a list of EU-supported civil protection exercises and their budgets is published online at: https://ec.europa.eu/echo/funding-evaluations/financing-civil-protection/civil-protection-exercises_en [accessed 10 December 2020].

insights into, for example, the operational readiness of emergency response organisations. Ideally, they provide systematic support for the direction of, and investment in, future learning and development (Abrahamsson et al., 2010; Alexander, 2015; Borell & Eriksson, 2008; Jongejan et al., 2011; Ritchie & MacDonald, 2010).

Currently, however, several issues limit the use of evaluation in crisis management. Firstly, disaster preparedness practices are largely based on anecdotal evidence, and lack systematic study or objective validation (T. L. Thomas et al., 2005). Crisis management practice has been accused of producing untested, fantasy documents related to various areas of planning and evaluation (Birkland, 2009; A. A. Bowen, 2008; Clarke, 1999; Hutchinson et al., 2018; McConnell & Drennan, 2006; Sinclair et al., 2012). Secondly, many emergency management practices are not validated because of a lack of materials to assess their performance and provide empirical feedback to participants (Biddinger et al., 2008). Thirdly, it is widely believed that evaluations are simply put into drawers, or lie on shelves gathering dust. Too often, they are seen as paper-pushing activities.

Kirschenbaum (2003) highlights that statements regarding the need for better coordination and communication appear repeatedly, but actual implementation comes down to personal experience. Similarly, evaluation of the effectiveness of the disaster response, and related training exercises, has been given little consideration from a scientific perspective. Consequently, there is no comprehensive overview of research into the evaluation of the operational response during disaster management exercises. This leads to a lack of clarity regarding the contribution of theory to practice.

Evaluations can provide evidence-based recommendations that might help organisations make their disaster management and response methodologies more effective. However, this requires a structured approach. The evaluation should provide insights into the functionality of the system, and identify any lessons to be learned that can be incorporated into future preparedness or response activities (A. A. Bowen, 2008). Learning is optimised when information is presented in a way that users find useful. *Usefulness* can be defined as the extent to which an evaluation achieves its intended purpose, as defined or perceived by the user. If the purpose is to support learning, then it is related to the extent to which it helps actors to learn from the exercise. If it is to support accountability, then it is related to the extent to which it helps actors to justify their actions.

Evaluations should not be seen as the holy grail for improvement (i.e. if we execute and use them, they will deliver effective solutions to all problems). In reality, this is not possible. The evaluation is only one of many tools or factors. Individual shortcomings or organisational restraints can hinder the process, and even the most incisive evaluation

will not lead to improvement if the identified outcomes are not implemented. In some cases, it merely serves a symbolic purpose: evaluation for the sake of evaluation. So, the question remains, to what extent do evaluations fulfil their purpose? What are the weaknesses? And why? Finally, an equally important question is how can we improve evaluations so the they are more readily perceived as making a useful contribution to preparedness or response?

1.2 Aim, objectives and focus

The overall aim of this research project is to enhance our understanding of the role of evaluation in disaster risk management (DRM). The primary focus is to evaluate operational responses, and use the product of this evaluation to identify activities that improve DRM preparedness and response, by supporting: i) learning; and ii) accountability.

As exercises are held regularly, this research primarily examines their evaluation. These simulated events provide a realistic experience for the operator, but are delivered in a controlled and relatively safe environment. Since the evaluation of real events shares many similarities, it is reasonable to expect that findings from a simulated exercise are relevant to real life. The analysis is focused on the evaluation product – the report – and its usefulness. When reports from real events are used, this is made explicit in the text.

The study aims to provide knowledge that will support further theoretical study of evaluations in the context of DRM. It also seeks to provide practical insights and guidance for professionals who develop and use evaluations. The overall aim is divided into the specific objectives outlined below:

- to map the scientific literature on disaster exercise evaluation in order to identify:
    - key concepts;
    - gaps in the literature; and
    - types and sources of evidence;
- to increase knowledge of how operational response evaluations are performed and reported in practice, and whether they actually meet their intended purpose;

- to identify which factors (components) of operational response evaluations influence their usefulness for crisis management professionals (users) in operational and supervisory positions, and support learning and accountability;
- to provide guidance that improves the usefulness of evaluations in emergency, disaster and crisis management practice;
- to inform professionals, policymakers and researchers about past, present and (possible) future advances with regard to operational response evaluations.

1.3 Research questions

In order to achieve the objectives described above, the following overarching research question was developed:

How can the usefulness of disaster response evaluations be improved with respect to their contribution to disaster risk management?

This overall question is broken down into the following four sub-questions that more precisely describe how the research is structured:

RQ 1 (paper 1):

What is known about the evaluation of disaster management exercises in scientific literature?

Conducting research requires standing on the shoulders of giants. The first step was, therefore, to map the scientific literature using a scoping methodology. The aim was to identify key contributions, and provide an overview of existing research. This phase also identified avenues for future research. The results highlighted that it is unclear how much progress has been made, and how evaluation practices have been implemented. The next step was, therefore, to investigate how evaluations of (simulated) emergencies are performed in practice, and how they are documented. This step led to the following research question:

RQ 2 (paper 2):

How are disaster management exercise and real-life response evaluations documented in the Netherlands?

RQ2 extends RQ1. It provides a comprehensive overview of the current state of practice in the Netherlands. A document analysis provided evidence of how multi-organisational emergency exercise and real-life response evaluations are designed,

implemented and documented. It was also important to identify how the evaluation is defined, and what aspects are deemed important. This study highlighted that although a range of reports are produced, it is unclear how they are used or whether they achieve their purpose. These outcomes were to be addressed by the third research question, which focuses on the usefulness of evaluations for crisis management professionals:

RQ 3:

What makes evaluations (descriptions/texts) more or less useful to professionals?

RQ3 assumes that a useful evaluation produces a product that can be used for either learning or accountability purposes. This RQ builds on the findings of the previous RQ, which identified a commonly used format. This format is referred to as the *evaluation description*, and it contains four elements: Purpose (P), Object description (O), Analysis (A), and Conclusions (C). These elements capture how evaluations communicate both the process and its findings and, thus, how they contribute to achieving a purpose.

The evaluation should be seen as a means to an end, and not an end in itself. It should be both useful and used. Usefulness can relate to: a) the information users want; and b) expectations. Thus, RQ3 is divided into two sub-questions. RQ3a investigates how O, A and C influence perceived usefulness for learning or accountability purposes (P), and is formulated as follows:

A. (paper 3) *(How) does the clarity of the presentation of the object (O), the analysis (A) and/ or the conclusion (C) in an evaluation description influence its perceived usefulness (P) for the purposes of: (i) learning and (ii) accountability?*

Here, the evaluation description is used to investigate usefulness in an experimental setting. The clarity (clear/ unclear) of O, A and C were manipulated to explore their effect on P. This experiment also provided qualitative empirical data regarding users' expectations. RQ3b was added to obtain more detailed data:

B. (paper 4) *What do crisis management professionals expect to find in a useful crisis management evaluation report?*

RQ3b assumes that a useful evaluation meets the expectations of users and contributes to achieving the higher-level purposes of learning or accountability. A thematic analysis of written quotes identified common expectations that could be used as a basis for describing needs.

Finally, the fourth RQ combines theoretical insights with practical implications, and seeks to bridge the gap between theory and practice. It builds upon the theoretical

insights gained from the previous RQs, and provides guidance on how evaluations can be improved:

RQ 4 (this thesis, mainly chapters 6 & 7):

How should we design evaluations of simulated or real disaster responses (including the product) in order to make them useful and relevant to a variety of users?

It should be noted that RQ4 does not, and cannot have an unambiguous answer, due to the wide-ranging and fluid nature of crises and their simulations. It does, however, identify a range of factors that can form the foundations for an evaluation framework, either for a specific situation or for a generic policy. Therefore, this question is mainly addressed in the Discussion (Chapter 6). The latter section offers a tentative solution, by integrating knowledge from this research and other studies. Answering this question is a first step in anchoring evaluation in the broader context of DRM.

Figure 2 illustrates the abovementioned research aim and questions and their relationships. It provides a visual overview of the connections between papers and RQ's.

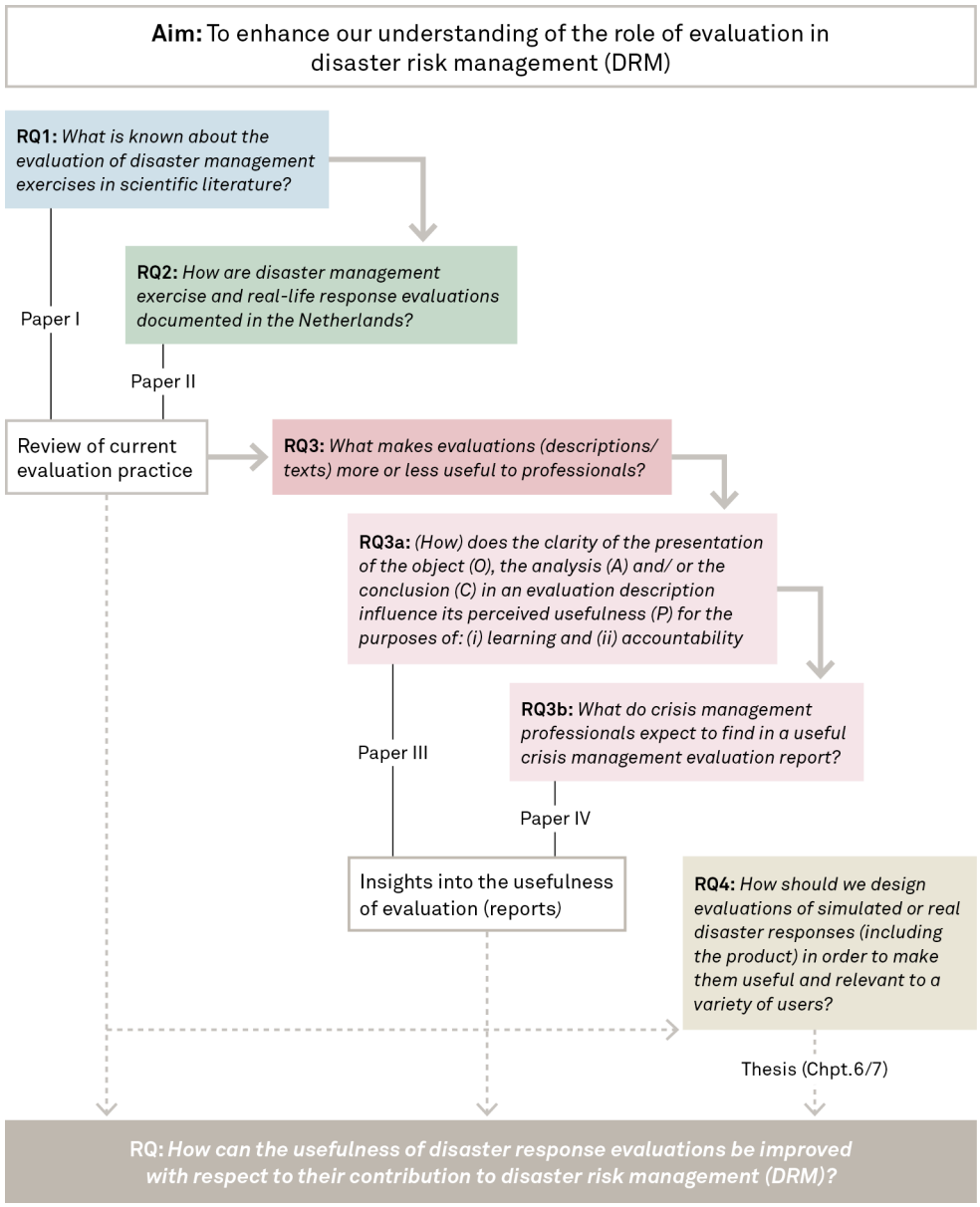


Figure 2: Research overview

The figure presents a schematic outline of how the four papers contribute to answering their respective research questions (RQs) and how they then contribute to answering the overall research question, thereby achieving the aim of this research. It shows that the research consists of two blocks: I) a review of current pratice; and II) an experiment that helped to gain insights into the usefulnness of evaluations and their reports. Although the RQs and papers that are connected to a block are independent, they inspire and built upon each other and, therefore, are closely related. In addition, it shows that answering RQ4 builts upon these blocks and is addressed in this thesis.

1.4 Geographical focus

Most of the practical research was conducted in the Netherlands. There were three reasons for this choice.

Firstly, the selection of one country provided an overview of response evaluations in the context of an overall crisis management system. A comparison with local evaluations from different countries or other systems would be misleading, as there are likely to be cultural, political, organisational or systematic differences. The focus on the Dutch system created a common context that mitigated such problems. More details are provided in section 3.3.

Secondly, and closely related to the relationship between (New) Public Management and crisis management, is the prevalent evaluation culture. The Netherlands has a long history of investigations into the causes of disasters and accidents. A notable example is the transport sector, where historical investigatory committees have formed the basis for the creation of a national, independent Safety Board. Moreover, recent emergencies, with substantial social and political impacts, have been thoroughly evaluated by other bodies, such as independent research agencies and inspectorates. These investigations are in most cases a legal requirement; thus, evaluations are well-established and there is a wide selection available for review.

Thirdly, this research project was funded by the national Institute for Safety (IFV), which has a particular interest in the Dutch context. Its mission is to strengthen the country's Safety Regions and their partners, in terms of professionalisation. It develops and shares relevant knowledge, has expertise in acquiring and managing communal equipment, and supports local authorities and councils. The topic of this research is highly relevant to this mission, and the Institute provided access to data from practice via its networks.

1.5 Thesis outline

This thesis provides a synthesis of research outcomes, and outlines how the project developed, in terms of theoretical and methodological considerations. This first chapter established the rationale for conducting research in this area, and set out the purpose and the main research questions. Chapter 2 outlines the theoretical background. This chapter introduces the key concepts and terminology. Chapter 3 provides additional, practical background information with regard to the context. Chapter 4 outlines the research design; in particular it clarifies how the overall research question was broken

down and systematically investigated. It sets out which research questions were addressed, and how. This chapter also reflects on the scientific quality of the research. Chapter 5 presents the key findings and contributions. This chapter is complemented by the four papers included in Annex E. The findings are then synthesised and discussed in Chapter 6. Chapter 6 also introduces various models that strengthen their conceptual foundations. It implicitly addresses the main research questions, in particular RQ4. Following this discussion, Chapter 7 proposes some ideas for future work in both research and practice. Finally, the overall conclusions are presented in Chapter 8.

2 Theoretical background

This section provides an overview of the key theoretical concepts that are central to this thesis. First, it defines *risk* and *disaster risk management*, before describing the core concept of this thesis: *evaluation*. It creates a common theoretical point of departure for the subsequent discussion of disaster exercises and evaluation practice presented in the next chapter.

2.1 Risk

The notion of risk is fundamental to this work. Although the focus of this research is not on risk *per se*, the concept helps to understand key elements such as events, uncertainties and consequences that influence evaluation, and *vice versa*.

2.1.1 What is risk?

Risk can be defined in various ways (see e.g. Vlek, 1996), as it is used in disciplines or contexts that encompass engineering, economy or sociology. Consequently, there is no single understanding (Aven, 2012; Aven & Renn, 2009; Haimes, 2009; Van Asselt & Renn, 2011). However, from an ontological perspective, it is possible to distinguish three categories (Aven et al., 2011): (1) as a concept that describes events, consequences and uncertainties; (2) as a modelled, quantitative concept (reflecting random uncertainties); and (3) as risk measurement. The latter concepts (2 and 3) are often viewed as narrower, technical definitions in which the probability of a hazard that causes a certain harm is rationally calculated. However, this approach is difficult to apply in some situations, notably events that have not yet happened. Thus, 'uncertainty' is seen as a broader notion than 'probability'.

This research does not focus on risk *per se*. Instead, it uses it to provide context. Therefore, it is better to view it from the broader perspective suggested by (1). Here, risk is defined as "uncertainty about and severity of the consequences (or outcomes) of an activity with respect to something that humans value" (Aven & Renn, 2009, p. 2).

The definition highlights the three building blocks proposed by Aven (2010): (A) events/ scenarios, (C) consequences, and (U) uncertainties. Risk = (A, C, U). The definition also emphasises that risk is related to something that humans value. Examples include life, health or property, which are threatened by an event, leading to unwanted consequences.

In order to reduce these consequences, and protect what is valued, people make preparations to be able to respond effectively. However, uncertainties surrounding the level of risk and, therefore, the scale of any consequences, make it difficult to predict with any degree of accuracy when, how and what will, or is likely to happen. It is therefore important to put in place a range of preparatory actions in order to deliver a response that is flexible and can be scaled to meet the threat. Moreover, if possible, these actions should be tested and evaluated.

The above definition reflects the 'new' risk perspective, which distinguishes between the concept of risk and its description, notably, the concept and the results of the risk assessment (Aven, 2010, 2012). It implies that the risk description includes a presentation of consequences, a measure of uncertainty, and the background knowledge that the uncertainty measure is based on (Aven, 2012). The approach is relevant to this research as these elements make it possible to evaluate risks.

2.1.2 How can consequences and events be defined? Emergencies, disasters and crises

This research examines operational response evaluations, and how the product can identify and support activities that improve DRM preparedness, and limit the consequences of an adverse event.

But how are consequences defined? In the risk context, they can be seen as the adverse effects of an event or activity (Aven, 2011). But how do we define the events that we are focussing on? Here, the focus is on emergencies, disasters or crises. Table 1 highlights that although these terms can mean different things to different people in different cultures, they can all be seen as a distressful situation in which (a series of unwanted) events have, or can have, very negative consequences for human beings, societal functions or fundamental values (Uhr, 2009). It should also be noted that these terms may refer to a unique, non-routine or rare event, with a high degree of uncertainty that has no precedent in an organisation's history or policies. Not least because, if it did occur regularly, it would be considered as a routine incident or accident (Deverell, 2012).

Table 1: Comparing emergencies, disasters and crises. All are examples of distressful situations in which series of events have, or can have, very negative consequences for human beings, societal functions or fundamental values (Uhr, 2009).

TERM	DEFINITION
Emergency	Unforeseen but predictable, narrow-scope incidents that regularly occur. Can also refer to a future event that is expected to cause significant damage and disruption (Perry & Lindell, 2007).
Disaster	Sudden onset events that seriously disrupt social routines, lead to the adoption of unplanned actions to adjust to the disruption, are delimited in social space and time, and endanger valued social objects. Disasters are more rare than emergencies and are defined by human casualties, property damage, and severe social disruption (Perry & Lindell, 2007).
Crisis	A situation that is perceived to threaten the core values or life-sustaining functions of a social system, and which requires urgent remedial action under conditions of deep uncertainty. Crises affect multiple jurisdictions, undermine the functioning of various policy sectors and critical infrastructures, escalate rapidly and morph as they unfold. In a crisis, past experience provides policymakers with little guidance (Ansell et al., 2010).

As this research does not focus on the situation (emergency, disaster or crisis) *per se*, but on the response, the distinctions given in Table 1 are used to illustrate that severity and uncertainty play an important role in all cases. The evaluation should provide a detailed characterisation, operationalisation or description of the situation. Therefore, the terms emergency, disaster and crisis are used interchangeably throughout this thesis, and only specified when needed. In most cases, the word 'disaster' will be used, except in the Netherlands context, where the word 'crisis' is more commonly used (see also section 3.3).

2.2 Disaster risk management

The previous sections suggest that risk can be reduced by either limiting how often adverse events lead to unwanted consequences, or by limiting the consequences of events. In both cases, uncertainty is an important factor. A systematic approach is used to deal with this combination of events, uncertainty and negative consequences, referred to as disaster risk management (DRM).

2.2.1 What is DRM?

DRM can be broadly described as the implementation of a process or approach that aims to mitigate risks. More specifically, it is defined here as "the application of disaster risk reduction policies and strategies to prevent new disaster risk, reduce existing disaster risk and manage residual risk, contributing to the strengthening of resilience and reduction of disaster losses" (United Nations Office for Disaster Risk Reduction (UNDRR), 2021, para. 1).

Like risk and its outcomes, DRM can be described in various ways. For example, it can be seen as a *process* that addresses the likelihood and consequences of risks (the approach taken in ISO 31000), or it can be seen as a *cycle* that addresses consequences. The DRM cycle has been seen as a crucial instrument for the management of disasters and their effects since the 1970s. Coetzee and Van Niekerk (2012) state that the idea illustrates the ongoing process by which governments, businesses and civil society plan for, and reduce the impact of disasters, plan the response during and immediately following a disaster, and take steps to recover after a disaster has occurred. This description highlights that organisations must anticipate and prepare, in order to respond effectively and efficiently to the consequences of disastrous events (Tierney et al., 2001).

Coetzee and Van Niekerk (2012) demonstrate that descriptive, linear models of disaster phases have evolved into normative models for their management; cycles are used as a tool to manage disasters and their consequences. Once again, the present research does not specifically focus on DRM, but the DRM framework provides an overall structure that guides the process and limits the focus. More specifically, a simplified model is used to identify concepts that are logically connected, and that form the context for risk management functions. Table 2 provides an overview of the typical functions found in the process, which are often referred to as stages in the DRM cycle.

Table 2: Functions in the disaster risk management cycle.

FUNCTION	DESCRIPTION
Mitigation (& prevention)	Disaster prevention and loss reduction activities that try to eliminate the causes of a disaster. This is done either by reducing the likelihood of occurrence, or limiting the magnitude of any negative effects. The aim is to prevent the event before it happens and reduce the impact of future disasters (Alexander, 2002; McEntire, 2007; Perry & Lindell, 2007).
Preparedness	Refers to efforts to increase disaster readiness. Activities aim to protect lives and property when (forecasted or imminent) threats cannot be controlled, or when only partial protection can be provided. They can be divided into two categories: (1) alerting members of response organisations and members of the public to the timing and extent of a potential disaster; and (2) actions designed to enhance the effectiveness of the response (Alexander, 2002; McEntire, 2007; Perry & Lindell, 2007).
Response	Refers to attempts to limit damage from the initial impact, which can be performed just before or during the disaster impact (Perry & Lindell, 2007). It can be seen as a group of individuals – perhaps specialists or experts, but often line managers or subordinates – who come together (in the immediate aftermath) of a critical situation to protect life and property (Borodzicz & Van Haperen, 2002; McEntire, 2007). The basic goal of a response organisation is to minimise the impact of the disaster.
Recovery	Recovery begins after the disaster impact has been stabilised and includes activities that aim to return the affected community to its pre-disaster or, preferably, improved state by restoring lost functions. It can be divided into short-range (relief and rehabilitation) and longer-range (reconstruction) measures (McEntire, 2007; Perry & Lindell, 2007).

The original cycle concept referred to the temporal stages of a disaster (pre-disaster, disaster and post-disaster), but many variations have emerged over the past 50 years (Coetzee & Van Niekerk, 2012). Table 2 presents some commonly-used terms: mitigation (and prevention), preparedness, response and recovery. It should be noted

that these phases or functions do not always, or even generally occur in isolation, or in this order. For example, they can occur simultaneously, as in the current Covid-19 pandemic, where mitigation, preparedness and response are all taking place during the disaster. Moreover, they are indistinct, as there is no clear beginning or end to each phase (Perry & Lindell, 2007). Within this cycle preparedness is a vital, continuous and innovative element as it links preventive measures and the response. These phases often overlap, and their length depends on the severity of the disaster (Khan et al., 2008). Therefore, it is better to view the cycle from the functional perspective of what is done in each phase, as this supports the identification of its tasks, capacities and capabilities.

2.2.2 Controlling risks and managing disasters: Cycles and loops

DRM can be seen as providing an overall structure for actions that are intended to lessen the adverse impacts of hazards, and a possible disaster (UN Secretary-General, 2016; United Nations International Strategy for Disaster Reduction (UNISDR), 2009). It is important to note that 'adverse impacts' can mean many things, and understandings can differ from person to person. A related term is 'severity', which is part of the definition of risk used here. The latter definition (see section 2.1.1) makes it clear that unwanted events can be described in terms of their severity. However, how severe a specific event is judged to be depends on who you are; therefore any effort to manage risk assumes that there is broad agreement regarding how severity is measured. Usually, this is not a problem. Many would agree, for example, that the more people who die due to a disastrous event, the more severe the event. But in other cases such agreement might be more difficult to achieve.

Assuming that there is agreement on what DRM is supposed to protect, and how adverse impacts are interpreted, it can be approached as a so-called 'control problem' (Brehmer, 1992; Rasmussen, 1997). From this perspective, the aim of risk management is to try to gain control, which is similar to trying "…to achieve some desired state of affairs" (Brehmer, 1992). Gaining control or achieving something also implies a continuous, rather than a one-time process. Thus, risk control refers to ongoing efforts to try to protect something that is valued (e.g. human lives) in the face of uncertainty: when a disaster strikes, the response system tries to minimise losses as a function of what is valued. When the response phase is over, it adjusts to the event that just happened (and the lessons that have been learned) to ensure that what is valued will be better-protected in the future. These activities extend to mitigation and preparedness. In sum, DRM can be characterised as an open system, as it tries to achieve goals, or a

desired state of affairs, using feedback (Brehmer, 1992; Coetzee & Van Niekerk, 2012; Rasmussen, 1997).

A closer look at the DRM functions reveals a distinction between feedback and/ or control. From a control perspective, it can be argued that feedback loops lie within and between the functions in the cycle. A specific example can be found in the notion of preparedness. Here, the evaluation of exercises provides a range of information regarding, for example, the capacity and capability of operational units, the effectiveness of policies, and the co-ordination of multi-agency activities (Skryabina et al., 2017, 2018). This information feeds back into, and informs the development of generic or specific DRM elements. It also provides a starting point for effective pre-planning and feeds into not just the broader DRM system but, in a learning organisation, the training and development cycles that underpin it.

In addition to pre-planning, it can provide a desired baseline in the mitigation phase, and help in matching capabilities and capacities to tasks. The overall effect is that the system is better-able to cope with events. More broadly, feedback is used to stimulate goal-seeking behaviour in order to gain control. More specifically, it can be used to gain a better understanding of an activity, i.e. to evaluate its functionality, and, where necessary, improve its quality. In all cases, feedback supports the purpose of DRM, which is to control the problem (protect what humans value) and achieve a desired state.

2.2.3 Preparing for disasters

As illustrated in the previous sections, the DRM cycle consists of various interdependent functions, which, ideally, contribute to the overall purpose of protecting what is valued. Preparedness connects activities related to the elimination of the causes of a disaster (mitigation) to actions that limit damage due to the consequences of an event (response). For example, if prevention measures fail to protect lives and property from risks or threats, preparedness ensures that systems, organisations or individuals are ready to deal with, and respond to the effects or consequences. Preparedness activities should, thus, aim "to build the (response) capabilities needed to efficiently manage all types of emergencies and achieve orderly transitions from response to sustained recovery" (UN Secretary-General, 2016, p. 22). In order to achieve this, they should be based on a sound risk analysis and clear links with prevention capability.

Preparedness activities are undertaken before a disaster response occurs, to: (1) improve the response capability; (2) foresee potential challenges and develop solutions; or

(3) build capabilities, abilities or readiness to improve the effectiveness of the response (McEntire, 2007). In this context, existing plans, procedures and resource management can be analysed in the light of insights or feedback from previous activities such as exercises (simulated events) or the response to real disasters. As noted above, such activities should not only support a timely and effective response to the threat, and address the consequences of unwanted events, but also guide recovery by, for example, ensuring a swift return to normality (Lindell, 2013b). Preparedness can, thus, be seen as an ongoing, heterogenous approach that encompasses a range of activities, such as drawing up plans, running exercises, conducting seminars and learning from previous experience (Eriksson, 2010). Ideally, they form a sequential and iterative process that leads to improved capabilities.

Evaluation can support preparedness activities by, for example, measuring any gaps in the response capability. As highlighted above, the present research focuses on the evaluation product resulting from emergency preparedness and response exercises. These exercises are a safe opportunity to observe and evaluate the response and develop structured feedback that supports the ongoing development of the DRM system.

What is the role of exercises in DRM, in particular, preparedness?

In the absence of real-life events or responses, suitably-designed exercises are often seen as a practical way to simulate disasters, either partially or in their entirety. Examples range from full-scale field exercises, to small-scale, table top simulations (Skryabina et al., 2017). The aim is to test the emergency response system and its capabilities. Given the random nature of real disasters, these exercises are an important part of the DRM process and, more specifically, improving preparedness. They are a useful way to identify or demonstrate qualitative improvements, and should be seen as supporting an integrated and continuous approach in which lessons identified are incorporated into training programs, and tested in exercises to become lessons learned.

Both exercises and simulations must reproduce reality as closely as possible. The aim is that participants are already familiar with the crisis management process, should a real disaster occur (Borodzicz & Van Haperen, 2002; Gebbie et al., 2006). Biddinger et al. (2008) note that exercises can significantly improve the preparedness of systems and their capabilities, and they distinguish two levels: (1) at the individual level, exercises are an opportunity to educate personnel on disaster plans and procedures through hands-on practice, while offering constructive critiques of their actions; and (2) on an institutional and/ or system-wide level, well-designed exercises can reveal gaps in resources and inter-agency coordination, uncover planning weaknesses and clarify roles and responsibilities.

Exercises seek to recreate real events, either in their entirety, or a few key elements. In some cases, they may simulate a recent emergency with a well-documented scenario. However, they are inherently artificial, notably with respect to time and resources. Moreover, exercises are typically less noisy than a real disaster, with far fewer mental and external distractions. Human factors must also be taken into consideration. For example, participants are often aware of both the purpose of the exercise, and of being under observation. This may have positive and negative influences on performance. Heath (1998) refers to it as the *compresence* effect. The latter study also highlights other biases, such as hindsight (related to evidence) and time distortion. Furthermore, it is unlikely that participants will ever work together to manage a real crisis (Borodzicz & Van Haperen, 2002). Finally, Wybo (2008) notes that the lack of reality could impact the commitment of participants, who do not react as they would in a real-life, stressful situation.

Despite these shortcomings, exercises need to be evaluated. Lindell (2013a) states that evaluations are an integral component, because they help participants and other stakeholders to identify deficiencies in plans, procedures, training, equipment and facilities. These weaknesses can serve as the basis for developing measurable and achievable objectives when revising emergency response resources. However, Wybo (2008) sees this practice as naïve, arguing that it does not reflect the complexity of emergencies and crises. The former approach can be compared to single-loop learning: measuring gaps and correcting them (Argyris, 1976). As such, it identifies specific issues that are addressed by changes to procedures or policies. However, it overlooks unobserved, systemic issues that affect the whole, and which are the root cause of a repeating cycle of suboptimal outcomes. A deeper analysis of causal factors may provide a more comprehensive and lasting solution, and improve efficiency and effectiveness. The narrow approach generally adopts a single viewpoint: managers observe operators, while a more holistic overview would be achieved with a broader perspective. A well-designed evaluation structure is a key element in this broader approach.

2.2.4 Responding to disasters

As noted above, DRM addresses the full spectrum of risk prevention and reduction activities. The focus is on managing residual risk, which remains even when effective prevention/ reduction measures are in place. Residual risk drives the need to develop and support emergency services' preparedness and response capability. Following Lindbom, et al. (2015), capability can be seen as the ability to do something with the purpose of positively influencing the outcome of an adverse event. It is related to the notion of readiness, which can be seen as the ability to quickly and appropriately

respond when required (UN Secretary-General, 2016). The capability of an actor will influence the severity of any consequences (outcomes). For example, high capability might reduce severity, or *vice versa*.

Given uncertainty with regard to future events, here, a broader definition of capability is used, namely, "the uncertainty about, and the severity of, the consequences of an activity given the occurrence of the initiating event and the performed tasks" (Lindbom et al., 2015, p. 45). This definition acknowledges that uncertainty, consequences, events and tasks are key. Uncertainty, in particular, is a natural component of capability, since we cannot know for sure what the consequences of the activity will be. It also underlines that capability can be measured in terms of success or failure with regard to intended performance. These core aspects make it possible to evaluate capability in terms of the consequences of an activity, and the performed tasks. Thus, it must be analysed or evaluated as a possible explanation for change (and hopefully improvement) in the behaviour of systems, people and organisations. More precisely, there is always uncertainty with respect to how successful an actor will be when responding to a certain scenario or event. Exercises and evaluations can reduce some of that uncertainty by testing the actor's capability with respect to a scenario of that type.

Response systems

Here, the focus is on the ability of an organisation, or organisations, to provide an appropriate and timely response, by mobilising suitable and sufficient resources to meet the needs of the incident. This complex socio-technical system (Abrahamsson et al., 2010) is composed of multiple actors and resources. The latter include, but are not limited to, official agencies such as fire and rescue services, police and emergency medical services, actors from the private sector, volunteers and non-profit organisations (Uhr et al., 2008). Together, they form a critical set of specialised agencies that have a specific responsibility to serve and protect society (UN Secretary-General, 2016).

More broadly, this implies that responders co-operate in order to perform tasks that mitigate the severity of consequences following a disaster (Uhr et al., 2008). These organisations are also engaged in other phases of the disaster cycle. For example, an effective, efficient and timely response relies on mature, risk-informed preparedness measures, notably the development of the response capability of individuals, communities, organisations, countries and the international community. This thesis focuses on the evaluation of this emergency response system, from a broad perspective – the system's capabilities and how they can be improved.

2.3 Evaluation

The previous sections introduced theoretical concepts related to the context in which this research is performed (DRM) and the subject of the evaluation (the disaster response system). This section focuses on the key concept, namely evaluation. Evaluation supports DRM by rigorously demonstrating the utility, quality and efficacy of response capabilities, which are measured against agreed benchmarks, with mature methods. A generic approach can be found in the general evaluation literature, which addresses, for example, definitions. The starting point for the present investigation is to establish what to focus on, and which aspects of the evaluation to analyse.

2.3.1 What is evaluation?

Stufflebeam and Coryn (2014) state that evaluation can be seen as an essential characteristic of the human condition. It permeates all areas of human activity, with important implications for maintaining and improving services, and protecting citizens. Evaluation, in the broadest sense, provides data that is needed for the development of quality assurance and improvement activities.

The concept can only be fully understood with reference to the 'logic of evaluation' (Scriven, 1980), which is considered as a meta-theory in the field of program evaluation (Shadish Jr. et al., 1991). In the present study, it forms the theoretical basis for defining evaluation. The process consists of the following basic steps (Hurteau et al., 2009):

1. selecting relevant criteria: identifying elements or components that influence the performance of the object being studied (the evaluand);
2. setting performance standards based on criteria which, in turn, become the anticipated level of performance;
3. gathering data pertaining to the performance of the evaluand (analysing the extent to which performance meets established standards); and
4. integrating the results into a final value judgment (synthesis).

Hurteau et al. (2009) note that Scriven used this logic to conceptualise an 'evaluation double pyramid' that encompasses two processes. The first, comprising steps 1 and 2, is an *analysis process*. This consists of assessing the merit of the evaluation object by identifying its purpose, the general criteria and indicators required to describe it, along with benchmarks or other data relative to each of the criteria (i.e. standards). The second, comprising steps 3 and 4, is a *synthesis process*. This consists of inferring conclusions by analysing each indicator based on its performance data, in relation to

each dimension, and moving from these inferences to a judgment – a conclusion about overall merit. The evaluation is, therefore, a careful, systematic process that seeks to prevent erroneous interpretations of the object's value. This ensures both the collection of high-quality data, and provides a rationale for the interpretation and communication of the findings, and any judgements/ conclusions (Stufflebeam & Coryn, 2014).

One of the earliest and most enduring definitions of evaluation is "determining whether objectives have been achieved" (Stufflebeam & Coryn, 2014, p. 6). However, this is a relatively narrow definition, as success cannot be equated to meeting objectives, and poorly set objectives can lead to failure. Another popular broader definition calls it "a systemic assessment of the worth or merit of an object (or evaluand)" (Joint Committee on Standards for Educational Evaluation, 1994, p. 3). Both Scriven (1991) and Vedung (1997) add 'value' to the definition. However, value can also be seen as the evaluation's root term and evaluations are not value-free (Scriven, 1993; Stufflebeam & Coryn, 2014). In the end, the evaluation should assess the object's standing against referenced values. The reason for adding 'value' is that governmental evaluations often measure the outcome of an intervention as whether it has achieved its objectives in terms of value for money (Vedung, 2010).

From a broader perspective, evaluation can be defined as a systemic assessment of the worth, merit or value of an object. In this respect merit (or intrinsic value) can be understood as 'does the object or evaluand perform well and achieve the desired outcome?' However, an object that scores high on merit might not have worth. Worth (or extrinsic value) refers to a combination of excellence and service in a clearly-defined context, with consideration given to costs, thus including merit (Stufflebeam & Coryn, 2014). It considers both context and costs, and must be linked to an assessment of need, with the aim of achieving a defensible purpose, within a particular time period.

Although the latter definition can be used as a starting point for identifying and structuring evaluations, it should be noted that the literature is rich in concepts, standards, guiding principles, practical guidelines and approaches. History is littered with attempts to structure and classify them, using a range of methodologies. The next sections will focus on the most common techniques, which have influenced the present research.

Formal and informal evaluations

A key difference relates to how they are performed. Formal evaluations are systematic and rigorous, while informal evaluations are of a more *ad-hoc* nature (Stufflebeam & Coryn, 2014). This is an important distinction, notably because in practice evaluation is ubiquitous. It is inherent to our daily life, and each individual performs it

continuously, often subconsciously. However, the present study concentrates on formal evaluations. A key requirement is the collection of accurate information that supports any conclusions. Rigorous findings should be based on appropriate, credible and reliable information (Stufflebeam & Coryn, 2014). In this context, in 1994, the Joint Committee on Standards for Educational Evaluation defined five fundamental concepts for program evaluation standards: utility, feasibility, propriety, accuracy and accountability. These concepts can be applied to formal DRM exercise evaluation, as shown in Table 3.

Evaluators of formal evaluations should make selective use of both qualitative and quantitative data collection tools and strategies. It is therefore interesting to identify what information is collected, using which method, in which context, and how reliability and validity are guaranteed. Multiple information sources can be especially important in ensuring the validity of observations made in a dynamic environment, such as the evaluation of DRM exercises. It is also important to ensure integrity and credibility. This can be achieved *via* a meta-evaluation, which is defined as the process of evaluating an evaluation. A meta-evaluation involves isolating, obtaining and applying descriptive and judgemental information that makes it possible to identify the initial evaluations' strengths and weaknesses (Stufflebeam & Coryn, 2014; Table 3). The meta-evaluation can serve various purposes, but it is often used to scrutinise evaluations, and assess the need to adjust or amend systems.

Table 3: Five fundamental concepts related to program evaluation and their relevance to DRM. Based on Stufflebeam and Coryn (2014).

CONCEPT	EXPLANATION	DRM-RELEVANCE
Utility	*An evaluation should be useful:* Utility relates to the recognition, in the design phase, that any process ultimately delivers information of a quality, in a format, and at a sufficient level of detail to support future developments regarding policy, procedures, knowledge or skills.	Emergencies, disasters and crises continue to occur. In order to prepare for them, response organisations plan, educate, train and exercise. These preparedness activities ideally help systems, organisations, teams, and individuals to be better prepared for the next disaster, thus responding more effectively or efficiently. Evaluations can effectively deliver information that can be used by a wide variety of users for a variety of purposes.
Feasibility	*An evaluation should be feasible:* The evaluator should employ procedures that meet the needs and restrictions of the area or areas under evaluation. The evaluation should be conducted as efficiently and cost-effectively as possible.	When setting up and running DRM evaluations, procedures must be workable and applicable to the context and conditions they are used in. The chosen methodology should avoid disrupting or biasing the activity under consideration, be it an exercise, or a real-time intervention.
Propriety	*An evaluation should meet conditions of propriety:* An evaluation should be conducted legally, ethically and with respect for the welfare of those involved, as well as those affected by the results. This implies that it should be grounded in clearly-defined, written agreements setting out the obligations of the evaluator and client in regard to supporting and executing it.	DRM evaluations should promote sound principles of disaster management, fulfil the aims and objectives, and ensure the effective performance of activities in order to mitigate the effects of a disaster situation. Their design and implementation should be guided by any applicable organisational, local, regional or national policy, and clearly-established guidelines and methodologies. They should clearly define roles and responsibilities, and must be consistent and equitable. The performance standard should be specified, in order to avoid money, time and quality being lost or wasted. The evaluator should protect all parties' rights and dignity, and the findings must be honest and unbiased. The process should be impartial and independent. It is therefore important that all stakeholders are encouraged to share their views, for example, both organisational and citizens' perspectives.
Accuracy	An evaluation should ensure that any comments made will convey technically-accurate information that will assist in determining the merit and/ or worth of the object under evaluation (i.e. the evaluand/ evaluee). The evaluator should clearly describe the object as it was planned and actually executed, describe the object's background, and report valid and reliable findings. An evaluation should be clear, systematic and transparent, containing a clear logic that shows how conclusions were reached.	A DRM evaluation should not only focus on achieving and measuring indicators (tick boxes) but also facilitate the identification of why something happened, in order to derive generic emergency preparedness and response norms that can support future organisational development. Thus, it is more qualitative than quantitative. Various methods and techniques can be used. However, it must be clear why certain methods were selected for a specific situation.
Accountability	*An evaluation should be fully accountable:* An evaluation should be designed in such a way that it supports the standardised and repeatable reporting of the environment, thus supporting comparability across a range of evaluations of similar events. Evaluations should be related to a specific purpose.	DRM evaluations should be cumulative. Evaluations of real-life emergencies and exercises should not be independent, but should build upon each other. They should be comparable and accessible. Complex constructs should be avoided, in order for them to be used by a broad audience in a variety of contexts. This implies that the process and its product must be as open and transparent as possible. Results must be made available and disseminated in an appropriate format.

2.3.2 Why do we perform evaluations?

Another important question is related to why the evaluation is being performed? What is the overall purpose or aim? Stufflebeam and Coryn (2014) relate evaluation to programs, and identified four main uses: improvement, accountability, dissemination and enlightenment. Hertting and Vedung (2012) relate it to governance and learning, and saw development and accountability as two further purposes. Venable et al. (2016) identified six different, but related purposes and linked them to design science. These are: (1) to determine how well an artefact is expected to achieve its expected environmental utility (an artefact's main purpose); (2) the quality of knowledge outcomes (will the artefact be useful in solving a problem or making an improvement); (3) to determine whether the new artefact/ theory improves the state-of-the-art; (4) utility, which is a complex, composite concept that goes beyond the simple achievement of the main purpose (examples include functionality, completeness, consistency, accuracy, performance, reliability, usability and fit with the organisation); (5) to assess other (undesirable) impacts such as side-effects; and (6) to elaborate on knowledge outcomes by discerning why an artefact works or not.

Other purposes include learning to learn, developing professional networks, creating shared understandings, strengthening (the project), boosting morale, quality assurance, supporting dissemination efforts, or to foster enlightenment (Forss et al., 2002; Stufflebeam & Coryn, 2014). Here, improvement or development (including learning) and accountability are seen as two very distinctive, but common purposes (Boin et al., 2008; Bovens et al., 2008; Hertting & Vedung, 2012), notably when using formative and summative efforts as defined by Scriven (1967) as a basis (see section 2.3.3). It is interesting to examine evaluations, and relate their purpose to the information they are expected to provide, and how they are conducted. In the context of DRM exercises, learning and accountability are two abstract purposes. Learning can be achieved through evaluation outcomes that provide feedback to participants; accountability may be supported by providing feedback on performance to decision-makers and other stakeholders. These points are developed in the next sections.

Learning

Learning is a broad and abstract concept. There are many ways to define it and, unfortunately, there is a lack of consensus. The most basic definition refers to "a relatively permanent change in behaviour as a result of practice or experience" (Lachman, 1997, p. 477). However, it can be argued that this is very narrow. In particular, the idea of a change in behaviour is open to debate, as learning is not a linear process (Lachman, 1997). Evaluation can support learning in various ways: (1) it can

provide feedback on practice and experience, in order to stimulate change, and enhance effectiveness, and/ or efficiency; or (2) it can verify whether learning has taken place through identifying any changes in behaviour.

Learning can occur along three dimensions: personal (or individual); interpersonal (or team); and institutional (organisations) (Borodzicz & Van Haperen, 2002). It develops in many ways, and every dimension can have its own requirements. Consequently, various models of learning processes can be used, depending on the dimension. It should be noted, however, that institutional learning can be closely related to accountability. This point will be discussed in the next paragraph.

In the context of DRM exercises, the work of Piaget (individual), Lewin (group) or Kolb (experiential) is often cited as the starting point (Borodzicz & Van Haperen, 2002), but many other authors have made substantial contributions to the literature (see Cassidy, 2004 or Driscoll, 1994 for a non-exhaustive overview). Learning theories can be used when developing exercises, in particular, the work of Kolb is often applied to designing simulations. The latter theory is focused on planning simulations that enable students to acquire knowledge, competencies and skills, but also gives them space to craft their own mental model, try it out, and observe and evaluate the results. Borodzicz and Van Haperen (2002) developed a generic synthesis of these approaches, and identified that in the context of exercises, the following prerequisites must be taken into account:

- understand the prior knowledge of learners;
- their social and operational context; and
- the degree to which they are able to reflect on previous experience and training in order to develop new mental models.

It can therefore be argued that evaluations that seek to support learning must consider their users or participants. A key element is clearly-specified objectives that are developed through discussion with end-users. This approach ensures that both the content and format meet end-user requirements, and enhances the usefulness of any documentation.

Accountability

Like learning, it is difficult to identify a single definition of accountability. It can mean different things to different people, and it is generally best to refer to it as 'being accountable' (Bovens, 2010). Both learning and accountability can be seen as elusive concepts that can always be made to respond to a need (Bovens, 2007). It is important

to take this observation into account when developing evaluations, as the results may be blurred any imprecision.

Bovens (2010) states that in an American context, accountability is often seen as a virtue and used as a normative concept, in the form of a set of standards used to evaluate the behaviour of public actors. Organisations and officials are expected to 'be accountable'. In Europe, it is often used more narrowly to mean an institutional relation or arrangement in which an agent can be held to account by another agent or institution. It can thus be used, incorrectly, as a synonym of evaluation. In general, it can be argued that the focus of accountability studies is not whether agents have acted in an accountable way, but whether they are, or can be held accountable *ex-post* (Bovens, 2007, 2010).

As the focus of the present research is on how evaluation can contribute to the process of being accountable, it is important to understand that accountability can be seen either as a virtue or as a mechanism. Evaluation can use the normative concept, or it can focus on the agent who is being held accountable. In classifying accountability it usually helps to ask two questions: accountability to whom? and accountable for what? This also applies to evaluation outcomes.

In the DRM context, accountability is often applied to public administrations, as response organisations are, mostly, governmental organisations. Populations, notably in developed countries, have high expectations regarding safety and security (e.g. Boin et al., 2017; Clarke, 2005; Kapucu & Van Wart, 2006) and thus of such organisations. The public expects to be safeguarded by their state if something out of the ordinary happens (Boin & 't Hart, 2003; Eriksson, 2010). The existence of organisations that are capable of responding to future crises is something that the public, taxpayers and potential victims expect (Boin et al., 2017; Eriksson, 2010). Emergency management organisations are held accountable by the public or their representatives, and must answer to those they serve, the potential victims of disaster (Kirschenbaum, 2003). The reason for this is simple – the public legitimises and financially subsidises them. Through much of history, disaster agencies were judged successful or not by standards that were either developed internally, or set down by government. This approach led to standards that reflected internal performance criteria based on organisational needs with limited, if any, consultation with their 'client group', the general public.

Accountability can also be a part of learning (Bovens, 2007). More precisely, it is an essential part of deutero-learning, an institutionalised capacity to learn. This blurs the distinction between the two concepts, as accountability can also be seen as an element of any learning cycle. A specific example is political accountability, which may result in scapegoating, blame deflection, and defensive routines, instead of policy reflection and

learning. Heath (1998) emphasises that organisations need to make clear that judgements regarding guilt will be made by another group of people, and not by the evaluation process itself, which is designed to gather data to measure the effectiveness of people, processes, or policies.

The users of the evaluation determine its use. Its design should, thus, take these concepts into account as a preliminary step.

2.3.3 How can evaluation outcomes be used?

In addition to the purposes given above, it is also possible to classify evaluations based on their uses. The previous section addressed why evaluations are conducted, this section focuses on how they can be used.

Formative or summative evaluations

Scriven (1967) makes many conceptual, methodological and practical contributions to evaluation. His work makes a key distinction between *formative* and *summative* evaluations. Formative evaluations provide information that is used to develop a service, and ensure or improve its quality. Venable et al. (2016) note that they are often regarded as iterative or cyclical, and are used to produce empirical interpretations that provide a basis for improving the characteristics or performance of the evaluand. They are both prospective and proactive as they can be undertaken during the development of a program, or during its operation. Summative evaluations are retrospective assessments of completed projects, established programs, finished products or services rendered. They provide an overall judgement of the effectiveness of the individual, team, organisation or policy, given the expected outcomes. Venable et al. (2016) state that they are used to measure results, and produce empirical interpretations that provide a basis for creating shared meanings about the evaluand in different contexts.

Formative evaluations often form the basis for, and supplement summative evaluations. A summative evaluation can build upon information from a formative evaluation, by retrospectively compiling and assessing data once development is complete. Formative evaluations are often linked to improvement or development (by insiders), while summative evaluations are used for accountability or selection (by outsiders) (Hertting & Vedung, 2012; Stufflebeam & Coryn, 2014; Venable et al., 2016). It is reasonable to expect that both types of evaluation are applied in DRM exercises.

Closely related on the continuum are *ex-ante* and *ex-post* evaluations. An *ex-ante* evaluation can be seen as a predictive assessment that is used to estimate and evaluate the impact of future situations. It is a calculated appraisal of the consequences of

proposed interventions that is performed before the intervention is adopted (Vedung, 2010). An *ex-post* assessment evaluates the implemented system or evaluand on the basis of various measures. It is an assessment of an adopted, ongoing or completed intervention (Vedung, 2010). Venable et al. (2016) note that while it may seem intuitive that *ex-post* evaluations are always summative, and *ex-ante* evaluations are always formative, the terminology only refers to timing.

Functions of evaluations

The first attempt to classify evaluations was undertaken by Guba and Lincoln (1989). They classified them into four generations, as shown in Table 4. However, it should be noted that they introduced the fourth generation themselves, as a demonstration that their approach went beyond the three earlier generations. The authors argue that there are at least three major flaws in the three previous generations: (1) a tendency to managerialism; (2) a failure to accommodate a conflict of equally correct and fundamental values (value pluralism); and (3) overcommitment to the scientific paradigm of inquiry. In their fourth generation, stakeholder participation is a core element in all phases. The purpose is to identify and clarify the variety of constructions that exist or emerge among stakeholders.

Table 4: Guba and Lincoln's (1989) four generations of evaluation.

GENERATION	DESCRIPTION
1. Measurement	In the first generation, the role of the evaluator was technical. They were expected to be aware of the full array of available instruments, so that any variables named for investigation could be measured. If appropriate instruments did not exist, the evaluator was expected to have the expertise necessary to create them.
2. Description	The second generation is characterised by descriptions of patterns of strengths and weaknesses with respect to certain, stated objectives. The role of the evaluator was that of describer, although the earlier technical aspects of the role were retained. Measurement was no longer treated as the equivalent of evaluation, but was redefined as one of several tools that might be used in its service.
3. Judgement	This generation is characterised by efforts to reach judgements. The evaluator assumed the role of judge, while retaining the earlier technical and descriptive functions.
4. Responsive-Constructivist	The fourth generation sees the claims and concerns of stakeholders as organisational foci (the basis for determining what information is needed). These foci are implemented within the methodological precepts of the constructivist inquiry paradigm. If it is the case that people act in accordance with their constructions, then the evaluator is a leading agent in the process of changing action, and action for change.

Following Guba and Lincoln (1989), Hansen and Vedung (2010) identified four *waves*, reflecting what they saw as different styles that had swept over parts of the world at different times. Their work captures the passage of time, showing how evaluations are used, and the implications regarding how they are conducted (Table 5). The study is another illustration of what evaluations are intended to achieve, and how they should achieve it.

Table 5: Hansen and Vedung's (2010) four waves of evaluation.

WAVE *PERIOD (EST.)*	DESCRIPTION
Science-driven *1950–1975*	Evaluation was expected to provide trustworthy scientific findings regarding adopted policies and programmes. It was based on a means-ends rationality. Goals and objectives were set by bodies outside the scientific community, and expressly recognised as subjective. Researchers examined, in experimental settings, the ability of various means to reach these externally-set ends. These experiments were expected to deliver objective generalised truths.
Dialogue-Oriented *1975–1990*	Evaluation was more pluralistic. Participants other than politicians, upper-management and academic researchers were involved. Also known as 'stakeholder evaluation', groups or individual actors had an interest in the intervention to be evaluated. The claims, concerns and issues of stakeholders served as points of departure. Interest could be measured in terms of money, status, power, face, opportunity, etc. Far from a rigorous, scientific two-group experiment, evaluation was supposed to be based on discussion, dialogue and communication among equals.
Neo-Liberal *1980–1995*	Evaluation supported results-based management. It was used to provide insights into accountability. In addition it was customer-oriented, focused on value-for-money or cost-effectiveness. They took the form of accountability assessments, performance measurements and consumer satisfaction appraisals.
Evidence *1995–*	Evaluation is focused on what works and (empirical) evidence. It is based on a means-ends rationality. The task is to enhance and disseminate knowledge of means.

There are many other ways to classify evaluations (e.g. Stufflebeam & Coryn, 2014). However, the generic methods outlined above are used as a point of departure in the present study. These overall classifications underline that the intended use of an evaluation plays a key role in its design, and that it is important to consider the historical timeframe. In addition to this theoretical or methodological classification, other concepts can be distinguished that may influence the evaluation design.

Usefulness

Guba and Lincoln (1989) highlight that stakeholders play a crucial role (in particular, in their fourth-generation evaluation). They emphasise the need to make (or collect) judgements about the merit and/ or worth of the object being evaluated, instead of simply measuring or determining whether objectives have been met (Stufflebeam & Coryn, 2014). Closely related to their work is the so-called 'constructivist evaluation', which is also referred to as responsive or stakeholder-centred evaluation (Stake, 1974). These approaches are seen as a radical departure from earlier generations. In this context, it is possible to identify many sets of stakeholders or users. Guba and Lincoln (1989) identified three broad classes, each with subtypes:

1. agents: people who are involved in producing, using and implementing the evaluand;
2. beneficiaries: people who profit in some way from the use of the evaluand; and
3. victims: people who are negatively affected by the use of the evaluand.

Another, more common way of identifying these users is based on their role in the evaluation process. They can be broadly called users, producers and evaluators (see Table 6).

Table 6: Overview of three generic roles in the evaluation process.

GROUP (ROLE)	DEFINITION OF THE (GROUP) ROLE
User(s)	An individual or a group that uses the evaluation as a means rather than an end, in the context of their organisational position, tasks and responsibilities. They use tangible outcomes/ products/ results for a specific purpose. The user can also be the object of the evaluation (the evaluee), or attend it, or be the person who commissioned it.
Producer(s)	An individual or group (e.g. team) that is responsible for setting up the process (development, design and execution). The producer is also responsible for ensuring that the evaluation addresses all relevant questions, meets information requirements and compiling data into a tangible product. They are responsible for staffing, finding a location and funding the evaluation. An evaluator is part of this group, but their specific role is investigated separately (see below).
Evaluator(s)	An individual or group that operationalises/ conducts the evaluation. The evaluator is responsible for the execution of the evaluation task, and follows the process set out by the producer. They collect data and can also be the producer.

Evaluation can be compared to information management. In particular, the product can be seen as shared information. Users play a crucial role in information management, as their willingness to adopt new systems determines the success of the project. Various management information systems studies have examined the notions of 'use' or 'usefulness' (see e.g. Davis, 1989; Franz & Robey, 1986; Hendrickson, et al., 1993; Karahanna & Straub, 1999). In the present study, Papers III and IV involve the user.

Davis (1989) introduced two theoretical constructs to describe usage: 1) perceived usefulness (PU); and 2) perceived ease of use (PEU). These constructs are still in use, and have been studied and retested several times (Hendrickson et al., 1993; Karahanna & Straub, 1999). Perceived usefulness can be defined as the tendency of people to use or not use an application, to the extent that they believe it will help them perform their job better. In the case of evaluations, we can assume that they are more likely to be used (i.e. perceived as useful) if the user believes that it will help him or her to be better prepared or perform better in the future. Perceived ease of use can be defined as the belief that the application is more or less onerous to use, and that any performance benefits must be weighed against the effort required to use it. In the case of evaluations this suggests that those that are difficult to use/ read, where the user must invest significant effort for a relatively small return, are not likely to be used.

Davis (1989) found that perceived usefulness has a strong correlation with user acceptance and should, along with perceived ease of use, be seen as a fundamental consideration in evaluation research and design. However, he also highlighted that the two components are subjective, and should be seen as 'beliefs' – meaningful behavioural determinants, rather than surrogate measures of objective phenomena. The key point

to note here is that even if an evaluation does objectively lead to a positive outcome for the end user, if the user does not perceive it as such, they are unlikely to use it and, in the end, it will not achieve its purpose.

Stufflebeam and Coryn (2014) emphasise that useful evaluations are grounded in descriptive and judgemental information. In general, users want to know what the object under evaluation was, and how well it performed. This requires the evaluator to collect and report both descriptive and judgemental information. Descriptive information should objectively describe the object in terms of its goals, plans, operations and outcomes, supported by factual statements. It should be kept separate from any judgements. Judgements go further, and are typically reached through the integration or synthesis of facts (descriptive information) and values.

2.4 Synthesis: evaluation in DRM

The above paragraphs introduced the concepts central to this thesis: risk, DRM, preparedness exercises and evaluation. It began with a discussion of the concept of risk. Risk, in this context, is uncertainty about and severity of the consequences of an activity with respect to something that humans value. It can be reduced by lowering the likelihood of an unwanted event, or by managing the severity of any consequences. The concepts of disaster preparedness and response focus on the post-event situation. Three important points emerge: firstly it is difficult to predict when, how and what will happen in the future; secondly, it is difficult to estimate how severe the consequences will be; and, thirdly, it is difficult to know how response efforts will perform and what their effects will be.

Despite these uncertainties, societies feel the need to prepare, predict and protect what they value. It should be noted that this research is focused on disasters—major events that, due to their magnitude, have significantly greater consequences than everyday incidents and emergencies. These severe events are unlikely to occur, which means that responders are unlikely to have a wealth of recent and relevant practical experience in dealing with them. The problem is usually addressed by designing and implementing a range of preparatory actions. Governmental disaster response systems play a crucial role. It could be argued that these systems are put in place to protect what society values, and that they do this by continuously reducing uncertainty and, more precisely, by reducing the likelihood and severity of consequences using feedback loops based on experience. Thus, in addition to dealing with real disasters, disaster response systems engage in activities such as exercises that aim to improve preparedness. In this context, the object of this study, evaluation, plays an important role.

This chapter underlines that evaluation should be seen as a means to an end, and not an end in itself. It can be used to support learning, or it can be used to identify developmental needs and deliver reliable reports to individuals within organisations. The information can support future development, and prevent a repeat of any negative impacts in the future (this may include establishing liabilities).

In the context of this study, and to link the concepts and definitions presented in this chapter, evaluation is redefined as: i) a systematic assessment that determines the value or performance of an emergency response actor with respect to the intended purposes, and the match between expected and actual outcome(s) in a given scenario; or ii) the product (e.g. report) of that assessment. This definition covers both the process and product of an evaluation. It highlights that the assessment of value or performance is directly linked to expected outcomes for an actor or a group of actors. Secondly it clarifies that it is the relationship between the purpose of the actor and outcomes that are the subject of the analysis. Thirdly, it highlights that the performance of an actor might be judged differently depending on the scenario. Fourthly, it illustrates that the process should be supported by a systematic assessment, which implies that there is a formal evaluation. Finally, it distinguishes between the evaluation process and its output, the product. This definition will be used throughout the remainder of this thesis when referring to an 'evaluation'.

Both the process and product contain elements that contribute to a useful evaluation. It should be noted that various authors have focused on operational response evaluations, and identified the elements that should be included. An overview of these frameworks, and their common elements can be found in Annex C. Such frameworks typically seek to identify how evaluations should be performed, and should be seen as complementary to the theoretical approaches and concepts introduced in section 2.3. Although there is some overlap between them, they draw upon the material used in the present study and, therefore, are seen as a beneficial conceptual starting point for this research.

No matter how operational response evaluations are performed, success depends on a variety of elements, notably use and users. It should, however, be noted that the various uses, users and purposes can conflict, which requires careful consideration of different needs. What might be useful for person A might be less so for person B. Nevertheless, users determine whether an evaluation will make an impact. Will it achieve its purpose, or will it be shelved as another useless product? Well-constructed, worthwhile evaluations should always make a positive impact. However, if the evaluation is really so useful, how can its full potential be exploited? The next chapter provides more insight into how exercises and their evaluation are currently performed in practice, in the specific context of the Netherlands.

3 Practical application of exercise and evaluation strategies

This section outlines the practical approaches and experience that are referred to in this thesis, and links them to the appended papers. It presents the physical context, the Netherlands, in order to provide a better understanding of the background.

3.1 Why, and how are exercises run?

It is important to understand that exercises, like their evaluations, are a means to an end. This implies that they support an overarching purpose, which is often learning or development. Callan (2009) notes that in the absence of an actual event or response, exercises have proven to be an effective way of evaluating and improving emergency management arrangements at all levels. They can have a variety of functions or benefits (Skryabina et al., 2017). For example, they can be used to plan, train and organise resources before a disaster, enabling the response system and its components to become more effective. They can also be used to generate (and validate) data related to disaster events, particularly in cases where there is little local experience. Finally, they are an invaluable source of information about less-well-known emergency management situations (Kelly, 1995). A sample of DRM exercise functions is illustrated in Table 7.

Table 7: Overview of DRM exercise functions. Adapted from Gov.uk (2021), Skryabina et al. (2017) and the United States Government (2021).

FUNCTION	IDENTIFY OR ASSESS	ENHANCE OR IMPROVE	TEST	VALIDATE	MEASURE
EXAMPLES	… planning and procedural gaps or deficiencies.	… the visibility and reputation of the agencies involved through publicising their work in the community.	… the adequacy of a disaster plan.	… training and education.	… improvement compared to performance objectives.
	… roles and responsibilities.	… the likelihood of the organisation or business surviving a disaster.	… the adequacy of personnel training.	… the preparedness program.	… participants' satisfaction or confidence.
	… the capabilities of existing resources and identify needed resources.	…coordination between internal and external teams, organisations and entities.	… communication systems, equipment and other resources.	…collaboration.	
	… an organisation's areas of vulnerability.	…awareness and understanding of hazards and their potential impacts.	… the viability of the emergency response network relative to the threat.		
	… limitations in plans or response (procedures).	… participant's knowledge, behaviour and understanding	… recently-changed procedures or plans.		
	… policies, plans, procedures, training, equipment and inter-agency agreements.	… the training of personnel in emergency roles and responsibilities.	… emergency response capabilities.		
	… gaps in resources.	… inter-agency coordination and communications.			
	… opportunities for organisational and regional improvement.				

At an individual level, exercises can help responders or participants to understand what they will have to do, and develop experience. They provide organisations, teams and individuals with a realistic opportunity to practice the skills, behaviour or knowledge previously acquired through training or education. They play a vital role in developing and keeping the skills of rescue personnel and voluntary groups up-to-date. In this context, it is important that exercises take into account the user's or participant's learning style (as described in section 2.3.2). They can also have a social function, as interactions can create networks of people who know who they can trust or rely on (Berlin & Carlström, 2014). Exercises can have many purposes or benefits, and this is largely a function of how they are organised (Lonka & Wybo, 2005). In the best cases, they involve multiple stakeholders and serve as real(istic) learning experiences for all participants. In the worst cases, they become the exercise of an exercise, reducing participant motivation and limiting any lessons learned.

The functions illustrated in Table 7 are found in several national policies, for example, Gov.uk (2021), the Swedish Civil Contingencies Agency (MSB) (2017) and the United States Government (2021). These policies often provide a common approach, or a set of (fundamental) principles regarding the design, development and conduct of exercises. One example of a comprehensive policy document (or program) is the United States Homeland Exercise and Evaluation Program (HSEEP). Its purpose is to provide a set of guiding principles for exercise programs, and a common approach to their management, design and development, conduct, evaluation and improvement. The HSEEP sees exercises as a key component of national preparedness, as they provide senior leaders and stakeholders with an opportunity "to shape planning, assess and validated capabilities and address areas for improvement" (United States Department of Homeland Security, 2020). It also explicitly notes that exercises are informed by the findings from risk and capability assessments, corrective actions from previous events, and external requirements such as regulations and grant guidance. The HSEEP is supported by other policies (United States Government, 2021), which emphasises that the development of a program starts with an assessment of needs and current capabilities.

In the United Kingdom (Gov.uk, 2021), national policies underline the role of exercises in reducing uncertainty. These documents explicitly state that planning for emergencies cannot be considered reliable until it has been exercised, and has proven to be workable, especially as false confidence may be placed in the integrity of a written plan. These examples illustrate that the theoretical link between the concepts of risk and DRM (outlined in the previous chapter) is also found in practice, and that exercises are seen as an important way to prepare for future disasters. Finally, it should be noted that they

facilitate lessons being learned as these controlled simulations make it easier to identify any areas of difficulty.

Types of exercise

Table 8 presents the different types of exercise, which vary from the individual level to full-scale simulations (Peterson & Perry, 1999). The scale is a function of the anticipated outcome and the skillsets of participants. For example, the needs of a team of rescue dog handlers are significantly different from teams responsible for the higher-level strategic co-ordination of operations. The size, level and complexity of any exercise must be a close match with the skills or knowledge being tested. Full-scale exercises test large-scale deployment, cooperation, coordination and methodologies at all responder levels. Higher-level tactical and strategic exercises can be run as a scripted, moderated tabletop simulation. Exercises can even be run in the normal working situation, but in the absence of the usual resources.

It is important that both the aim and outcomes of an exercise are clearly identified in the prevention phase (risk analysis), either based on the initial programme specifications or from previous evaluations. These outcomes give participants a common understanding of its purpose. In terms of frequency, the limiting factors are need, money, time and logistics. A plausible scenario will encourage a natural response from participants (Alexander, 2000). The success of an exercise depends on two factors: 1) the motivation of participants; and 2) the similarity with real-life conditions.

Table 8: An overview of discussion-based and operations-based exercises. Adapted from United States Department of Homeland Security (2021).

TYPE OF EXERCISE	UTILITY/PURPOSE	TYPE OF PLAYER ACTION	DURATION	REAL-TIME PLAY	SCOPE
Discussion-based	To familiarise players with current plans, policies, agreements and procedures; to develop new plans, policies, agreements and procedures	Notional; player actions are imaginary or hypothetical	Rarely exceeds 8 hours	No	Varies
Seminar	Provide an overview of new or current plans, resources, strategies, concepts or ideas	N/A	2–5 hours	No	Multi or single agency
Workshop	Achieve a specific goal or build a product (e.g. exercise objectives, SOPs, policies or plans)	N/A	3–8 hours	No	Multi-agency or multiple functions
Tabletop exercise	Help senior officials to understand and assess plans, policies, procedures and concepts	Notional	4–8 hours	No	Multi-agency or multiple functions
Game	Explore decision-making processes and examine the consequences of these decisions	Notional	2–5 hours	No (some simulations provide real- or near real-time play)	Multi-agency or multiple functions
Operations-based	To test and validate plans, policies, agreements and procedures; clarify roles and responsibilities; identify resource gaps	Actual; player action mimics the reaction, response, mobilisation and commitment of personnel and resources	May be hours, days, or weeks depending on the purpose, type, and scope	Yes	Varies
Drill	Test a single operation or function	Actual	2–4 hours	Yes	Single agency or function
Functional exercise or Command Post Exercise	Test and evaluate the capabilities, functions and plans of Incident Command, Unified Command, Intel centres or other command/ operations centres	Command staff actions are actual; movement of other personnel, equipment, or adversaries is simulated	4–8 hours or several days or weeks	Yes	Multiple functional areas/ multiple functions
Full-scale exercise or Field Exercise	Implement and analyse plans, policies, procedures and cooperative agreements developed in previous exercises	Actual	1 full day or longer	Yes	Multiple agencies or multiple functions

3.2 Why, and how are evaluations run?

The effectiveness of exercises and, more precisely, their effect on the response is difficult to determine, as there is little objective data (such as pre- and post-exercise scores or statistics) in the literature. It therefore appears that we do not know to what degree they are effective. In addition, the time, effort and resources that are required to design, conduct and evaluate an effective exercise should not be underestimated, and it can be seen as a time-consuming, expensive process that diverts resources away from other important needs (Callan, 2009; Hsu et al., 2004).

Although there are examples of evaluation standards or guidelines at local and national levels, and in related subject areas such as education, policy or project management, there is no widely-applicable standard approach in DRM. It is clear that academics and professionals need to work together to develop the concepts and principles that could underlie such an approach. Such efforts must acknowledge that it is difficult to make a direct comparison of systems and outcomes, due to the broad range of approaches across situations, organisations and nations (Rüter et al., 2006). Any proposals should ensure that the evaluation process is rigorous across the full range of scenarios, in order to make a meaningful assessment of the extent to which expected outcomes have been achieved (Klein et al., 2005). Beyond the exercise itself, a well-structured report can support development at all levels by drawing attention to actions that should be taken, and providing a starting point for future planning (Dausey & Moore, 2014).

Worldwide, there are few examples of general evaluation standards. Moreover, there are no comprehensive standards that specifically address response evaluation. Nevertheless, several scientific and practical frameworks could be used as a basis. Table 9 presents some evaluation standards or policies that are currently used in practice. These policies are mainly process-oriented, and provide some quality criteria related to the design and execution of evaluations. Finally, other national, regional and organisational instruments also provide guidance.

Table 9: Overview of evaluation standards.

ORGANISATION OR AGENCY	STANDARD AND ORIGIN	SHORT DESCRIPTION
International Organization for Standardization (ISO)	Societal security – Guidelines for exercises (ISO 22398:2013,IDT) https://www.iso.org/obp/ui#iso:std:iso:22398:ed-1:v1:en	The standard describes a generic approach to planning, conducting and improving exercise programmes and projects. Evaluation is defined as a "systematic process that compares the result of measurements to recognised criteria to determine the discrepancies between intended and actual performance". It mainly provides descriptive information about the role of evaluation within exercises along with some examples of indicators. It sees evaluation as part of a process of continual improvement and specifies the following steps: (1) initiating, (2) planning and organisation, (3) formulating questions and the basis for analysis, (4) training, (5) observing and directing feedback, (6) analysing data and developing the after-action report, (7) presenting the after-action report. The annex refers to the Plan-Do-Check-Act cycle.
Centers for Disease Control and Prevention (CDC)	Framework for program evaluation in public health https://www.cdc.gov/eval/framework/index.htm	The CDC emphasises that program evaluation is a systematic way to improve and account for public health actions through procedures that are: (1) useful, (2) feasible, (3) ethical and (4) accurate.
United Nations (UN Evaluation Group)	Norms and Standards for Evaluation (2016) http://www.unevaluation.org/document/detail/1914	The document contains norms that should be adopted, and institutional standards that should be reflected in the management and governance of evaluation functions. The norms are: (1) internationally-agreed principles, goals and targets, (2) utility, (3) credibility, (4) independence, (5) impartiality, (6) ethics, (7) transparency, (8) human rights and gender equality, (9) national evaluation capacities, (10) professionalism, (11) enabling environment, (12) evaluation policy, (13) responsibility for the evaluation function, and (14) use and follow-up. The associated standards are: (1) institutional framework, (2) management of the evaluation function, (3) competencies, (4) conduct and (5) quality.
The Organisation for Economic Co-operation and Development (OECD)	Quality Standards for Development Evaluation http://www.oecd.org/dac/evaluation/qualitystandards.pdf	OECD quality standards focus on development evaluation processes and products. Ultimately, they aim to strengthen the contribution of evaluation to development outcomes. This document provides process-related information with regards to: (1) purpose, planning and design, (2) implementation and reporting, and (3) follow-up, use and learning.
National frameworks		
The Federal Emergency Management Agency (FEMA), United States	Homeland Security Exercise and Evaluation Program (HSEEP) https://www.fema.gov/hseep	The HSEEP sees evaluation as the cornerstone of an exercise and an element that maintains the functional link with improvement planning. Through evaluation, organisations assess the capabilities needed to accomplish a mission, function or objective. Effective evaluation involves planning, observing, collecting and analysing data, and reporting outcomes. Exercise evaluation guides (EEGS) guide exercise observation and data collection. These guides refer to the 32 core capabilities identified in the National Preparedness Goal, along with national planning frameworks, threat/hazard identifications, risk assessments, or the organisation's own plans and assessments. The EEGS are structured as a function of phases or tasks (e.g. prevention, protection, mitigation, response and recovery).
The Civil Contingencies Agency (MSB), Sweden	Handbook 'Evaluation of exercises' https://www.msb.se/RibData/Filer/pdf/25885.pdf	This practical tool describes how a crisis management exercise may be evaluated. It covers both the behaviour of participants and the impact of the chosen format on outcomes. It presents a process consisting of eight stages: (1) appoint a head; (2) plan and organise in cooperation with management; (3) formulate questions and determine the basis for analysis; (4) train evaluators; (5) observe the exercise and collect direct feedback; (6) analyse the material collected and compile the evaluation report; (7) present and disseminate the results; and (8) use the lessons learned to start planning the next initiative.

In theory, exercise evaluation provides an excellent critique, however, in practice it can be based on qualitative impressions and verbal descriptions that are not amenable to quantitative analysis (Klein et al., 2005). Poorly-designed, or badly-executed exercises, along with an unevaluated or poorly-evaluated plan, may do more harm than good if they lead to a false sense of security and poor performance during an actual emergency (Gebbie et al., 2006). This can have significant consequences, as citizens expect to be safeguarded by organisations that have an assigned duty of care; should something out of the ordinary happen, the public may be put at risk (Boin & 't Hart, 2003). Conversely, a well-constructed and well-managed evaluation process is the key to collecting evidence-based feedback on performance across the full range of exercise activities.

3.3 The Dutch context

The Netherlands is the context for much of this research (see also section 1.4). This comprehensive, single-country approach made it possible to examine the work of most of the relevant actors, and provided an overview of how multi-organisational emergency exercises are designed, implemented and documented within a country. In addition, respondents shared the same experience of incident and crisis management response. Papers II, III and IV draw upon the situation in the Netherlands to provide case study material. In order to better-understand the geographical context, it is important to link the previously-introduced terms to the specific Dutch context. In particular, the next section provides a general overview of the crisis and DRM system in the Netherlands.

3.3.1 Crisis management and DRM in the Netherlands

In order to explain the Netherlands crisis management system, it is essential to start with the definition of crisis[5] in Dutch policy. A crisis is defined in the Safety Region Act as "a situation where the normal functioning of identified areas of vital interest is threatened". Brainich von Brainich Felth (2004, p. 13) argues that this definition could be supplemented by the observation that, in this situation, "the normal resources (capacities) are insufficient to deal with the threat". Four areas of 'vital interest' can be distinguished in Dutch policy: (1) Public safety; (2) National security; (3) Economic security; and (4) International security (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2004). Protection against threats to these vital interests takes the

[5] In the Netherlands context, the term 'crisis' is commonly used because of its wider applicability.

form of a process model that structures crisis management and its capacities. This model is called the safety chain, and resembles the well-known disaster cycle. As with many other crisis management systems, there are four phases: prevention, preparedness, response and recovery (see also section 2.2.1).

The crisis management system can be further divided into two decision-making structures that relate to the above-mentioned vital interests. On a national level, the generic crisis management structure mainly applies to issues related to public safety and public order. Here, crisis management is decentralised to local governments, while the Ministry of Justice and Safety is responsible for the overall system.

However, if a vital interest related to economic security (e.g. electric power), or public safety (e.g. infectious diseases) is threatened, a sectorial ministry (e.g. the Ministry of Economic Affairs in cooperation with power suppliers, or the Ministry of Health together with private healthcare institutions) become involved and/ or take the lead. Therefore, a second structure can be identified, referred to as the functional crisis management structure. The latter deals with sectorial crises and, ideally, collaborates with the generic structure.

Although two structures can be identified, as shown in Figure 3, this division is mainly related to the context of the crisis and the involvement of ministries and sectors. The two structures overlap and collaborate during crises that cross sectoral boundaries. In the response phase, the generic structure is complemented by the sectoral structure, and responsibility for responding to a crisis is shared.

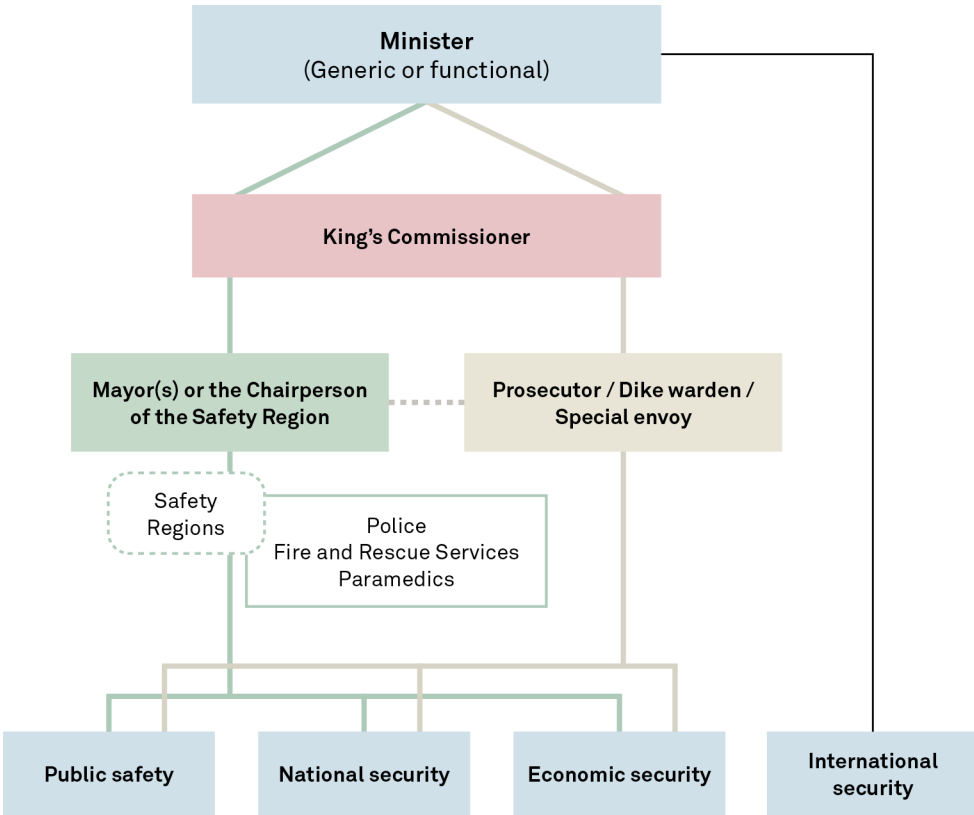


Figure 3: Overview of the Dutch crisis management decision-making structure (based on Rimbo-Gilde.nl, 2021)
The figure illustrates the complexity of the Dutch crisis management structure that is built around four areas of vital interest: I) Public Safety, II) National Security, III) Economic Security and IV) International Security. It shows that on a national level there can be 'generic' and 'functional' responsibilities, each triggering a slightly different crisis management structure (green/brown lines). For example if I) Public Safety is threatened, the generic crisis management structure will be most important (green line); here the mayor(s) or the chairperson of the Safety Region plays a key role. However, if another area is threatened, other entities such as prosecutors, dike wardens or other special envoys will also play an important role (brown line). To further complicate things, these two structures might be in place at the same time and need to collaborate or integrate (for example, as has been the case during the Covid-19 crisis). In addition, the GRIP procedure operates at the (supra-)regional level to ensure coordination.

As Figure 3 shows, the Dutch crisis management system has a more governmental than operational orientation. Most of the command-and-control function lies within governmental entities, with crisis management and decision-making taking place at various levels (local, regional, supra-regional and national). In the Netherlands, disaster and crisis management is decentralised and managed locally (Scholtens, 2008). At the local level, the Mayor is the responsible authority, while other entities provide guidance at various levels. Within this structure, there is a division between operational decision-making and governing, or policy-related decision-making. The Minister of Justice and Security is responsible for the overall system, and individual ministries are responsible

for their own sector. At the national level, ministries coordinate their response to crises, while at the (supra)regional level the Incident Command Procedure (GRIP[6]) ensures coordination.

Safety Regions

After two large-scale disasters (the Enschede fireworks[7] disaster in 2000, and the Volendam New Year's fire[8] in 2001), it became clear that the Dutch DRM system needed to be revised in order to be better-prepared for the future. Changes mainly related to the organisational structure, and new legislation was introduced. The *Fire Services Act*, the *Medical Assistance at Accidents and Disasters Act,* and the *Disaster and Major Accidents Act* were merged into one new law, the *Safety Regions Act*. This led to the creation of several new entities with an overarching, networking role, known as the Safety Regions.

The Netherlands is currently divided into 25 Regions, which are responsible for improving DRM and crisis management, and protecting citizens (Ministry of Security and Justice, 2013). A geographical overview of the regions is shown in Figure 4. It should, however, be emphasised that the Safety Regions are not themselves response organisations; instead, they bring together resources from traditional response organisations such as fire and rescue services, paramedics and/ or the police. They are, thus, responsible for the execution and maintenance of the generic crisis management structure on a regional level. They play a key role with respect to preparing for, and responding to emergencies and crises, and are mainly responsible for conducting emergency and crisis management response evaluations.

[6] The Coordinated Regional Incident Management Procedure (GRIP) is a nationwide emergency management procedure. It is used to scale coordination as a function of the area affected by an incident. There are four levels: the higher the level, the more complex the response (Van Duin & Wijkhuijs, 2015) .

[7] On 13 May 2000 a fire in a fireworks depot in Enschede led to an enormous explosion that killed 23 people and injured nearly 1000. A total of 400 houses were destroyed, and another 1500 buildings were damaged (Wikipedia, 2021b).

[8] On New Year's Eve 2000–2001, a fire broke out in a café in Volendam that was packed with young people aged between 13 and 22 after a sparkler hit the Christmas decorations hanging from the ceiling. A total of 14 people were killed, 200 suffered serious burns, and 241 were admitted to hospital (Wikipedia, 2021c).

Figure 4: Map of the Netherlands showing the 25 Safety Regions
This map of the Netherlands illustrates the division into 25 Safety Regions and their main cities. The names of most regaions relate to the province they are located in, however, there are some exceptions, notably regions with larger cities like Amsterdam, Rotterdam or The Hague.

3.3.2 Evaluation in the Netherlands

The Safety Regions Act (Ministry of Security and Justice, 2013) provides the basis for the work of the regions. It contains a description of the operational performance of the various emergency services, their organisation, roles, tasks and responsibilities, the risk analysis, and the format of plans and policy. Separate Orders in Council specify that information should be provided regarding the professionalisation of emergency services personnel and the quality standards of their equipment. The Safety Region is responsible for gathering this information, and must provide updates on the execution of their tasks.

There is only one direct reference to evaluation in the current Act. Article 23 notes that they must implement a quality assurance system, while Article 56 mentions a 'cost evaluation' and 'visitation'. Surprisingly, although Paragraph 18 is titled 'Evaluation', it only refers to the evaluation of the Act itself. In fact, while it is reasonable to expect that the basis for an evaluation (tasks, roles and responsibilities, purposes) would be set out in the Act, that is not the case, and it appears that the opportunity to establish a uniform process has been missed. Finally, Paragraph 14 specifies that the Inspectorate for Security and Justice has supervisory powers over the system, which could be seen as related to accountability.

The Netherlands also has an independent Safety Investigation Board (the Dutch Safety Board) instituted by a Kingdom Act (Overheid.nl, 2021). The Board's mission is to prevent or limit the consequences of future unwanted events by investigating and establishing probable causes and, if necessary, making appropriate recommendations. However, the focus is on post-accident evaluation rather than preparedness and, thus, exercise evaluation.

Exercise evaluation is, therefore, mostly a local responsibility. Safety Regions are able to design and structure their own version. They have their own exercise policy, and evaluate these activities as part of their preparedness policy. An overview of the multi-organisational evaluation reports found in the Netherlands system is shown in Table 10.

Table 10: Dutch crisis management evaluations.

TYPE	EXPLANATION
Evaluation of a simulated multi-organisational emergency exercise	Documents provide information on the results of an evaluation of a multi-organisational emergency response exercise within a Safety Region.
Evaluation of a systemic test exercise (simulation)	Dutch legislation (Overheid.nl, 2017) obliges Safety Regions to hold an annual crisis or disaster simulation. The Inspectorate verifies that they have been held, and prescribes the setup. These exercises aim to test the Dutch crisis management system (including the GRIP procedure).
Evaluation of a real multi-organisational emergency response	Documents provide information on the results of an evaluation of a multi-organisational emergency response within a Safety Region.

Key users of evaluation documents in the Netherlands

The Dutch crisis management system consists of users at various organisational levels (e.g. operational, tactical or strategic), who use evaluation reports in different ways. Three categories of primary users (people who use evaluation reports in their day-to-day work) were initially targeted by this research: mayors, (regional) operational leaders, and directors of Safety Regions. These groups are clearly identifiable within the Dutch crisis management structure and have different, but closely-related roles and responsibilities.

Mayors have specific legal responsibilities with respect to public security in their municipality. In their role as commander-in-chief, they have overall responsibility for command and control during local crises (Broekema et al., 2019). For events that extend beyond municipal boundaries, they can either collaborate with mayors of adjacent municipalities, or, in the case of more serious incidents, the chairperson of the Safety Region (who is selected from among the mayors of municipalities making up the region) can take over.

Both mayors and the chairperson are supported, on a practical level, by (regional) operational leaders. The latter operate under the responsibility of the mayor. On the one hand, they provide advice and implement his or her orders; on the other hand, they direct and support, for example, on-scene incident commanders. Organisations that are directly involved in the response (fire and rescue services, paramedics, and police) remain responsible for their own performance. Thus, operational leaders have a complex task, as they must manoeuvre in a multi-organisational operational and administrative environment. In addition, this group forms the link between strategic and tactical operational levels, which makes them, along with mayors, one of the key users of evaluation reports. It is, however, important to note that in the political arena, it is the mayor – and not the operational leader – who is held accountable for the (operational) decisions that were made, and the outcome of the response.

Safety Region directors (not to be confused with the chairperson) are responsible for day-to-day activities, and assist mayors. Their role is administrative rather than operational, and they are responsible for ensuring that their region is prepared to respond to disasters and crises by leading the daily operations of Safety Regions.

There are clearly far more (end) users or stakeholders. Table 11 presents a sample. Although non-exhaustive, it does illustrate the wide variety of potential stakeholders (individual roles and related organisations), and the scope of the challenges that must be faced when preparing an evaluation. It is important to identify all stakeholders in order to be able to determine whether evaluations achieve their purpose and are useful.

Table 11: An illustrative overview of crisis management evaluation stakeholders.

ROLE	ORGANISATION(S)
First Responders	Various response organisations such as fire & rescue services, police, ambulance, Local government.
Teamleaders, crisis managers or coordinators	Various response organisations such as fire & rescue services, police, ambulance, Local government.
Teachers, trainers or educators	Various education, training and exercise organisations such as educational agencies, training institutes, schools.
Evaluators	Various
Policy officers	Various entities such as safety regions, municipalities, ministries, the European Commission.
Mayors	Municipality
Council members	Municipality
Safety region directors	Safety Region
Lawyers	Legal organisations
Researchers	Various research organisations such as universities, research institutes, the Dutch Safety Board
Inspectors	Inspectorates (e.g. Justice and Security)
Consultants/ advisors	(Private) consulting agencies
Politicians	Parliament / government
Minister	Ministry
Reporters	Media or news agencies
Citizens	General public

Case selection: the Netherlands

Although this research focuses on the Dutch context, and its Safety Regions, the overall crisis management system can be compared to generic DRM systems (see section 2.2) and other national crisis management systems. Consequently, the findings can be applied to other, similar systems.

The purpose of the Dutch system is to protect the Netherlands and its population against threats. This purpose is common to DRM (as defined in section 2.2). The Netherlands also uses a model similar to the disaster cycle (as introduced in section 2.2.1). The functions that make up the Dutch safety chain model can also be found in other DRM systems. The Dutch system uses a variety of exercises and evaluations to either prepare for future events, or to derive insights and hold people accountable. Within the system, particular stakeholders have specific roles and functions. It should be noted that stakeholders might be named differently in other systems, or a stakeholder with the same job title might fulfil a different role in another organisational structure. For example, in the Netherlands, the mayor plays a key role within the crisis management system, while in other systems these responsibilities might be given to, for example, a governor, a chief of police, or a fire commander. Nevertheless, it is reasonable to assume that, in general, all crisis management systems have stakeholders who fulfil similar functions and play similar roles, as otherwise it is impossible to protect

the population. Non-Dutch readers can compare the information provided in this section with their own system (notably regarding evaluation stakeholders), and apply the results of this research to their system.

4 Research process and methodologies

The previous chapters addressed the 'why' and 'what' questions that play a central role in introducing and defining the research topic or phenomenon, and establishing its context. The next step is to describe and justify a set of defined and measurable activities, methods and techniques that are expected to lead to a deeper understanding (or greater knowledge) of the observed phenomena or outcomes.

This chapter describes and justifies these methods, before addressing the question of 'how'. It begins by introducing the underlying approach and philosophical assumptions underpinning this study, as this justifies the research design that was developed and implemented. Next, the various methods that were applied are presented, and the section ends with a reflection on the quality of the research.

4.1 Design science

In general terms, research can be defined as "an activity that contributes to the understanding of a phenomenon" (Vaishnavi, et al., 2004, p. 2). The phenomenon is typically a set of behaviours of an entity (or entities) that the researcher or the research community finds interesting. Often it ends in making a new and valid contribution to knowledge.

The phenomena that are the object of this research (the operational response and its evaluation) do not occur naturally; they are created by humans. In other words, they are artefacts. This research seeks to go beyond the creation of new knowledge about these objects, and aims to improve their usefulness (i.e. how they meet their purpose). As design research is part of design science, their combination is appropriate.

Design science focuses on how the design of artificial objects and phenomena meets certain goals. It can be contrasted with the natural sciences, which are focused on describing how objects or phenomena behave and interact with each other. The general goal of design science research is to create or contribute new and interesting knowledge in an area of interest, in other words, "a body of intellectually tough, analytic, partly formalizable, partly empirical teachable doctrine about the design process" (Simon,

1969, p. 58). It aims to understand the design and, subsequently, evaluate the effectiveness or performance of artefacts for the purpose of problem-solving (Hevner et al., 2004; Sein et al., 2011; Simon, 1969).

Nobel laureate Herbert Simon proposed that organisation and management research is a science of design (Van Aken & Romme, 2012). According to Van Aken (2005, p. 20) several sciences, such as medicine and engineering are considered design sciences since they focus on "develop[ing] knowledge that the professionals of the discipline in question can use to design solutions for their field problems. The mission can be compared to that of 'explanatory' sciences (such as the natural sciences and sociology), which is to develop knowledge to describe, explain and predict".

Whether DRM research should be considered as design research (or not) is likely to depend on the perspective adopted by the researcher. But, as solutions to problems encountered in the field are a very important part of current efforts, a design research approach is suitable. For example, emergency response organisations can be seen as artificial, purposeful systems that are associated with bureaucratic structures to maximise efficiency and effectiveness (Kirschenbaum, 2003). Design science research systematically builds and uses one or more artefacts that contribute to the understanding of a specific problem and its solution (Hevner & Chatterjee, 2010). It can be seen as a lens, or a set of synthetic and analytical techniques and perspectives for carrying out research (Vaishnavi et al., 2004). It aims to understand the design of an artefact and subsequently evaluate its performance or effectiveness, with the intention of improving it. Vaishnavi et al. (2004) note that the two primary activities are:

a. creating new knowledge through the design of novel or innovative artefacts (things or processes); and

b. the analysis of the artefact's use and/or performance *via* reflection and abstraction.

Hevner et al. (2004) argue that truth or justified theory (the goal of natural research), and utility or use (the goal of design) are inseparable, and should inform each other. Expanding on the combination of design and natural research, Baskerville and Pries-Heje (2010) go further, and argue that design research itself should include a descriptive, as well as a prescriptive element. From this perspective, design science can be used to understand and evaluate response organisations, support problem solving, suggest future improvements, and contribute to maximising efficiency and effectiveness, which makes it an appropriate approach in the present research.

Designing and finding solutions is an activity that is inherently connected to evaluation, as it provides feedback on the design and contributes to finding an optimal solution.

Together with building artefacts, evaluation is one of the two key activities that form the basis for design science research (March & Smith, 1995). It is therefore argued that it can contribute to understanding and analysing evaluation approaches—notably, the models and theories that are currently used in DRM practice. In addition, it can be used to build a new, common approach as it focuses on producing prescriptive knowledge in order to solve problems, rather than describing phenomena, which is the aim of traditional research.

4.1.1 An abstraction hierarchy

The design of artefacts such as exercises or disaster responses, and evaluating them can be broken down into various stages or phases. In order to evaluate a system and/ or other artefacts, it is important to understand how these elements are hierarchically related, if at all. Rasmussen's abstraction hierarchy (Rasmussen, 1985) provides a basis for analysing and understanding a system, as it can be used to describe top-down levels of abstraction. Rasmussen demonstrated how this hierarchy influences the design of a physical artefact. In a similar vein, Brehmer (2007) applied this concept to dynamic human systems, such as Command and Control, that are less static than, for example, power plant design. Brehmer demonstrated that the three conceptual levels of a design logic (purpose, function and form) can be used to analyse other, non-physical systems because they are constructed in a similar way, using a logic of design. If this logic is adopted, specific questions can be associated with each level (see Table 12, column A2) and it can be used as a basis for both designing exercises (Table 12, column A4) and their evaluations (Table 12, column A3).

Table 12: The logic of design science applied to exercises and evaluations. Adapted from Paper II.

A1 LEVEL OF ABSTRACTION	A2 QUESTION	A3 EVALUATION PROCESS EXAMPLE QUESTION	A4 EXERCISE DESIGN EXAMPLE QUESTION	A5 PROCESS-RELATED ANSWER
Purpose	Why?	Why do we need to evaluate?	Why do we need to run this exercise?	(1) Improvement or development (2) Accountability
Function	What?	What does the evaluation need to do in order to fulfil its purpose?	What does the exercise need to organise and demonstrate in order to fulfil its purpose?	Guba and Lincoln (1989) use the term 'generations' to denote the focus of a specific type of evaluation: (1) Measuring e.g. performance (2) Describing (3) Judging (4) Stakeholder-centred (extra) Testing (a combination of 1, 2 and 3)
Form	How?	How does the evaluation carry out the necessary functions?	How is it ensured that the exercise carries out the necessary functions e.g. what type of exercise is needed?	A systematic and rigorous exercise design, and ensuring that evaluative information can be collected in a similar manner.

In addition to this abstraction hierarchy, design science also uses so-called 'design propositions' (Denyer et al., 2008; Romme, 2003; Van Aken, 2004). Design propositions can be seen as the product of design research. They support professionals in the process of designing solutions to problems in the field (Van Aken, 2004, 2005). Design propositions can be formulated in various ways. The simplest is a rule of the form: 'if A do B'. This is also referred to as IO-logic, as it relates an outcome (O) to an intervention (I). However, Denyer et al. (2008) suggested that they should present information in a logical order: what to do, in which situations, to produce what effect, and to offer some understanding of why this happens. This is known as the CIMO-logic (Context, Interventions, Mechanism and Outcome). These CIMO-logic rules can be described as 'if you want achieve O in context C, you need to do I by creating M'.

Both design propositions and the abstraction hierarchy can be used to describe, design and evaluate artefacts.

4.2 Scientific paradigm

Remenyi, et al. (1998) noted that there are several questions that researchers should pay close attention to; for example, 'How do we do research?' and 'What do we research?' Questions like these can be linked to design science approach, and relate to describing the research paradigm. However, these questions cannot be answered without considering the question 'Why do we do research?' The latter question is addressed below, in the section that discusses the philosophical assumptions underpinning this study. As noted above, this type of research is not only aimed at describing and explaining the present (what is), but is also aimed at the future (what can be? or what ought to be?), which is the methodological approach taken here.

4.2.1 Philosophical assumptions

This research, like any other, is consciously, but also unconsciously influenced by certain beliefs and philosophical assumptions. It is important to be aware of them, and the role they play. Typically, they form the foundation for developing a study. For example, they shape how the problem and research questions are formulated, and how they are addressed (Crotty, 1998). However, this was not explicitly the case in this research project, and decisions were taken partly based on experience, logic and instinct as the work evolved. This does not mean that philosophy was not important, on the contrary, it was more or less implicit from the start.

In retrospect, the present research is guided by the premise that the ontological realm must exist independently of our knowledge of it. The epistemological idea that knowledge develops and changes based on what the researcher knows, is important. From this perspective, the world exists independently of our knowledge of it; it can only be understood using specific descriptions, and our knowledge is fallible (Easton, 2010; Mearns, 2011; Sayer, 2000). Furthermore, there is a clear distinction between the natural world and the social world. While the natural world can be measured and statistically analysed, the social world is not as simple, as it is made up of unique individuals whose emotions and behaviours cannot be accurately predicted or controlled. Knowledge is gained from an independent reality, which can be accessed by individuals who use their ability to reflect and learn from experience using their senses rather than quantifiable statistics. As Johnson and Duberley (2000, p. 164) frame it, "While the truth may well be 'out there' we may never know it in an absolute sense because we lack the necessary cognitive and linguistic means of apprehending it. [However]… we can develop, and indeed identify, in a fallible manner, more adequate social constructions or reality by demonstrating their variable ability to realise our goals, ends or expectations since our practical activities allow transactions between subject and object". This view is reflected in the research process and design described below.

4.2.2 Research design and process

This research focuses on the practical problem of the usefulness of evaluation products, rather than being a quest for basic knowledge about them. It seeks to produce explanations that can guide, and may be evaluated by human interventions in social worlds. The approach resembles other problem-solving disciplines that have a long history. One example is medicine, which seeks to diagnose, prognose, treat and prevent disease. More precisely, the first part of this research aimed to identify current evaluation practices, both in theory and in practice, whilst the latter part aimed to suggest improvements that can increase the usefulness of evaluations. Although this combined approach may make it more difficult to relate the research to one interpretive framework, it is possible to divide it into parts that are investigated from a common point of departure. The research strategy that was deployed was not set in stone from the beginning, but gradually evolved in an iterative process (indicated by the research questions presented in section 1.3). In general, this strategy consists of three phases, shown in Figure 5 and adapted from Runeson, et al. (2012).

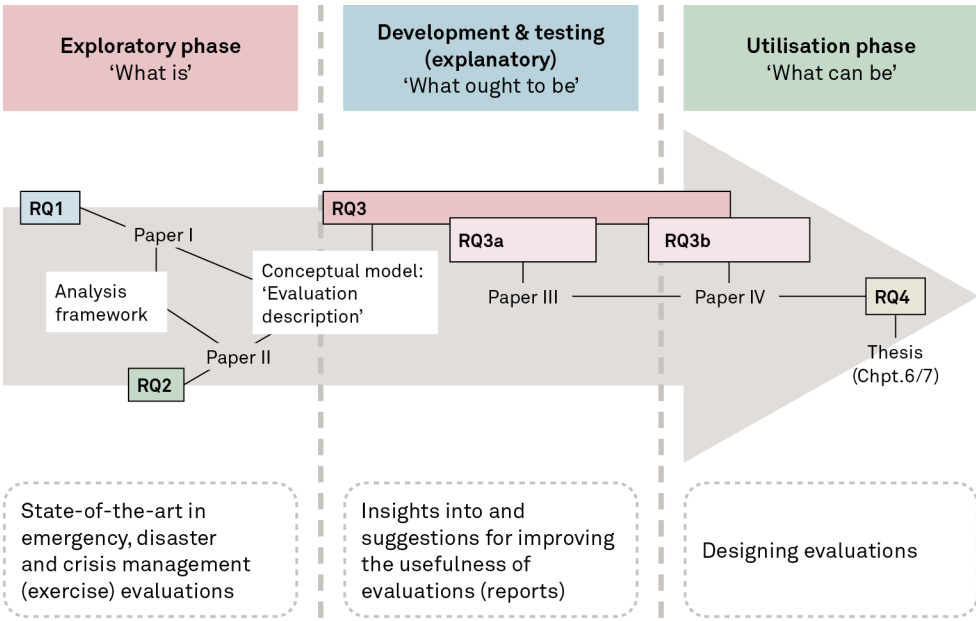


Figure 5: Overview of the research process and -design
This figure shows the various research phases: I) Exploratory – what is?, II) Development & testing / Explanatory – what ought to be?, and III) Utilization or application – what can be? It also relates the various research questions, the papers, and their relevant output to the various phases in order to show how they contributed to research in the respective phases, and how the research evolved.

The first stage can be seen as *exploratory*, describing what is. RQ1 and RQ2 laid down the theoretical and practical foundations for the research that followed. This phase focused on an investigation of the state-of-the-art in emergency, disaster and crisis management exercise evaluations in the Netherlands (see section 3.3). During this phase, Papers I and II were developed. These papers are closely linked and build upon each other, as Paper II uses a similar approach with a focus on practice. In Paper I, a framework was developed to run an in-depth analysis of the content of the literature focused on evaluation design. In Paper II, this framework was adapted and re-used to investigate evaluation design in the Dutch crisis management context. This exploratory phase highlighted a diverse literature, and a topic that is gaining more research attention. But it also revealed that professionals lack guidance or scientific foundations to support the design of their evaluations. The results from this phase motivated the development of a conceptual model to investigate the usefulness of evaluations.

The exploratory stage was followed by the development and testing, or *explanatory* stage, which aimed to identify and describe 'what ought to be'. The insights and findings from RQ1 and RQ2 led to the development of a conceptual model—the evaluation description—that was used as a basis for investigating RQ3 (Paper III). This

model was also used in order to identify the relation between usefulness and two distinct purposes: learning or accountability. Paper III reports the results of an experiment with crisis management professionals. The findings illustrate how various aspects of an evaluation influence its usefulness with respect to these two purposes, and they have implications for the way evaluations should be documented in practice.

The final stage can be seen as the *utilisation* or *application* phase. RQ3(b) and RQ4 play crucial roles. In this phase, practical observations and findings were combined with theoretical insights from RQ1–RQ3 to provide an insight into 'what can be'. This phase is partially reflected in the latter part of this thesis (Chapters 6 and 7), and supports the transfer from theory to practice. It provides an initial approach to anchoring evaluation in the broader context of DRM, and highlights the need to use a systematic, standardised approach to the design of future evaluations, as this would enable exercises to be compared.

Mixed design

Both the fields of evaluation and DRM are interdisciplinary, complex and dynamic. Therefore, a relatively open research design was applied from the beginning. However, this became more specific and focused as the process evolved. Although this research project tends to be qualitative, it applied a mixture of qualitative and quantitative approaches in order to develop a clear understanding of evaluation, and research on the topic. Initially (RQ1 & RQ2) a more exploratory, qualitative approach was adopted to describe 'what is'. This was complemented by a quantitative approach (RQ3) that tested and explored a concept. Together, these two approaches allowed information to be gathered from a wide variety of sources and a number of perspectives, using complementary rather than conflicting methodological techniques. They were then used to develop a comprehensive framework (to answer the question 'what can be') and address the issues surrounding evaluation design (RQ4).

It could be argued that the project sought "to choose the combination or mixture of methods and procedures that works best for answering research questions" (Johnson & Onwuegbuzie, 2004, p. 17). The selection of a variety of methods was another attempt to reduce uncertainties through triangulation. Triangulation seeks to identify convergence and corroborate results from different methods and designs while studying the same phenomenon, in order to improve the quality of the research (see also section 4.4). Overall, the present research can be seen as an example of a mixed methods approach, which opens the door to different worldviews and assumptions, as well as different forms of data collection and analysis (Creswell, 2003). It strengthens research findings through the combination of information sources and analytical approaches.

Nolan and Walsh (1995) state that the multi-method approach combines analytical strength and qualitative reflection drawn from a number of social science disciplines, thus increasing the variety of data gathered and improving understanding. The selected methods generate valid and reliable data that can be used to explore issues and seek out causal explanations.

Section 4.3 discusses the various data collection methods and instruments in greater detail. It also discusses the strengths and weaknesses, in order to justify the use of the mixed methods approach.

Case study design

This project also (partially) applied a descriptive and/ or exploratory case study design (Baxter & Jack, 2008). In particular, questions focused on professional practice used the Netherlands as a single case. In line with Yin (2003), the aim was to explore phenomena in context. This comprehensive approach covered most of the relevant actors in the country, and provided an overview of how multi-organisational emergency exercises and response evaluations are designed, implemented, documented, shared and used.

A key objective was to find answers to functional (how) questions. Such questions cannot be addressed without taking into account the context, as this influences both the evaluation process and its design. A more holistic approach was adopted to address RQ2 as different Safety Regions were analysed, while RQ3 is an example of an embedded case study. The latter examined one case (the Netherlands crisis management system) at multiple units of analysis (mayors and regional operational leaders). This case is described in more detail in section 3.3.

4.3 Research methods, approaches and activities

This section gives an overview of the methods used to address the four RQs and answer the main RQ. It is important to balance the use of qualitative and quantitative methods, as it can be argued that valuable data may be lost or ignored if the research focuses too much on one at the expense of the other. An imbalance could also limit theoretical development. The design approach encourages the use of multiple methods for gathering and analysing data, in order to gain a more rounded view of the area under investigation. In the present research, a scoping study (including snowballing), document analysis, in-depth (content) analysis, and expert judgement were used to elicit knowledge to support an exploratory and descriptive approach to RQ1 and RQ2.

Then, a survey experiment (RQ3b) tested concepts that were developed in the exploratory phase. In order to explore the expectations of crisis management professionals, the findings from RQ3 were complemented by a thematic analysis. Finally, conceptualisation was used in the later stages (RQ3 and RQ4) of the project to encourage creative and innovative thinking.

Figure 6 and Table 13 provide an overview of these methods and the empirical data that was collected, and relate them to the RQs.

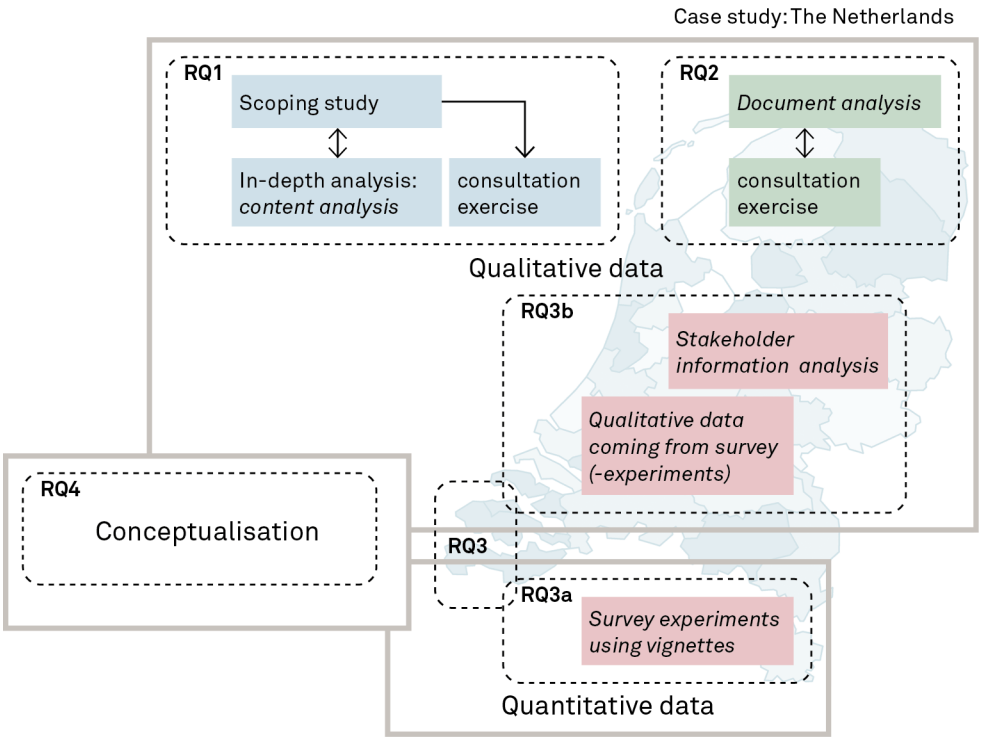


Figure 6: Visual overview of research methods and approaches
This figure presents the various RQ's and the methods and approaches that were used to address them. It also shows the mixed nature of this research and how the various methods and approaches relate to each other.

Table 13: Overview of RQs, methods and empirical data.

RESEARCH QUESTION (RQ)	RESEARCH METHODS AND SAMPLING	EMPIRICAL DATA
RQ 1 (Paper I): What is known about the evaluation of disaster management exercises (in the scientific literature)?	Scoping study In-depth analysis (i.e. science mapping, combined with a content analysis) Expert judgement (consultation exercise)	246 scientific articles (overall analysis) 43 scientific articles (in-depth analysis)
RQ 2 (Paper II): How are disaster management exercise and real-life response evaluations documented in the Netherlands?	Document analysis (content analysis) Expert judgement (consultation exercise)	62 documents (18 exercise evaluations, 23 systemic test evaluations, 21 real emergency response evaluations) 55 participants
RQ 3a (Paper III): (How) does the clarity of the presentation of the object (O), the analysis (A) and/ or the conclusion (C) in an evaluation description influence its perceived usefulness (P) for the purposes of: (i) learning and (ii) accountability?	Survey experiments using vignettes	84 participants
RQ 3b (Paper IV): What do crisis management professionals expect to find in a useful crisis management evaluation report?	A stakeholder information analysis including a thematic analysis of qualitative survey data coming from the factorial survey performed for RQ3.	84 participants
RQ 4 (Chapters 6 and 7): How should we design evaluations of simulated or real disaster responses (including the product) in order to make them useful and relevant to a variety of users?	Mainly conceptualisation	Findings from Papers I–IV.

4.3.1 The scoping study

Grant and Booth (2009) identified fourteen types of literature review and associated methodologies, and concluded that there is no internationally-agreed, coherent and mutually-exclusive categorisation. It is therefore up to the researcher to decide, as a function of the purpose of the research, whether a specific methodology is suitable or not. The scoping study applied in Paper I aimed to provide a comprehensive overview of research in the field of crisis management exercise evaluation, and suggest ways to improve it. This is in line with the general use of scoping studies, which is "to map the literature on a particular topic or research area and to provide an opportunity to identify key concepts; gaps in the research; and types and sources of evidence to inform practice, policymaking, and research" (Daudt, et al., 2013, p. 8). Thus, they are both broad, and provide a basis for identifying further research needs.

Although a scoping study shares several of the characteristics of a classical literature review (it is systematic, transparent and replicable) it differs in that it is broader in scope and characterises the quantity and quality of research. In particular, it does not seek to critically review the corpus based on analysis, synthesis and conceptual innovation. In this context, Grant and Booth (2009) state that scoping studies lack rigour and may foster bias. For example, they do not include a quality assessment process.

In order to mitigate these weaknesses, the 'six step framework' (Arksey & O'Malley, 2005) was developed as a research protocol. The aim was to ensure that results were obtained in a structured manner, and reproducible. The process begins with the research question, then describes how relevant studies were identified, selected, and how data were recorded. Steps five and six concern the presentation of results and a review of the protocol. In order to mitigate some of the weaknesses of the method, step six was complemented with a consultation exercise (see also section 4.3.3). Figure 7 provides a schematic overview of the protocol, including the six-step framework, and its application to this research.

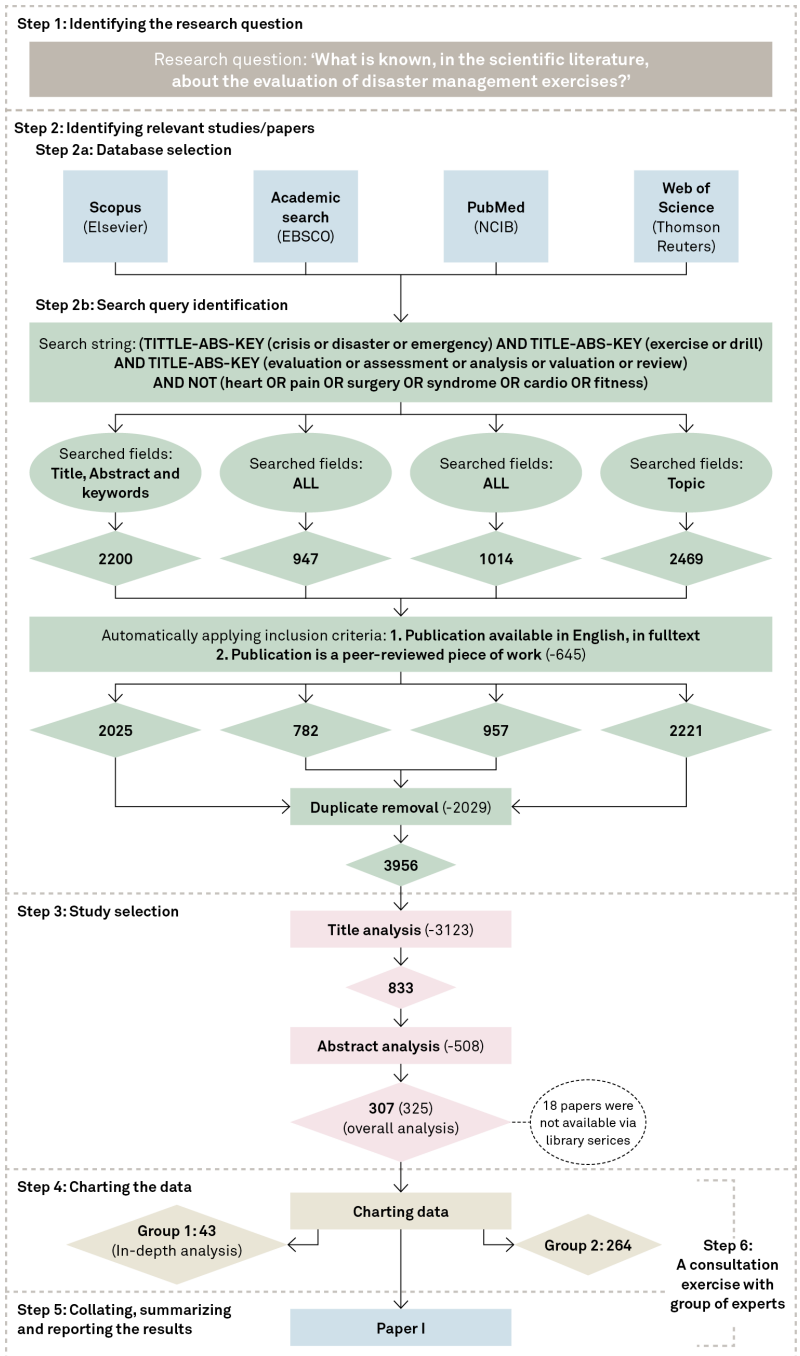


Figure 7: Schematic overview of the scoping study approach
The figure illustrates the application of the six-step scoping study framework used for this research and how the analysis leads to the results discussed in Paper I.

4.3.2 Document/ content analysis

Documents such as research articles, policy papers or evaluation reports are an effective means of gathering data that provides an understanding and develops empirical knowledge of a specific or generic situation. They can be said to be stable, as they are unaffected by the research process. An analysis of their content provides a basis for in-depth research regarding design and use (G. A. Bowen, 2009). Content analysis can be defined as "a systematic and replicable process, or technique, for organising many words or information streams into fewer content categories related to the central questions of the research and by explicit rules of coding" (Stemler, 2001, p. 5). It can reduce a large amount of text into fewer categories that have been specified in advance (Bengtsson, 2016; Bryman, 2016; Weber, 1990). Additionally, it can identify trends and patterns (Stemler, 2001). It should be emphasised that content analysis is more than a word-counting process (Downe-Wamboldt, 1992).

Paper I included an in-depth analysis of the study's core scientific documents. This sought to mitigate the weaknesses of a scoping study. The first step took the form of a network citation analysis (Lecy & Beatty, 2012) . The aim of this analysis was to determine the most influential studies (in terms of citations), and identify clusters of researchers who cited each other's work. The method is also referred to as 'science mapping', and basically consists of visualising bibliometric networks (Van Eck & Waltman, 2014). The second step extracted information related to the purpose, function and form of the evaluations described in the selected papers. The coding scheme was inspired by various frameworks, and the design science approach (section 4.1). Design aspects (purpose, function and form) were supplemented with more detailed aspects that formed the background for the coding scheme. The content analysis also investigated and identified opportunities for improvement.

Complementary to the scientific literature (Paper I), many evaluation documents are produced in the course of day-to-day professional practice. The purpose of performing a content analysis of evaluation reports in Paper II was to increase knowledge relating to how crisis evaluations are performed and reported in practice, and to identify whether they actually meet their intended purpose. These documents, requested in December 2016, are mostly in the public domain, and provide background and contextual information on how evaluations were organised and executed. Their analysis made it possible to develop an overview of emergency and crisis management evaluation practice in the Netherlands. Documents related to emergency responses were also included. These emergency response evaluations are very useful when developing disaster exercise evaluation practice and ideally, there is a cyclic relation.

G. A. Bowen (2009) notes that both content and document analyses have their own limitations. Firstly, many documents are produced for a purpose other than research. They are created independent of a research agenda and might not provide sufficient detail to answer the research question. Thus, the provided documents were not considered to be an accurate or complete record of events, whether incidents or exercises. Instead, they were viewed as containing background information on how different types of formal evaluations and systemic exercises are currently performed and documented. They offered a valid starting point for analysing evaluation practice, and developing discussions based on its design and use. Secondly, access can be a challenge. However, in the context of this study, access was provided through institutional networks.

Paper II builds upon the aspects of an evaluation (its purpose, goals, analysis, context and scenario) identified in Paper I. This material was complemented by process-related information, such as users and stakeholders, completion dates and follow-up. The consistent use of the coding scheme helped to develop an understanding of current evaluation design, and identify common approaches in Dutch practice.

An overview of the documents that were provided and analysed is presented in Table 10. In some cases, a Safety Region provided more than one document of a particular type. In these cases, the table of contents of selected documents was compared with previous versions in order to identify any structural differences. In other cases, additional supporting documents provided more detailed information or a summary (e.g. a policy or factsheet). This information was integrated into the regional-level analysis. A final consideration was that an incomplete collection of documents might lead to selection bias. A few regions did not provide the full range of documents. In order to reduce bias and limit the weaknesses of the method, the preliminary findings of the document analysis and the general overview were cross-checked and discussed with a group of experts from the Safety Regions (see section 4.3.3). This group session contributed to developing a more reliable interpretation of the data.

4.3.3 The consultation exercise (expert judgement)

Both the scoping study (Paper I) and the document analysis (Paper II) were complemented by expert judgement sessions in the final stages. These sessions resembled focus groups (Kitzinger, 1995; R. A. Powell & Single, 1996). However, participants were not subjects who were used to generate or collect new data; instead they were consulted to verify, complement and comment on preliminary research findings based on their professional experience.

Following Hughes (1996), an estimate was developed based on the experience of one or more people who were familiar with the object of the research. The group was invited to discuss and comment on the outcomes of the research from the viewpoint of their professional experience. This step was included as both primary methods (the scoping study and the document analysis) can be subject to interpretation or selection bias. Involving active professionals can mitigate these biases. In addition, the sessions were an opportunity for participants to interact with each other, which is not possible in a one-to-one interview. This characteristic is a key advantage of group sessions (Kitzinger, 1995). Sessions were guided by the preliminary findings of the study that served as a basis for the discussion. To combat the risk that the debate might become irrelevant, or that one or more participants dominated, the researcher facilitated each session or instructed other facilitators to intervene, if necessary.

With regard to Paper I, the group consisted of seven people from the Netherlands with a background in research/ academia (public administration, organisational science and medicine). Participants were familiar with the key topics of the research (disaster management and/ or evaluation). Five had carried out incident and/ or exercise evaluations in the Netherlands. Two had carried out other types of evaluation (policy evaluations or evaluative research).

As for Paper II, the preliminary findings and general overview were cross-checked and discussed by a group of representatives from the Safety Regions[9]. Here, the aim was to gather complementary information and develop an overview of multi-organisational crisis management evaluation practice in the Netherlands. Initially, statements derived from the preliminary findings were checked and discussed with all participants using a digital tool. Then, experts and professionals responsible for evaluation discussed the results in smaller group sessions and provided additional background information. The aim was to reduce investigator bias and verify that the material in the documents matched general views of practice.

4.3.4 Survey experiments

The second phase adopted a more quantitative approach. Here, the focus was on the extent to which professionals believe that a specific format of evaluation description enhances its ability to achieve its purpose. Survey questions are often used to capture respondents' opinions. However, techniques that involve asking direct questions can

[9] Fifty-five participants attended an afternoon session at the Netherlands Institute for Safety. In addition to representatives from 16 of the 25 (64%) Safety Regions, crisis management partners and other stakeholders from the police, the coastguard, the Rijkswaterstaat, the Ministry of Security & Justice, the Inspectorate and consultants participated.

pose methodological problems, with it remaining unclear whether the results obtained reveal respondents' true opinions or whether they simply reflect desirable answers. Thus, it is best to avoid single-item questions. Instead, a survey experiment approach was applied. The combination of survey and experimentation is the key characteristic of the survey design, which optimises the advantages of both approaches. Experiments increase internal validity, while survey studies increase the generalisability of results, increasing external validity.

Auspurg and Hinz (2015) note that a situational description that manipulates various dimensions leads to more subtle questioning and, therefore, responses are less likely to be influenced by social desirability bias. In addition, a more detailed description of a (real-world) situation helps to standardise stimuli and provides deeper insights into respondents' judgements.

Vignettes

Vignettes were a crucial element in the survey experiment. Vignettes are "short, carefully constructed descriptions of a person, object, or situation, representing a systematic combination of characteristics" (Atzmüller & Steiner, 2010, p. 128). The vignettes that were used in the survey experiment were based on actual evaluation reports (used in Paper II). A key element was that they were clear and consistent, and that they contained the most crucial components, namely: purpose (P), object (O), analysis (A) and conclusion (C) (see Table 14).

O, A and C were developed as factors and experimentally manipulated. The number of factors and levels meant that the total vignette population was too large and the vignettes were too long for each one to be judged by each respondent. Thus each participant was only exposed to a random subset, resulting in hierarchical data.

Participants were asked to read and rate the vignettes that were presented to them, focusing on their perceived usefulness with respect to the purpose of learning or accountability, while individual components were either presented in a *clear* (1) or *unclear* way (0). Thus, some documents contained very clear descriptions of O, A and C, while others did not address them explicitly (Table 14). If the way O, A and C are expressed does influence the perceived usefulness of an evaluation description, it is reasonable to expect that such a manipulation would be detected. Perceived usefulness was defined as 'the extent to which professionals believe that a specific form of evaluation description enhanced its ability to achieve its purpose'. The overall hypothesis was that variation in the clarity of O, A and C would influence the usefulness of an evaluation description with respect to a specific purpose (P).

Table 14: Overview of vignette components (O, A and C), including examples.

COMPONENT (INCLUDING A GENERAL DESCRIPTION)	CLEAR DESCRIPTION (1)	UNCLEAR DESCRIPTION (0)
Object description (O) What or who is evaluated? This key element contributes to the response in a specific DRM context. It could be an individual, part of an organisation, an entire organisation, or even multiple organisations operating together. An important aspect is the relationship between the object, and the context or scenario.	A clear object description should clarify: (1) Who or what is the subject of the evaluation? *The Operational Leader (OL) within the Regional Operational Team (ROT).* (2) Why does the object exist? What is its role and responsibilities? *The OL is (ultimately) responsible for the process of decision-making within the ROT.* (3) What can be expected of the object? What is its function? *The OL ensures that data are collected and shared, the situation is judged, and a well-founded decision is made, shared, and documented.*	*(1) The object (the OL) was not specifically mentioned as being the focus of the exercise evaluation, and other actors were also introduced.* *(2) No details regarding the role and responsibilities of the OL were provided, and only a generic description of the response organisation was given.* *(3) No details were provided regarding what to expect from the OL, and a generic description of the scenario was repeated.*
Analysis (A) The analysis supports the arguments put forward. What happened during the exercise and why?	A clear analysis should indicate: (1) How information was collected. *(Observation) notes and reports of the evaluator, meeting reports, and other documents and data.* (2) What (value-free) results do this yield? *A (f)actual description of the decision-making process was detailed.* (3) The action(s) performed by the evaluation object: *The OL took a number of decisions with regards to the incident and informed the mayor.*	*(1) No data collection methods were described; only a general description of the response to the incident was provided.* *(2) No results were presented and, again, only a generic description of the response to the incident was given.* *(3) No specific actions (regarding O) were mentioned, only a generic description of the incident was provided.*
Conclusion (C) The conclusion determines value or performance. How (well) did the object perform? It is the logical consequence of the P, O, A chain, and indicates how a judgement is reached.	A clear conclusion should: (1) Give an opinion on the functioning of the evaluation object: *The OL correctly implemented the three-phase decision-making process.* (2) Judge whether the evaluation object has fulfilled its purpose (see also the object description) in the described context: *Although the process was correctly executed, some incorrect decisions were made.*	*(1) No opinion was given, only a generic description of the exercise setup and an explanation of the response.* *(2) No judgement was formulated and the emergency response process was only vaguely presented.*

Participants

Two groups of professionals who use evaluation descriptions in their day-to-day work were studied. The first group consisted of people holding a 'governing' (responsible) position such as mayors (N=34), and the second group were people in operational (executive) positions, such as operational leaders (N=50). These groups are clearly identifiable within the Dutch crisis management structure and have different, but closely related, roles and responsibilities (see 3.3.2). Safety Region directors were also invited, and a fourth group of 'other' respondents was identified. An overview is presented in Table 15.

A qualitative analysis showed that the roles of this fourth group were primarily operational. Examples include (municipal) crisis management advisors, incident commanders, preparedness experts, emergency planners and crisis coordinators. They were included as their roles indicated that they might be users of evaluations and, therefore, able to provide input to this research. Participants were contacted via email and online community newsletters through their respective national networks. All correspondence stated the purpose of the research and included a link to the survey.

Table 15: Overview of respondents and the number of participants.

ROLE	N.	%	GROUP	N	%
Regional Operational Leaders (ROL)	39	46.4%	*'Operational' (=ROL + 'other')*	50	59.5%
'Other'	11	13.1%			
Mayors	28	33.3%	*'Governing' (=Mayors + Directors)*	34	40.5%
Directors of Safety Regions	6	7.1%			
Total	84	100%	Total	84	100%

Analysis

Analyses were run using IBM SPSS Statistics for Windows, version 25 (IBM Corp., 2019). Data related to participants' background, experience, the Safety Region they worked in and their use of evaluations. In order to verify whether the vignettes reflected a realistic scenario, respondents were asked to rate them. Given the nested structure of the data, various multilevel models were estimated. As it was reasonable to expect a positive relation with prior use of evaluation descriptions, data were controlled for fixed effects. Similarly, as the respondent's background (Operational or Governing) was also expected to affect the results, it was added as a control variable.

4.3.5 Stakeholder information analysis

Paper IV adopted a more qualitative and open approach in an attempt to involve stakeholders. Here, the aim was to examine the feasibility of various data collection approaches, or aspects of performance that were thought to influence the selection of design variables, without selecting a specific design. A second objective was to reveal any biases or dissatisfaction with respect to earlier designs that could be used to improve future designs. Attention to, and the involvement of key stakeholders was presumed to enhance the design and implementation of evaluations and the use of their results. It is important to identify the expectations of the target audience in order to increase the usefulness of evaluations. Thus, a stakeholder information analysis was used (Lawrence & Cook, 1982). This approach consisted of three steps: (1) identification and selection of stakeholders; (2) accessing and surveying them, and (3) analysing responses.

Step 1: Stakeholder identification and selection

In the first step, key users of crisis management evaluations were identified. This group included mayors, directors of safety regions and regional operational leaders.

Step 2: Accessing and surveying stakeholders

Because of similarities between user groups (see 3.3.2 and 4.3.4) and the objectives of the research, it was decided to supplement the survey experiment with open survey questions. It should be noted that the key questions, designed to identify the expectations of participants, were asked at the beginning of the survey experiment in order to avoid bias due to the use of the vignettes. Together, the survey and related questions can be seen as the second step in the stakeholder information analysis. In addition to the key questions presented initially, open-ended questions were asked throughout the remainder of the survey (see Table 16 for a full overview).

Table 16: Overview of qualitative, open-ended questions used in the survey experiment (translated from Dutch).

NUMBER	FUNCTION	PHASE OF THE SURVEY	QUESTION
Q5	Main question used to obtain information regarding the user's expectations.	Part 2: Current use of evaluations	Can you describe your expectations of crisis management evaluations based on your (operational) role/ position?
Q12, Q16, Q20, Q24	Control questions used to verify and complement the expectations expressed in answer to Q5. These answers could be biased as respondents had been exposed to other information as they completed the full survey.	Part 4: Evaluation descriptions	Can you provide feedback regarding the evaluation description (that was presented to you)?
Q28	Control question used to verify and complement the response to Q5.	Part 5: Optional feedback	You have read four evaluation descriptions. Are there any other (crucial) components that you think are missing?
Q29	Control question used to verify and complement the response to Q5.	Part 5: Optional feedback	Do you have any additional feedback or questions that you could not, or have not, mentioned elsewhere?

Step 3: Analysis

Qualitative data collected in response to the questions illustrated in Table 16 were used to check, clarify and identify any other expectations. The third, and final step, the analysis, used Atlas.ti software (Scientific Software Development, 2019) and applied a general inductive analysis (D. R. Thomas, 2006). The text was coded according to a number of themes related to the terminology used by respondents. This helped to identify both expected and unexpected expectations, while reducing the total amount of data. The overall analysis consisted of three steps.

In the first, raw data from Q5 was segmented into categories and themes by an independent analyst. This was to ensure that codes were based on the terminology used by respondents. It should be noted that stakeholders typically have diverse and often competing interests, which makes it hard to address all expectations equally well, or judge their importance. Therefore, the second step consisted of linking codes to concepts such as POAC and (user) design, and reviewing the results of the first step. This step reduced the amount of data by distinguishing between main and subcodes, and checking their relevance to theoretical concepts. The final step involved selective coding. Recurrent themes were selected and analysed with respect to how they related to one another.

The information that emerged from the stakeholder information analysis was judged to be likely to affect the design of the evaluation, notably underlying components such as the object's performance, or the selection of data collection methods. The analysis also highlighted biases or concerns based on experience with previous evaluations, such

as a time lag between initial and follow-up data collection, or the representativeness of data. Here, the findings should be seen as reflecting generic expectations based on a limited sample of crisis management professionals.

4.3.6 Conceptualisation

Conceptualisation, as it is used here, is not a method *per se* and should rather be seen as a broad and holistic research activity. It supports creative and innovative solutions, unlike empirical and related methods. Inspired by Savoia, et al. (2014), the process of conceptualisation was explored and partially applied in the later stages of this research.

The process begins with developing an understanding of the situation or problem. In this research, this was established via RQ1 and RQ2, which helped to identify a typical format used to document, transfer and present information from exercise evaluations. An even better understanding can be gained by identifying patterns or connections, and the key underlying properties (components) of concepts. The process can be seen as a form of conceptual thinking (MacInnis, 2011). It requires logical reasoning and argument with regards to concepts in order to identify, develop, or propose what is termed a 'conceptual framework'. The concepts that constitute such a framework support one another, articulate their respective phenomena, and establish a framework-specific philosophy. Such a framework can be seen as a network, or plane of interlinked concepts that together provide a comprehensive understanding of the phenomenon or problem.

In this research, the process led, *via* RQ3, to the creation of the concept of an evaluation description, which contains four elements: Purpose (P), Object description (O), Analysis (A) and Conclusion (C). These components provide a logical foundation for understanding the why-what/ who-where/ and when-how questions that frame an evaluation (Heath, 1998). For example, conclusions (C) make no sense unless the object (O) and how the analysis was conducted (A) have already been presented. For more details regarding the concept, see Paper III.

Conceptualisation was also used to create the various models and frameworks that are presented in the Discussion of this thesis. These models and frameworks can be related to RQ4. In addition to the practical experience of the researcher, qualitative information collected from crisis management professionals was used to identify the specific elements that must be considered when designing a disaster exercise evaluation. The conceptualisation applied here can also be seen as the synthesis of previous research methods, and is reflected in Chapter 6. It contributes to the understanding of evaluation in the context of a (simulated) disaster response, and suggests improvements

for its usefulness. It can also be seen as a point of departure for other researchers and professionals.

4.4 Methodological reflection and research quality

This research can be described as mixed, with respect to the basic types of research (Kothari, 2004). RQ1 and RQ2 focused on an analysis of the scientific literature and evaluation documents, and could be seen as *analytical*. RQ3 and RQ4, on the other hand, sought to improve the usefulness of evaluation documents; thus they are more *descriptive*. Both *qualitative* and *quantitative* methods were applied. RQ1, RQ2 and RQ3b used qualitative methods to gain a broader understanding of evaluation in the context of DRM. However, RQ3a required a more quantitative approach, in order to identify how people rate evaluations. The latter approach identified the range of factors involved, and allowed their influence on the usefulness of evaluations to be measured. Furthermore, it can be defined as *applied* as it was not aimed at natural phenomena or gathering knowledge for knowledge's sake, but at the practical problem of investigating evaluation practice in order to improve its usefulness.

The study is both conceptual and empirical. For example, the introduction of the concept of an evaluation description can be seen as conceptual, as can the models that are outlined in the Discussion (Chapter 6). Nevertheless, these concepts are built upon empirical data that was gathered in earlier phases (RQs 1–3).

The quality of any research effort merits further consideration. The concepts of *validity* and *reliability* are commonly used to evaluate quality. Validity refers to the extent to which the results are credible. This project evolved as an iterative and sequential process, guided by earlier findings. Methodological triangulation helped to overcome bias inherent in a single approach, and added value to the theoretical debate. A multi-method approach is known to increase the validity and reliability of data (Hammersley & Atkinson, 2019).

The following sections outline the primary threats to the validity of this research, and how they were mitigated. The various methods served different, but complementary purposes and, in doing so, also partially overcame their respective limitations.

4.4.1 Validity

Validity indicates the accuracy of a measure. Various types can be evaluated, through either expert judgement or statistical methods. Here, the focus is on three types:

construct validity, internal validity and external validity. *Construct validity* (Kothari, 2004) addresses the relation between underlying theory and observations. It looks at the choice and collection of measures used to examine the studied concepts in order to determine whether they are suitable. Construct validity can be improved by ensuring that indicators and measurements are carefully developed from existing knowledge. It can be increased using strategies such as: (1) multiple sources of evidence; or (2) establishing chains of evidence (Runeson et al., 2012).

RQ1 used the scoping study method to identify relevant literature. Here, construct validity can be undermined by factors such as the search string, which must correspond to the terminology used. This threat was mitigated using various strategies, which included: using digital software tools; adopting a systematic approach based on a six-step framework; consulting experienced librarians about relevant search strings; and combing various databases. In addition, snowball sampling was applied, and the results were cross-checked, followed by a final consultation exercise with external experts to validate the method and outcomes.

RQ2 used a document analysis as the primary research method. Here, construct validity is threatened by the analysis of contents. This threat was mitigated by considering the corpus as an imprecise, inaccurate or incomplete record. The coding scheme was inspired by various evaluation frameworks and the theoretical foundations were drawn from the design science approach. At a later stage, the preliminary findings were validated by a representative group.

For RQ3a, a survey experiment was used to investigate perceived usefulness. Here, threats to construct validity concerned the instrument. Therefore, texts were derived from real documents and pre-tested by students and colleagues. These tests sought to verify that the aspects under investigation (P, O, A, C) were present (or not), and whether they were clear or unclear. The qualitative feedback was used to refine the descriptions. RQ3b was primarily aimed at collecting expectations. Here, content validity played a role in the thematic analysis. It was important that the themes that emerged adequately reflected professionals' expectations. This point was addressed by asking an independent analyst to combine data from various related survey questions into categories and themes. The resulting codes were linked to theoretical concepts and reviewed by various researchers.

Internal validity is a second important type of validity (Kothari, 2004). This concept relates to issues that may affect any causal relationships between the treatment and the outcome. Techniques such as randomisation, random selection, blinding, experimental manipulation, and the use of a strict protocol can be applied to improve it. Most of the RQs (1, 2, 3b, 4) in this research are exploratory and do not make any causal claims.

Moreover, confounding factors are not considered to be a significant threat. Combined with the absence of causal claims, it is reasonable to consider that there are few threats to the internal validity of this work.

However, RQ3a assessed the perceived usefulness of evaluation descriptions, based on a survey experiment. Here, it is important to reflect more closely on internal validity. In the experimental setting, internal validity refers to whether the effect is caused by the independent variables, or other confounding factors. Here, internal validity was established by the manipulation, which sought to assess the causal effect of the vignette on the outcome variable. The method assumes that effect estimates are free of any bias (Atzmüller & Steiner, 2010). The eight vignettes reflected the content of current evaluation reports, and were developed to accurately reflect the key experimental factors. Participants were not informed about the intervention to avoid bias. Respondents volunteered to participate, and vignettes were randomly assigned to respondents, which increased internal validity. The statistical analysis applied multilevel modelling with a random respondent effect. Control variables were used to take account of the possibility that previous use of evaluations might affect the outcome. The nested data structure was considered. Together, these precautions ensured internal validity was as good as possible, and threats were limited.

External validity concerns the ability to generalise findings beyond the scope of the study, as the results obtained in a specific context may not be valid in other contexts. In this research, it is considered to be more important than internal validity. The reason for this is that qualitative studies such as this one must rely on analytical generalisation, unlike quantitative studies that can apply statistical and sampling strategies. Generalisation is enhanced by a comprehensive and realistic context description, and clear inclusion and exclusion criteria. Other strategies include studying multiple cases and replicating experiments.

Starting with RQ1, systematic inclusion and exclusion criteria were used to select relevant studies and software was used to identify bibliometric networks. RQ2 and RQ3 investigated the Netherlands context, and it would be interesting to evaluate their relevance to other nations and organisations (generalisability). However, the selection of one country supported a comprehensive focus on most of the relevant actors. In addition, it could be argued that the various Safety Regions constitute multiple units in the Netherlands case. Finally, different emergency management systems have certain similarities, in particular with regard to the use of evaluation.

The Discussion (Chapter 6) examines the broader applicability of the study's findings in greater detail, and highlights their generic nature. A key point to note regarding external validity is the applied nature of this research, which uses real(istic) examples

throughout all of the RQs. This increases psychological realism, which is a threat to external validity. In particular, in RQ3 participants were provided with contextual information from a realistic scenario, and vignettes were constructed using text from real evaluations.

4.4.2 Reliability

Reliability is almost inseparable from validity. It relates to whether the same outcome would be obtained if the study is repeated by another set of researchers, both in terms of data collection and analysis. Reliability is an indicator of the consistency of a measure, and should be ensured throughout the data collection process (Kothari, 2004). Methods must be applied consistently and the conditions for the research must be standardised to reduce the influence of external factors.

As much of this study is qualitative, and thus relies on the interpretation of the researcher, exact replication is difficult. However, particular attention was paid to making replication as easy as possible. In all cases, protocols are provided that structure data collection and analysis processes (see appended papers in Annex E) and a detailed description of how the study was executed is given. These protocols were discussed with methodological experts and are intended to enable other researchers to replicate the study in question, and identify and investigate any shortcomings.

However it should be noted that replicating the study with the same sample might lead to learning bias. Other reliability measures are discussed in detail in the appended papers. For example, papers related to RQ1 and RQ2 focused on reducing the influence of individual researchers (selection and researcher bias). The preliminary findings of the scoping study and the document analysis, and the general overview were reviewed by a group of experts (RQ1) or representatives from the Safety Regions (RQ2). Throughout the project, there was close consultation with various experts. Other researchers also contributed. For example, for RQ1 a colleague was asked to review the selection of articles. Any differences were discussed, and the final result was agreed between them. For RQ3b, an independent analyst who was initially unfamiliar with the theoretical concepts was asked to review the coding. This was to ensure that codes were based on the terminology used by respondents. Furthermore, all papers were reviewed by at least one other researcher. Despite these measures, the qualitative nature of this study means that other researchers may encounter other challenges. Finally, it should be noted that normative questions such as RQ4 do not have an unambiguous answer and that any outcomes presented here may need to be revised in the light of new knowledge.

5 Key findings

The four articles appended to this thesis address three of the four sub-questions (see section 1.3). All four seek to enhance our understanding of evaluation in the DRM context. However, each paper makes its own contribution. This chapter highlights the key findings. Outcomes are summarised in more detail in Annex D and form the basis for the synthesis that is presented in the next chapter.

5.1 **Paper I**: Scoping the field of disaster exercise evaluation – A literature overview and analysis

This explorative paper addresses RQ1: *What is known about the evaluation of disaster management exercises in scientific literature?* It aims to provide a comprehensive overview of the scientific literature regarding the evaluation of DRM exercises. More specifically, it maps the disaster exercise evaluation literature, in order to identify key concepts, gaps, and types and sources of evidence that can inform both practice and research. The overall analysis is presented in paragraph 4.1. of the paper in Annex E. The key findings are discussed below.

***Key finding 1**: Four research groups, four foci.*

The scoping study identified 43 papers that met the inclusion criteria and specifically examined the evaluation of exercises in a DRM context. A citation analysis distinguished four groups that built on each other's work, or at least cross-referenced each other. This was visualised (using a citation analysis) as clusters or nodes (papers) that were more highly connected to each other than the rest of the network. A further content analysis showed that they all had slightly different foci. These findings are illustrated in Table 17.

Table 17: The four research groups and their foci.

GROUP	AUTHORS	FOCUS	COMMENTS
I	Key authors: *Biddinger, Savoia and Agboola* (Agboola et al., 2013; Biddinger et al., 2008, 2010; Kaji & Lewis, 2008; Morris et al., 2012; Savoia et al., 2009, 2010, 2013, 2014).	Developing a **framework** for performance evaluation in the area of public health.	Research focuses on all aspects of evaluation—from measurement criteria and tool development, to post-action review analysis and lessons learned.
II	Key authors: *Rådestad and Rüter* (Djalali et al., 2014; Nilsson & Rüter, 2008; Rådestad et al., 2012; Rüter et al., 2006)	Developing and implementing performance **indicators**	Indicators are developed specifically for the evaluation of medical responses and, in particular, hospital preparedness.
III	(Cranmer et al., 2014; Ingrassia et al., 2010; Klein et al., 2005; T. L. Thomas et al., 2005)	**Data collection** during evaluations of medical management in mass casualty incidents	Data collection by developing observation tools, or training observers in medical management evaluation.
IV	*Sinclair, Latiers and Wybo* (Kim, 2013; Latiers & Jacques, 2009; Lonka & Wybo, 2005; Sinclair et al., 2012; Wybo, 2008)	**A broader, methodological perspective** on the subject of disaster exercise evaluation	Overviews based on other literature and research, for example accident investigation, in order to understand and assess disaster exercises in general and, thus, be able to evaluate them.

Key finding 2: *Single/ stand-alone cases lack systematic efforts to build upon each other, hampering lessons identified becoming future lessons learned.*

Although it was possible to distinguish four groups that cross-referenced each other, many contributions focus on single/ stand-alone cases. They do not build upon each other, meaning that suggested methodologies are not further tested or used. To illustrate this stand-alone aspect, more than 50% of papers did not refer to previous exercises or incidents. This highlighted a lack of systematic effort to build on existing work, for example, by assessing whether lessons identified had become lessons learned, or building a solid knowledge base using meta-evaluations.

Key finding 3*: Little, or no evidence of the effectiveness of the methods applied.*

The in-depth analysis sought to identify elements such as purposes, contexts, data collection methods, functions and timing. Various papers introduced methods, tools, process models or performance measurement frameworks. Most refer to a variety of general approaches, such as the United States HSEEP Guidelines (United States Departement of Homeland Security, 2020) or evaluation research (R. R. Powell, 2017) using a variety of standard qualitative and/ or quantitative, social science data collection methods. The latter mainly deal with methodological concerns, such as threats to reliability and validity. While several papers made an explicit link with supporting learning/ development, they did not specify how the selected methods support this. In

general, most papers provide no evidence of the effectiveness of the evaluation methods that were applied, making it difficult to determine if they were more or less likely to achieve their intended purpose. This lack of clarity makes it difficult to determine the value or utility of a method in a broader context.

Key finding 4: *A field that is gaining interest.*

The scoping study also mapped the broader literature. This found that most papers were published in the area of medicine, followed by the social sciences and engineering. This was also reflected in the scenarios that were presented. Mass casualty incidents were frequently used to evaluate a variety of objects or artefacts. Evaluation is often understood as: testing plans, processes and procedures; assessing the performance of tools or systems, equipment and personnel; enhancing or improving awareness; and, by identifying gaps or areas of limited capability, supporting the future development of understanding and knowledge through training activities or preparedness programs. It is often briefly touched upon, mainly as a final step in the design process, within the broader context of DRM exercises. Information is presented in a descriptive format, rather than detailing and prescribing how evaluations should be, or were, undertaken.

5.1.1 Contribution

Overall, the data showed that the broader field of disaster exercise evaluation has received increasing attention from researchers in recent times, although the literature remains limited compared to the overarching topic of disaster preparedness. There is little evidence regarding how evaluations are conducted, how and why they are used to achieve their purpose, and their effectiveness. In particular, usefulness is currently under-investigated. This also applies to their value or utility, which can be seen as a key point in both evaluations and the methods used. For example, relevant and robust arguments must be put forward to mitigate threats to reliability and validity. Furthermore, there appears to be a lack of enthusiasm for building a solid knowledge base that can support future developments both with regard to evaluation design and improving disaster preparedness and response. Studies in this field need to become more cohesive and build on each other's work. Empirical evidence should be used to support claims regarding usefulness, and the construction of a knowledge base.

The present research underlines that any claims regarding the effectiveness and usefulness of specific methods or approaches, and/ or how to improve them, requires more empirical data and/ or logical reasoning, with a specific focus on reporting. More precisely, the scientific discourse would benefit from using ideas and concepts from

evaluation research and design science. This could be achieved by providing greater clarity regarding: (1) the purpose and context in which a specific evaluation method is expected to be used; (2) what the method needs to do (or produce) in order for it to fulfil the purpose; and (3) how the method achieves its goal, and thereby fulfils its purpose.

Finally, this study identified a gap in knowledge regarding the value of exercises for professionals. Little is known about the impact of a disaster exercise on operational preparedness, and how this can best be evaluated. This suggests that evaluations are currently mainly designed using trial and error.

5.2 **Paper II**: Does the means achieve an end? A document analysis providing an overview of emergency and crisis management evaluation practice in the Netherlands

The second paper is also explorative, but more practice-oriented. It seeks to gain a better understanding of how exercise evaluation is currently executed in professional practice. The study aimed to increase knowledge related to how evaluations are performed and reported on in practice, and whether they meet their intended purpose. Although not explicitly mentioned in the paper, the study addresses RQ2: *How are disaster management exercise and real-life response evaluations documented in the Netherlands?* The paper provides a comprehensive analysis of the current state of evaluation practice. More specifically, it presents an overview of how multidisciplinary emergency exercises and response evaluations are designed, documented, implemented and used in the Netherlands (see section 4.3.2).

The study was based on an analysis of evaluation reports and supporting documents that are prepared by the Safety Regions within the Dutch crisis management context (see section 3.3.2.). These documents offer a valid starting point for developing an overview of emergency and crisis management evaluation practice. In order to build upon the research presented in Paper I, the framework was inspired by the in-depth analysis outlined in that paper (see section 4.3.2) complemented by process-related information that reflected the temporal progression through the design phases: (1) starting the evaluation; (2) executing it; and (3) finalising it.

Key finding 1: *Various contexts, various (main) purposes and uses.*

In general, exercise and response evaluations are seen as supporting either learning or developmental purposes. An exception, in the Netherlands context, are systemic exercises. The latter are used to report or test the extent to which the respective Safety Regions meet their legal obligations. Although they can be framed as serving learning purposes, their overall aim is related to summative evaluation, i.e. accountability. This is also reflected in the supporting documents, which seek to measure and assess whether legal requirements or standards are being met. A similar framing issue applies to some real emergency response evaluations. Although they state that they are used for learning purposes, they can be perceived as serving the purpose of measuring the actions of those held accountable.

***Key finding 2**: A lack of detail regarding data collection and evaluation methods makes it difficult to assess the quality (credibility, value, accuracy), usefulness and effectiveness of evaluations.*

Evaluation reports pay little attention to describing or justifying data collection and/ or evaluation methods. Few details are given regarding the design and execution of evaluations, and reports do not link this design to a specific purpose. Documents also rarely provide details of the methodology and its limitations, which makes it difficult to assess quality. If information is provided, it is mostly very generic, and describes the evaluation process, or parts of it. A variety of designs and data collection methods are employed on a regional level. For both exercises and real events there is no common framework to ensure that the intended purposes are met and facilitate the sharing of findings. The exception is systemic test exercises, where guidelines and requirements provided by the Inspectorate support the design phase.

Observations were a commonly-used data collection and evaluation method. The use of observation as a primary data collection method can be challenging, particularly if it remains unclear how conclusions are derived, or if observers lack the appropriate competences or experience. The analysis highlighted that data collection lacks transparency, and the underlying reasoning was unclear. Both of these elements are crucial, and weaknesses can have a serious impact on the credibility, value, accuracy and usefulness of the final document. Not to mention validity challenges from the end user.

***Key finding 3:** Evaluations are seen as stand-alone or independent, and do not built upon each other to ensure that lessons identified become lessons learned.*

The analysis found that each evaluation tends to be seen as an independent activity. Efforts do not build upon each other. For example, evaluation reports do not necessarily draw attention to lessons that were identified in one exercise as potential indicators of systemic errors across a range of similar scenarios, exercises or events. In addition, many reports provide no recommendations or insights into why, what and how improvements can be made, and in what context. Most are independent, single case evaluations. This may limit their applicability to the national context, and hamper the development of crisis management systems in the Netherlands.

5.2.1 Contribution

This study supports some of the theoretical findings reported in Paper I, with a practical perspective. Professionals lack guidance on the design of an evaluation, and the many variations make it difficult to justify the investments that are made. Thus, it is difficult to determine how effective current evaluations are, or verify their quality, usefulness or effectiveness. It is unclear how they contribute to the development of preparedness or responses to future crises. Current evaluations are more-or-less independent activities, and a more holistic, strategic view is needed to support development and learning.

This study identified weaknesses in the Dutch approach that may have implications for broader, nationwide learning and the ongoing development of the country's crisis management system. The limited links, where they exist at all, between the various types of evaluation, exercises, systemic tests and real responses limit the extent of any broader learning.

On a more detailed level, this study showed that it is unclear how the ease of use of an evaluation product can be quantified. This problem is a challenge to providing the user with information in the most accessible format. Evaluation reports fail to provide clear information regarding the structure of the analysis, notably the framework for the investigation, and this has implications for the reliability and validity of their findings. For example, if there is no clear framework to support normative judgements, it is difficult to determine, and justify, what is 'good' or 'bad' in any particular circumstance. It is important for both users and evaluators to be aware of biases, and to mitigate them.

Finally, the presentation of evaluation outcomes should be better-matched to the needs of users and stakeholders; this may require a new or improved approach. The study underlines that, despite current efforts, there is much to be learned about improving the credibility and usefulness of evaluations.

5.3 **Paper III**: How can we make crisis management evaluations more useful? An empirical study of Dutch evaluation descriptions

The third study can be seen as causal or explanatory research, with some prescriptive aspects. In part, it addresses RQ 3, *What makes evaluations (descriptions/ texts) more or less useful to professionals?* Paper III presents a quantitative investigation of the relationship between the usefulness of evaluations for learning and accountability purposes, and some key components. The notion of the evaluation description (see section 4.3.4) was introduced, which encompasses four components: Purpose (P), Object description (O), Analysis (A) and Conclusion (C). The latter are logically connected, and assumed to influence the usefulness of an evaluation. More specifically, the study aimed to gain insights into whether O, A and C influenced the usefulness of the report with respect to P.

It addressed the following sub-question, RQ3a: *(How) does the clarity of the presentation of the object (O), the analysis (A) and/ or the conclusion (C) in an evaluation description influence its perceived usefulness for the purposes of (i) learning and (ii) accountability?* A survey experiment was developed using vignettes (see section 4.3.4.). More detailed information can be found in paragraph 4 of Paper III (Annex E).

The three independent variables were O, A and C. P was implemented as the dependent variable. Analyses were run using IBM SPSS Statistics for Windows, version 25 (IBM Corp., 2019). Given the nested structure of the data, various multi-level models were estimated. This resulted in the following findings:

Key finding 1: *The clarity of the conclusions (C) has a significant effect on perceived usefulness for both learning and accountability purposes, and operational and governing users.*

It appears that judgements regarding usefulness are based primarily on the final outcome (i.e. A and C) rather than O. How the conclusions are presented have a significant impact for both learning and accountability. If the conclusion is unclear, the evaluation is less useful for both learning and accountability.

Key finding 2: *The analysis (A) was significant for learning purposes, with a marginally significant effect on accountability.*

How analyses are presented was found to have a significant effect on learning. A clearly presented analysis is perceived as more useful in supporting learning.

Key finding 3: *Although no significant effect was found for the object description (O), it is believed that clarity is important from a practical perspective.*

O (the object description) did not have a significant effect on usefulness. While there were differences between descriptions that contained clear/ unclear object descriptions, these results were not statistically significant. This is a remarkable finding if we consider that O is an important component of real evaluation descriptions. Notably, it can be seen as the logical point of departure for understanding A and C.

5.3.1 Contribution

The findings presented here indicate that the way emergency exercise evaluations are documented is important, and directly affects their usefulness. Different components are more-or-less useful depending on the purpose of the report. Specifically, the usefulness of an evaluation for learning purposes is improved when its analysis and conclusions are clearer. This implies that evaluations should provide information about what happened, and how and why this was the case. Such information can be used to support the transfer from lessons identified to lessons learned, and the creation of a knowledge base. In contrast, evaluations used for accountability purposes are only improved by the clarity of the conclusion. Here, they should focus on providing a judgement, outcome or result with regard to the evaluated performance.

The evaluation description introduced in this study offers professionals a way to make evaluations more useful for their users. It not only provides an agreed framework that ensures that findings can be shared, but also offers scope for situational customisation. It can support the creation of a common foundation that can be used to form a knowledge base, and support meta-evaluation. Scientists can use this concept to determine the best way to conduct an evaluation, and to optimise the presentation of its outcomes. These findings also indicate the importance of documenting emergency exercise evaluations, and underline the need for clear guidelines. Guidelines and/ or frameworks could help professionals to structure their work, notably by indicating the criteria used to arrive at any conclusions, and supporting arguments.

5.4 **Paper IV**: What do practitioners expect from an evaluation report? A qualitative analysis of Dutch crisis management professionals' expectations

The fourth paper is closely linked to the third, as it also partially addresses RQ 3. It is partially explorative but again, also contains prescriptive aspects. In contrast to Paper III, this qualitative study addresses perceived usefulness by focusing on the expectations of crisis professionals who use exercise evaluations in their work. It assumes that the perceived usefulness of an evaluation depends on the target audience/ users; this implies that evaluations might be perceived differently by different individuals. Gaining a better understanding of what users think an evaluation report should contain, and what they deem important in order for evaluations to be perceived as useful could improve utility, usability and outcomes.

This study addresses sub-question RQ 3b: *What do crisis management professionals expect to find in a useful crisis management evaluation report?* Again it was run in the context of Dutch crisis management. A stakeholder information analysis (see section 4.3.5.) was used to survey the expectations of professionals holding key roles (see section 3.3.2) in the Dutch crisis management system. For practical reasons, and in order to reach a wider group, the quantitative survey was supplemented with additional open questions. The analysis (see section 4.3.5) led to the following, key findings:

Key finding 1: *Five main themes can be distinguished regarding what users expect to find in evaluation reports.*

The analysis identified five main themes: (I) information on why the evaluation is needed or what it is used for (purpose); (II) information about what, or who, is being evaluated (object); (III) information that is needed to reach conclusions, or details about what happened, how and why (analysis); (IV) details about the outcome, or how well the object of the evaluation performed (conclusions); and (V) detailed information about how the data should be presented (design of the evaluation report).

Key finding 2: *Evaluations should be designed around their expected purpose. Typically, this is learning/ development. However, if the purpose is accountability then blaming/ finger pointing and scapegoating should be avoided.*

The findings highlighted that the purpose of the evaluation is the starting point for many users, and is often referred to when discussing other aspects such as its object, analysis or conclusions. The data showed that the majority of respondents expect evaluations to contribute to future learning and improvement. They see them as an opportunity to share experience, with the aim of avoiding mistakes being repeated in future activities. Evaluations are expected to support them in being, or becoming, better-prepared for future disasters. Respondents also reflected on the evaluation product and its form. The analysis revealed that the latter should provide clear, concise statements about required actions, if any, that would improve future preparedness or responses. Some respondents expected a useful evaluation to contribute to holding individuals or organisations accountable. In this case, it should provide insights into performance, by comparing what was expected with what was actually delivered.

Although it was possible to identify these two distinct purposes, further analysis showed that it is difficult to separate them. Even if the purpose is clear, the information that is presented can be interpreted and used in two ways, one leading to learning and the other to accountability. In any case, respondents noted that evaluations should avoid apportioning blame, as this has a negative impact on both the evaluee and users.

Key finding 3: *Evaluations can focus on a variety of objects, and these objects can influence the chosen approach (e.g. qualitative or quantitative).*

Respondents highlighted the importance of the object of the evaluation. Many objects can be identified at various levels, i.e. the system, the organisation, the team or the individual. Each might require a different evaluation approach; for example, the evaluation of a new piece of equipment might require a more quantitative approach, while the evaluation of teams might be more qualitative and discussion-based.

Key finding 4: *The analysis should be rigorous, take into account the context, and go beyond the individual.*

Respondents noted that analyses/ the evaluation should be rigorous, go beyond the individual, and adopt a broad, system-wide perspective. They noted that it is important to take into account the context in which the object is evaluated in the analysis and evaluation, as this gives perspective to the decisions and actions that were made, and any dilemmas that arose. They underlined that the scenario should be as realistic as possible, given that retrospective (*ex-post*) information might differ from the actual information that was available at the time of the event. Finally, the analysis and any

meta-analysis should make it possible to distinguish between case-specific (one time) failures and ongoing systemic failures, and provide directions for change.

5.4.1 Contribution

The variety of views and expectations noted in this study show that it is difficult to create one form of evaluation report that meets all expectations. However, it should be possible to adjust and augment a generic product to meet different needs. The study underlines that the majority of respondents expected evaluation reports to contribute to learning and support improvement. They want be provided with actionable, evidence-based feedback regarding what could be done differently or better, and indicate how this can be achieved.

User expectations need to be clarified in the early phases of the design, as they have implications for data collection, analysis and presentation. Existing guidance does not encourage the active involvement of users, and this could be improved in the future, for example by making a needs analysis an integral part of the process. This study also showed that users are aware that evaluations can be subject to biases such as hindsight, time distortion or selection. It is important that these biases are identified and, if possible, mitigated, notably during the analysis phase, in order to improve the credibility and reliability of the product. The correct application of scientifically-proven, rigorous methods can be seen as a step in this direction. Finally, it was noted that evaluation reports should try to overcome the hurdle of simultaneously delivering practitioner relevance and scholarly rigour, also referred to as the rigour–relevance gap (Hodgkinson & Rousseau, 2009; Kieser & Leiner, 2009; Pettigrew, 2001).

6 Discussion

This chapter adopts a broader perspective. It starts by providing an overview of the current status of DRM exercise evaluation in practice and in theory. The conceptual basis for future research and practical applications is strengthened by introducing new models that highlight key aspects, based on the underlying research. It addresses RQ4, in order to anchor evaluation in the broader DRM context. It is also reflective; patterns in earlier research are examined in order to draw conclusions about the overall outcomes, and what these findings mean.

6.1 Developments in DRM exercise evaluation

This research enhances our understanding of using evaluation to support DRM. The starting point was to review the state-of-the-art by mapping the literature on disaster exercise evaluation (see section 5.1 and Paper I).

6.1.1 Past, present and future developments

History teaches us that formal evaluation was first introduced in the mid-1840s in the United States (Stufflebeam & Coryn, 2014) as a tool for assessing student learning. More than a century later, one of the earliest academic articles on evaluation examined the DRM context, and addresses the evaluation of a hospital disaster plan (Letourneau, 1962). Even today, this paper is a representative example of evaluation in a DRM context, and reflects two findings reported in Paper I:

- most of the literature on disaster exercise evaluation is in the subject area of medicine/ public health;
- in a DRM context, evaluation is often used in combination with plan testing. A plan or procedure forms the basis for the design of the exercise and its evaluation.

It could be argued that these examples show us that little has changed. Heath (1998) claims that the reason for this lack of progress is that effective crisis management principles and practices are still evolving, and the DRM community is still learning how to manage crisis situations in practice. In this context, processes such as evaluation take a back seat. However, this appears to be changing, as it becomes clearer that evaluations are a tool that can directly support DRM. For example, they provide insight into the effectiveness of current practice, can identify weakness, and propose improvements.

In the case of the Netherlands, response evaluations started to attract attention in the early 1990s. Initially, the focus was on incident evaluations/ analyses. One reason for this limited approach is the social impact and the media attention given to incidents such as the Bijlmer disaster in 1992[10]. In the 2000s, operational evaluations that focused on the emergency response organisation and management became more common. This marked a shift from a simple description of what had happened, to examining the effectiveness of the response system, and analysing questions such as how did it work and could it be improved in the future? This approach was taken in response to the fireworks disaster in 2000 and the New Year's Eve café fire in 2001, ultimately contributing to the introduction of the Safety Regions concept. The recent Covid-19 outbreak is likely to trigger a range of investigations and evaluations that may lead to new DRM insights and developments. These examples demonstrate, as noted in the introduction, that the complexity and evolution of the evaluation context should be seen as a driver for the further development of evaluation systems.

6.1.2 Formal vs informal evaluations

Section 2.3.1 illustrated that evaluations can be either informal or formal. This research focuses on the latter, and their products. However, during exercises and real responses, and in life in general, individuals or participants constantly make their own informal evaluations. This is often done subconsciously, for the person's own use, ending in a low level/ non-critical product. They can be a very useful tool, as without them it would probably not have been possible for humans to evolve. However, they are mostly invisible, and without direct questioning an observer would be unable to know how, or if, they were performed, and, if they were, what their effects or purposes were.

[10] On 4 October 1992 a Boeing 747 cargo aircraft owned by the Israeli airline El Al crashed into flats in the Bijlmer neighbourhood of Amsterdam. A total of 43 people died, 11 others were seriously injured and 15 had minor injuries (Wikipedia, 2021a).

It is difficult to predict situations in which individuals need to have an evaluative capacity, and it would be even more difficult to find a single rigorous and systematic method to formalise their evaluations. Nevertheless, there is a logic that each individual implicitly follows, and that supports him/ her in determining the worth, merit or value of something. Examples include previous experience, discussions or general conversations that are used to obtain data that support value judgements and, if necessary, behaviour adaptation. This process is often far from explicit, and the purpose of the individual's evaluation remains unstated. Therefore, there will always be questions regarding the rigour of any internalised/ informal evaluation system compared to explicit, rigorous formal evaluations.

Informal evaluations invariably only present an individual snapshot of reality. This personal view reflects a single perspective, rather than a structured view of the entire situation. They will almost certainly lack information about the context or the system that the individual is part of. It could be compared to taking a single photograph of an incident. Although the image contains a significant amount of information, without further photographs, taken from other locations, it is almost impossible to determine the full scale and complexity of the situation. Even if we were able to collect these informal evaluations, they would cover such a vast range of subjects that one system would not meet the needs of all. These challenges mean that informal evaluations are beyond the scope of this study.

Formal evaluations, on the other hand, support the rigorous and systematic collection of data. They are a way to collect individual pictures and tease out broader, connected elements of knowledge (justified beliefs) that were not visible in a single image. This formal process is illustrated in Figure 8.

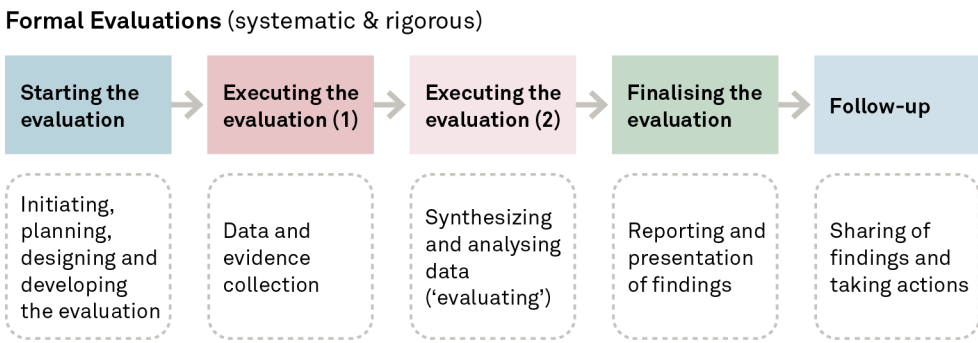


Figure 8: Formal disaster response (exercise) evaluation process
The figure illustrates the various steps in the evaluation process and provides examples of tasks or activities related to these steps.

Preparing for, and responding to disasters in a DRM context is not an individual effort. It is invariably a multi-agency collaboration, focused on the joint outcome of protecting collective values. It can be argued that, while an individual's understanding of a specific situation or context is only validated using evidence gathered at their specific viewpoint, if sufficient individuals have similar experiences, the organisation or system that they are part of could, eventually, align or organise itself to provide the most effective response. Even without a formal evaluation process, the shared experience of many individuals could improve effectiveness. An illustration comes from early settlements, at a time when there were no formal emergency response organisations. However, if individual views or informal evaluations are not systematically collected, particularly in multi-layered organisations, they are easily lost.

Formalising evaluations

Evaluation and, more specifically, formal evaluation, should be considered a science, as it is a structured, input/ outcome-based systematic approach. However, as illustrated in Papers I and II, poor or unstructured execution can make the process relatively unreliable.

In the DRM context, formal evaluations can provide us with reliable, evidence-based statements regarding questions such as what happened and what was the response? For this to happen, it is important that data collection methods are both suitable and support the formulation of such statements. Biases and any limitations must also be known. One example is observations, which are difficult to collect during real disasters. Data is often collected in post-incident interviews, leading to time distortion or confirmation bias. Data can also support statements such as why did something happen? On the other hand, normative questions such as what should have been done better are less straightforward to answer, as they rely on the establishment of a measure or benchmark to determine success.

Formal evaluations can be seen as the process of collecting all legitimate truth claims and beliefs. If there is a well-established knowledge base, it may be justifiable to draw conclusions as it provides evidence to support a decision. However, Hansson and Aven (2014) argue that the interpretation of the knowledge base is not always straightforward, and is complicated by the need to evaluate its contents in order to reach a summary judgement. The evaluation has to take values into account, address any uncertainty and is subject to the burden of proof. It could, thus, be seen as a combination of factual and value-based considerations. A systematic assessment should lead to the creation of reliable knowledge. In order to verify whether this is actually the

case, it should be looked at from an epistemological point of view that examine its methods, methodologies and claims (Hansson & Aven, 2014).

Paper II showed that there are a variety of evaluation designs in use on a regional and local level in the Netherlands, not to mention worldwide. Many provide no evidence regarding their impacts, and their methods can be questioned. To enhance credibility and improve reliability, simple controls can be implemented that would improve data validity, methodological reasoning, and any justification. One example is to validate individual statements by comparing and contrasting them with supporting statements or observations. Another is to implement multi-method approaches and use multiple data sources. A third is to run meta-studies that seek to identify trends and critically evaluate methods. Such meta-studies are currently very rare, as evaluations are mostly performed and presented as stand-alone case studies of a single exercise or incident.

Evaluations are created by various processes, each having a different point of departure. Sometimes the process is outlined in organisational policy or other documents. But in many cases, neither the process nor the method is systematically tested, investigated or reviewed, which makes it difficult to determine whether it is more or less likely to achieve its purpose. As highlighted in Paper II, this might have implications for the evaluation's quality and use. One simple measure of the success of an evaluation is the degree to which the end-user perceives it as useful and actionable.

6.1.3 Contributing to DRM

Evaluations contribute in three ways to both the evolution of DRM, and responses becoming more effective and efficient. Firstly, they can help to share findings, lessons learned, or lessons to be learned amongst a broader group. For example, in Australia there are large-scale wildfires that trigger a massive disaster response. The evaluation of the response might identify valuable lessons that are useful to the wildfire response community as a whole. Without a structured approach to information management, any record of a disaster will just be another story. Australian lessons learned could be reviewed by teams in, for example, Europe, where, even though there is a clearly structured response model, a comparison would be useful. A similar approach could be applied to the Covid-19 crisis in order to efficiently mitigate its effects. Transfer makes it possible to move to a higher, systems level of learning and formal evaluations are a way to share data, as illustrated in Figure 9.

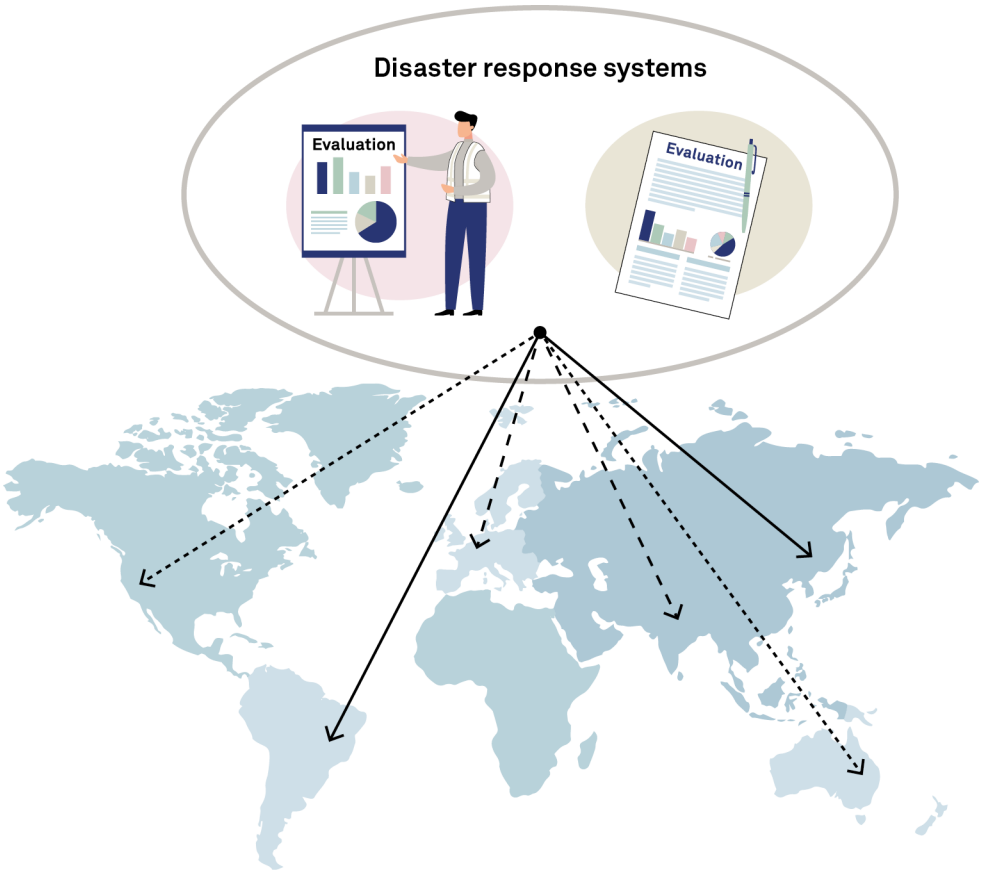


Figure 9: Evaluations as means to share experiences, lessons identified, and more.
The figure illustrates how evaluation outcomes can be shared (around the world) through its evaluation products.

Secondly, a collection of evaluations undertaken using a systematic, rigorous and comparable approach can form the basis for a meta-evaluation. This would help to manage risk at various levels and identify broader trends. Rasmussen (1997) notes that in a mature society, there is an inverse relationship between the accepted frequency and magnitude of disasters. Larger scale, but less frequent disasters are less accepted, and are likely to receive more attention than more frequent, smaller-scale events, even though collectively they might result in similar levels of damage. He introduced three risk management strategies: empirical, evolutionary and analytical that implicitly highlight the need for meta-evaluations. Infrequent, larger-scale disasters require an evolutionary and analytical strategy supported by a rigorous and systematic evidence-based approach. Formal evaluations can support meta or trend analyses as outcomes from smaller-scale responses are examined. In the Netherlands context, evaluations prepared by one Safety

Region could be used by another to improve their system. Similarly, evaluations from several Safety Regions could be collated and used on the national level, as suggested in Paper II. It is therefore of overriding importance that experiences and lessons to be learned are clearly documented, and that there is a structured explanation of how these conclusions were arrived at.

Thirdly, evaluations should be seen as an honest attempt to reflect on actions and processes during a real or simulated event, with the aim of increasing professionalism in the DRM community. However, they often fail to address the real problems revealed by the event or scenario in question (Birkland, 2009). While systematic data collection is a key element in a rigorous evaluation, it is critical that outputs are perceived as useful. We know that formal evaluations can support the direction of, and investment in learning and development, and provide insights into current practice (Abrahamsson et al., 2010; Alexander, 2015; Boin et al., 2017; Borell & Eriksson, 2008; Borodzicz & Van Haperen, 2002; Jongejan et al., 2011; Ritchie & MacDonald, 2010). However, a lack of rigour and transparency not only risks losing vital information, but also tacitly supports suboptimal outcomes, opening the way for political manipulation.

Theoretical contribution

Formal evaluations should be rigorous and systematic if they are to arrive at reliable conclusions. It is therefore important that the process itself, from a theoretical perspective, is improved. Any evaluation tool must be fit for purpose, and meet the needs of the end user. It should extract and record as much data as possible, in the best-possible way. The evaluation product needs to be designed and delivered in such a way that its outcomes are useful. As noted above, the scientific contribution to DRM exercise evaluation is limited. Papers I and II demonstrate that both theory and practice are evolving. While it can be difficult to determine whether an evaluation has had the desired impact, we must constantly ask, does the means achieve the intended end?

6.1.4 Evaluation quality

The quality of the evaluation influences its usefulness. Paper IV highlighted that users expect evaluations to be rigorous. But in practice (Paper II) it is difficult for users to determine their credibility, value, accuracy or usefulness. Many reports fail to provide sufficient detail on the process to enable the reader to assess its quality, notably regarding data collection and analysis (see Figure 8 – execution phase). Most current evaluations lack a solid theoretical foundation. This does not mean that they are useless, as they provide significant quantities of information, however, it is difficult to determine if this information is valid and reliable.

Data collection vs evaluation

An ongoing debate concerns justifying the applicability of selected evaluation methods and linking them to the overall purpose and specific context. In other words, which method works best for any particular case? Data collection quality assurance is achieved through the use of standard social science methods such as surveys and questionnaires, observations or interviews. There is little theoretical guidance regarding the choice of a particular method (Heath, 1998). The process of moving from data collection, through analysis, to reaching conclusions is key, and any weaknesses will devalue the outcomes.

In the Netherlands, evaluations mostly focus on providing a factual account of events. They describe how the object under evaluation performed, whether an evaluation framework was used, and often refer to legislation and related documents. At the same time, few provide details regarding the selected method or approach. Although it is often clear how data were collected, the evaluation process and any criteria are unclear. Therefore, it remains difficult to determine if the method worked well on that particular occasion, or whether other methods would have been more appropriate.

This research indicates that the weakest links in developing an evaluation are a lack of detail regarding justifications or reasoning, little rigorous testing of methods, and few hypotheses regarding the effects of different approaches. Consequently, professionals lack prescriptive, reliable and valid guidance regarding the design of evaluations. Greater clarity is needed in terms of how specific methods are linked to the overall purpose. Much progress remains to be made before research can deliver reliable, overarching, validated theories and methodologies that can guide best practice.

Evaluative reasoning

Hurteau et al. (2009) carried out a meta-analysis based on 130 evaluation reports produced for government departments in Canada. They concluded that 50% of these reports lacked credibility, as the results were not based on information relevant to the aim and, in 32% of cases, the supporting arguments were not sufficient to generate a judgement. Their analysis highlights three important aspects of evaluation: (1) its aim or purpose; (2) the analysis or reasoning process; and (3) the involvement of users and stakeholders. These observations can be linked to what Scriven (1980) called the 'logic of evaluation' or the 'evaluation double pyramid' (Hurteau et al., 2009). In Paper III, this was translated into an evaluation description.

The description should be seen as a way to conceptualise four key aspects of an evaluation: its Purpose, Object description, Analysis and Conclusions (P, O, A, C). Purpose can refer to either learning or accountability. This conceptualisation made it

possible to survey users' expectations and identify what, if anything, would enhance the usefulness of products (Papers III and IV). The concept is a generic description of the elements that users expect to find in an evaluation (see Figure 10). Notably, (I) information on why the evaluation is needed, or what it is used for (purpose), (II) information about what, or who, is being evaluated (object), (III) information that is needed to reach conclusions about what happened, how and why, (analysis), (IV) information that details the outcome, or how well the object of the evaluation performed (conclusions), and (V) details regarding how the data should be presented (design).

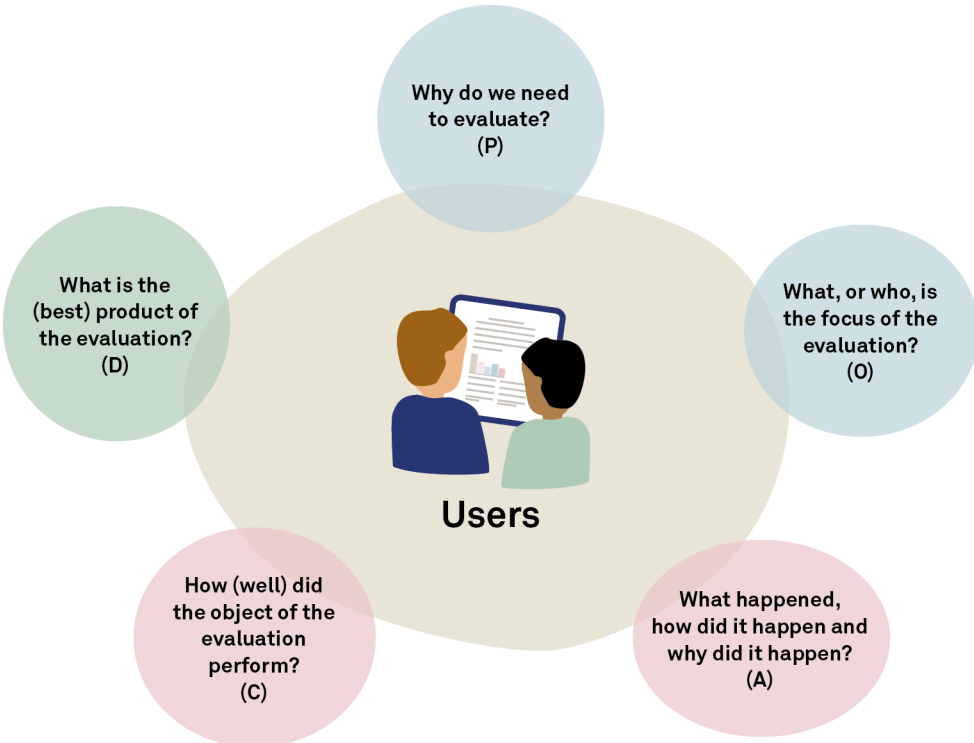


Figure 10: User-aspects of evaluations
This figure illustrates various user aspects, based on the analysis of user expectations explored in the context of RQ3b and discussed in Paper IV.

It is reasonable to assume that these aspects influence each other, and form a logical order. For example, the purpose of the evaluation is likely to affect the following steps. The results reported in Paper III showed that the clarity of the analysis and conclusion had a significant effect on how users perceive evaluations for learning purposes. However, only the conclusion affects accountability. Although important, the purpose

is only implicitly reflected in practice (Paper II), suggesting that although evaluations can have various purposes, they are not often explicitly mentioned in reports. In this case, if we consider that the purpose is the starting point, there is a risk of misuse—an evaluation designed for learning purposes could be used to hold people accountable.

Furthermore, the object under evaluation may be used for a purpose that was not initially considered. To take an extreme illustration: a truck can be used to transport goods, but it can also be used to knock down a door during a robbery. Thus, the purpose of an artefact is what we ascribe to it. Nevertheless, the purpose (or purposes) should be identified in order to establish a measurement baseline for the functionality of a system. It should be noted that we cannot force people to use the artefact (an evaluation, or any other object) in a certain way, but we can be clear with respect to how we have evaluated it and reached our conclusions. This implies that the full range of normal uses or expected outcomes must be clearly understood and agreed.

Finally, evaluations are subject to biases that need to be clearly identified and managed. Hindsight can play an important role when data is collected in real time, but the evaluation is performed *ex-post* when more information is available, potentially from other sources. It is important that any biases are taken into account, particularly in the analysis.

Making better use of evaluations

Each situation requires careful consideration, as there are a range of approaches and methods that can be used in various combinations. These combinations may improve the logic, and the subsequent usefulness of the evaluation product. If we see evaluation as a means that users use to achieve a purpose, it can be argued that usefulness is a function of the target audience. This implies that evaluations might be perceived differently by different individuals as a function of the intended purpose. Paper IV illustrates different user expectations and highlights that it is very hard to design a one-size-fits-all evaluation.

Paper II shows that Dutch professionals produce generic reports in a simulation setting. These reports do not clearly indicate a target audience, and do not make clear how the evaluation's design (why, what and how) makes it relevant and useful for this audience. We can ask whether they are written with a specific audience in mind, or if they are intended to be used by everyone or anyone? The latter case raises the question of what the purpose is? Are these documents really intended to contribute to learning and development, or are they just informative? Reports that do not guide users, and offer a one-size-fits-all approach are less likely to have an impact. Although they may be a good read, they are less likely to deliver useful, reliable and actionable information.

Users are averse to evaluations that seek to apportion blame, although references to specific individuals are sometimes unavoidable. Here, evaluators play an important role as they are responsible for designing and running an evaluation. Heath (1998) notes that evaluators need to seek hard information or objective facts, and move beyond convenient guesses or a search for a scapegoat. They need to be continually aware of the effects of biases, particularly hindsight and time distortion. In this context, the POAC format provides evaluators with guidance and support.

Evaluation is not an end, but a means to an end

Evaluations are not the sole contributor to change, learning or improvement. The evaluation product should be seen as a step on a much longer timeline, signalling the end of a period of reflection, or the beginning of controversy over what is claimed to have been learned (Birkland, 2009). In this research, evaluation is seen as a means, with a key focus on the product rather than the process. It is, however, acknowledged that the benefits of an evaluation extend much further. As stated by Stufflebeam (2003), the most important purpose of evaluation is not to prove, but to improve. From this perspective, it can be difficult to observe and, moreover, measure its direct effects.

The design approach (see section 4.1, Paper IV) might seem to suggest that evaluation always needs to have a defined purpose, such as better disaster preparedness. However, it should be emphasised that evaluations and their products alone might not lead to the desired outcome. Other circumstantial factors such as willingness, financial capacity, knowledge and understanding can be seen as prerequisites. In addition, evaluations are not always tangible.

This research identified that there should be a continuous cycle of evaluations that build upon each other. Papers I and II indicated that they are often performed as stand-alone cases, and are seen as a final step. Connecting them would enhance the value of each set of individual outcomes, and support broader learning by identifying themes. Learning from a disaster should not be seen as an outcome or goal of the process, but as an ongoing activity (Birkland, 2009). Lessons to be learned can be used to evaluate the implementation of recommended measures.

Paper II demonstrated that the Netherlands system combines formative (exercise) and summative (systemic test) evaluations. However, it lacks common ground that would encourage the two types to be linked. Evaluation descriptions can contribute to the development of this common ground. In addition, the introduction of meta-evaluations would not only improve the ability to build on previous outcomes and identify system errors, but they would also motivate the creation and use of frameworks that could ensure comparability. While this research provides some initial insights, it

remains unclear how effective current evaluations are, and how they contribute to the development of disaster preparedness. In this context, it is important that evaluation products are not seen as an end in themselves but as a means to achieve a purpose, or clarify if a purpose has been achieved.

This section underlines the notion that it is not the evaluation itself that leads to improvement, it is the *use* of evaluations *that can lead to* improvement. Evaluation should be seen as a means to an end. At the same time, it is not the holy grail. Many other factors influence our behaviour and, in turn, our response to disasters. Although evaluations are mainly used for learning/development or accountability purposes, there are others. In all cases, it is important to ensure that the evaluation is fit-for-purpose and addresses the user's needs. This research identifies various issues and provides several suggestions on how to improve the usefulness of evaluation products. While it is difficult to establish a causal relationship, it is reasonable to assume that a more useful product will have greater impact, and make a bigger contribution to preparedness and response.

6.2 Building a stronger conceptual basis for future research and the practical application of evaluation in DRM

This research seeks to contribute to improved preparedness. Its insights, conclusions and recommendations are expected to have benefits at individual, team, organisation and system levels. Although more work is needed, the contributions presented here serve as a point of departure. More precisely, the conceptual and empirical insights can be used to support a rigorous, systematic approach to evaluating disaster response exercises. While the present study is focused on exercises, it is reasonable to assume that the outcomes also apply to real responses.

This section demonstrates how the developed models can contribute to preparedness and response. It consolidates the key concepts that were introduced in Chapter 2, and reflects aspects of the studies highlighted in Chapter 5 and the appended papers. Although a starting point for further discussion, it should be noted that the models presented here are a conceptual simplification of real-life phenomena, and are only an initial contribution. It is highly likely that they will need to be revised as more knowledge is gained.

6.2.1 Evaluation in DRM

The first model relates to the broader context of this research—DRM (see sections 2.1 and 2.2) and the role of evaluations (see section 2.4). The DRM cycle was introduced as a way to structure activities or functions. In this approach, subsequent functions should, ideally, relate to each other. This reflects actual practice, where preparedness is built upon insights gained in the prevention phase (the red cycle in Figure 11), and the response uses insights gained in the preparedness phase (the green cycle in Figure 11). Techniques such as risk analysis, and exercise or response evaluations can be used to collect and share these insights and deliver effective DRM. In general, the aim is either to: (I) reduce the likelihood that adverse events lead to unwanted consequences; or (II) reduce the impact or severity of these consequences.

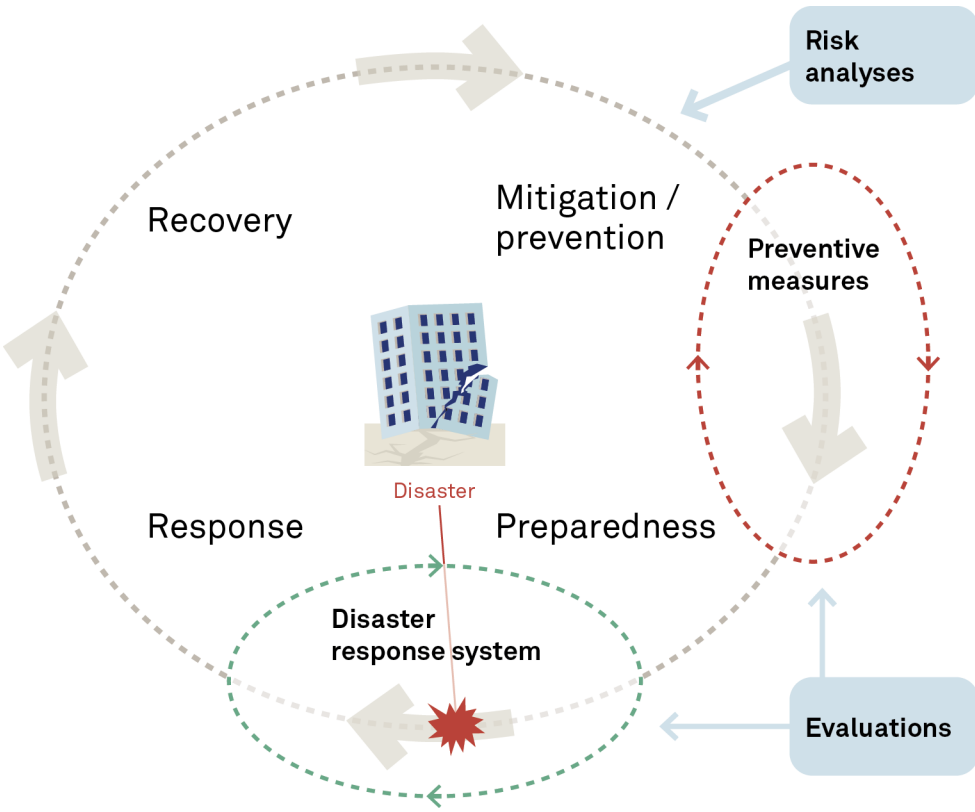


Figure 11: The role of evaluations in the DRM-cycle
This figure illustrates the DRM cycle and highlights two examples of how evaluation can contribute to its functioning.

(I) is most often found in prevention. Risk analyses seek to gain a better insight into aspects such as adverse events, consequences and uncertainty. The output of these

analyses can be used to take preventive action. However, the identified risk can remain. If this is the case, society needs to take action to prepare itself to deal with the impact of an adverse event. Here, the focus shifts from preventing disasters, to preparing for them in order to mitigate their consequences.

Events, uncertainty and the severity of consequences also play a central role in the analysis of capability. If the severity of consequences is described using the same global scale (e.g. lost lives, economic damage) a capability assessment can be an integral part of a risk assessment (Lindbom, 2020). However, it is often the case that uncertainty regarding how well the various actors can manage a disaster is much greater than uncertainty regarding how often a potentially disastrous event will occur. Evaluations of exercises and, if available, previous disasters, can help to reduce uncertainty. They can, for example, provide insights into how a response actor performs, and to what extent responders are capable of dealing with the disaster and its consequences.

Another factor to take into account is the 'spinning speed' of the various functions in the cycle. Routine events (incidents/ accidents) and rare events (disasters/ crises) have different characteristics. Routine events such as car accidents are more regular and many of the threats are known. Feedback from, and the evaluation of management activities can support the development of a knowledge base, which can be used in preparedness exercises. In practice, the response feedback loop spins much more quickly for routine events than disasters. As feedback from rare events is limited, they are necessarily associated with greater uncertainty.

Evaluations can play a crucial role in DRM by reducing uncertainty and providing a more solid foundation for decision-making. They are likely to have a greater role to play in development or learning when events are rare as it is likely that, in a routine situation, people and organisations can learn without a formal evaluation. This may either be because the cause/ consequences relationship is simpler, or because informal evaluations provide feedback.

An example: flood risk

Flood risk can be used to illustrate the relationships shown in Figure 11. If, in the prevention phase, a risk analysis identifies flood as a risk, preventive measures such as levees can be proposed, and eventually implemented. However, uncertainty remains. There is a residual risk that the levees will be insufficient, and the population will need to make its own preparations. This reflects a shift from influencing the likelihood (prevention) of a flood, to planning mitigation efforts. An incomplete analysis of the response capability creates greater uncertainty with respect to possible consequences should a flood occur.

An exercise, or a well-constructed and delivered simulation may reduce this uncertainty, as it provides more information regarding what can be expected of the response system. For example, the outcome of the risk analysis can be used to establish minimum response or readiness requirements. In the preparedness phase, this information can be used in planning, or to support the design of educational and training activities. Criteria can be tested to establish whether they are realistic, and whether response capacities are adequate. The data that are collected can provide insights that are fed back into the risk analysis. They can also be used in the response phase to support decision-making.

These insights can predict a realistic response, which can be used by strategic teams to develop a strategy and assemble resources. Requirements will be very different for a small local flood compared to a wide-scale inundation. This example illustrates that evaluation, as a means, not only supports connections between the various DRM functions, and the exchange of information (see Figure 11), but also the reduction of uncertainty.

6.2.2 The contribution of evaluations to preparedness and response

This section focuses on the relationship between real and simulated events (the green loop in Figure 11). Simulated events are seen as an opportunity to gain more knowledge about the capability of the DRM system. Figure 12 presents the green loop in more detail, and shows that the disaster response system plays a central role. The overall purpose is to protect what society values from the impacts of disastrous events. As highlighted in sections 2.2.3 and 2.2.4, it needs to be ready to respond effectively, with appropriate skills and resources, when disaster strikes. This is achieved through simulations that focus on all, or specific elements of the response system. In practice, teams or units are selected, and a scenario is constructed to test their capacity. The scenario sets out how, and when a team, or elements of it, are triggered in order to see how they respond. Ideally, the situation is designed so that there is still some uncertainty regarding the outcome. The evaluation establishes performance: what went well and what can be improved. This process is illustrated in the lower part of the left loop in Figure 12.

A similar process can be identified when the disaster response system needs to respond to a real event. However, in most cases, the event is unplanned and the response capability is unknown, as it depends on the severity or consequences of the event, and the values that need to be protected. The evaluation of the performance takes into account what happened, and the response. This process is illustrated in the lower part of the right loop in Figure 12.

Although Figure 12 places the disaster response system at the centre of the process, it should be noted that disaster response is part of the broader DRM system. It does not operate in isolation, as it must address not only uncertainty, but also a range of positive and negative influences such as technological advances, scientific insights and political shifts. It must be able to adapt itself to the evolving context. This is sometimes referred to as a control problem: response capabilities must match the risks in a changing landscape (Brehmer, 1992).

Section 2.2.2 highlights that control is maintained through a non-linear, cyclic feedback process. The latter ensures that the disaster response system is up-to-date, and fit to operate in the current iteration of a constantly-changing risk landscape. Evaluation provides feedback that the system can use to take action. It can be derived from simulated events that support learning and development (left loop, Figure 12) in order to become better-prepared in the response phase (right loop, Figure 12). Moreover, as a disaster response system inevitably operates in a political context, evaluations can also be used to hold components, or actors, accountable.

Papers III and IV underline that these elements structure the design of a useful and actionable evaluation product. This product, together with lessons identified and lessons learned, can not only be used by the disaster response system under evaluation, but also be shared with other systems. It is not a loop as such, but rather a connection between systems.

Finally, the model shows that there are three sources of material for evaluating operational disaster responses: (1) real, disastrous events in the current system; (2) simulated disastrous events (or exercises) in the current system; and (3) experiences, both real and simulated, in other disaster response systems. The successful transfer of knowledge relies upon the compatibility of the products produced by these various sources, as this is the primary way to support cross-system learning.

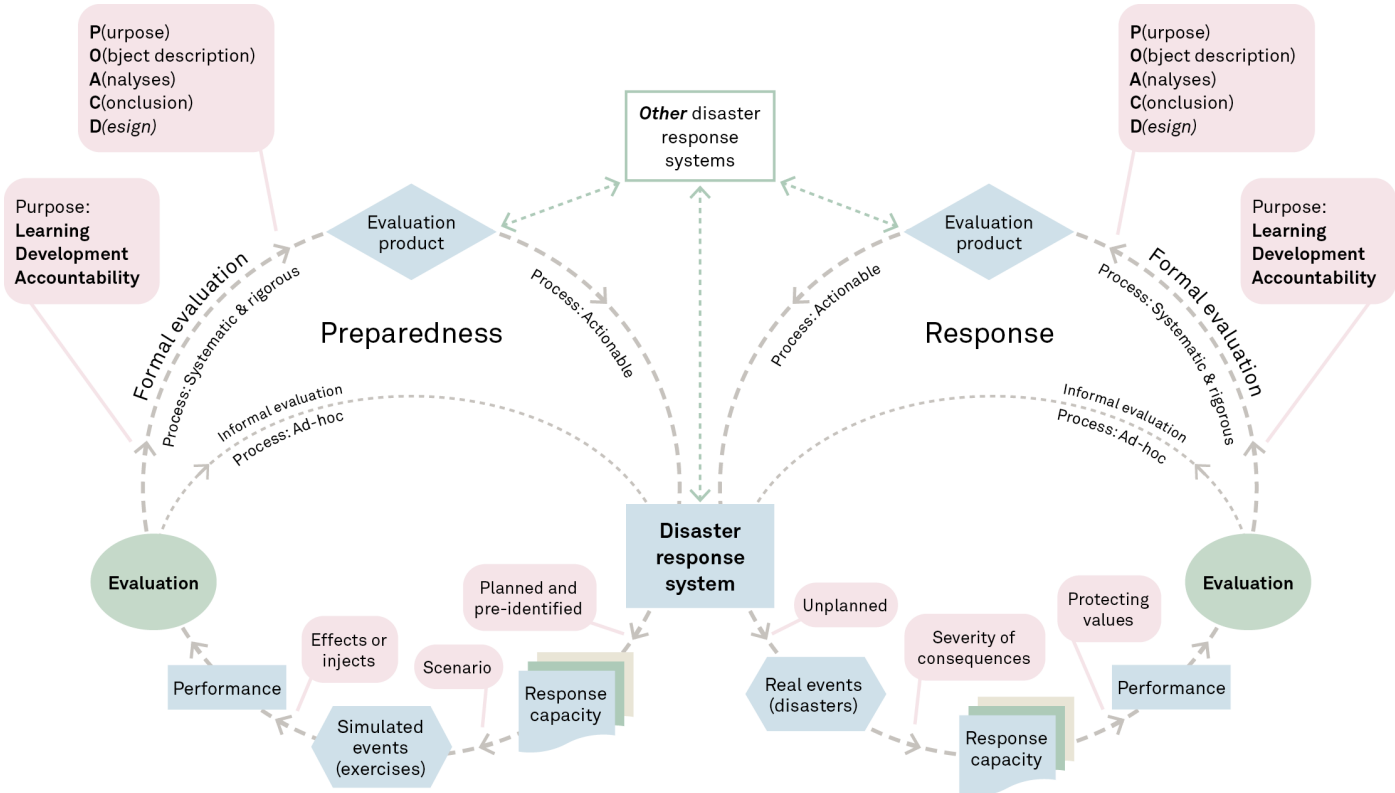


Figure 12: Evaluation contributing to improved preparedness and response

The figure shows a model that is based on two loops that are bound together by the disaster response system. The left loop provides details about the role of evaluation with regard to preparedness activities such as exercises or simulated events. The right loop illustrates the role of evaluation during real responses. Both loops also include a smaller loop that illustrates informal evaluations that also take place during these events. The bigger (formal evaluation) loop uses the outcomes of this research as a basis for demonstrating where improvements can be made. Finally, the model also shows that evaluations and its outcomes and products can also be shared amongst various (other) disaster response systems in order to stimulate wider development and learning.

The model shown in Figure 12 is not intended to provide a complete description of all of the factors that influence preparedness and response. Instead, it should be seen as a representation of the various components presented in this thesis, and an illustration of how they interact. The two loops show the role of evaluation in the preparedness phase (simulated events) and the response phase (real-life events). The two loops are bound together by the disaster response system box, which illustrates the state of the DRM system. This state records the current position. It is influenced by the two loops that pass through it and, in turn, influences them. This information changes the state of the DRM system, hopefully resulting in improvements that make it better-prepared for future disasters.

The dual loop model illustrates the similarities between the two loops. However, at the same time, it should be acknowledged that learning and development can be instigated by information from other, external disaster response systems. Therefore, in practice, the influence of the loops is likely to differ. In some contexts the influence of other disaster response systems might be very strong, while in others the loops dominate. The disaster response system therefore goes beyond a closed system that only relies on feedback from these two loops.

6.2.3 Designing exercises for operational response evaluation

The two-loop model illustrated in Figure 12 shows that the two phases, preparedness and response, each have their own, distinctive processes. However, despite their differences, three key aspects related to DRM play a crucial role: (1) the baseline scenario or trigger event; (2) the response capabilities available to deal with the consequences; and (3) the severity of the consequences or the final outcome. If we look more closely at the left loop (preparedness), it becomes apparent that the design of an exercise, including its scenario, is a key factor in its ability to contribute to reducing uncertainty. A significant difference between a simulated event and a real-life event is the level of control. In simulated events the nature of the event and the response system can be influenced and, to a certain extent, the design can be seen as a selection process. Figure 13 illustrates this process for a response exercise or simulation, and introduces the above-mentioned three variables.

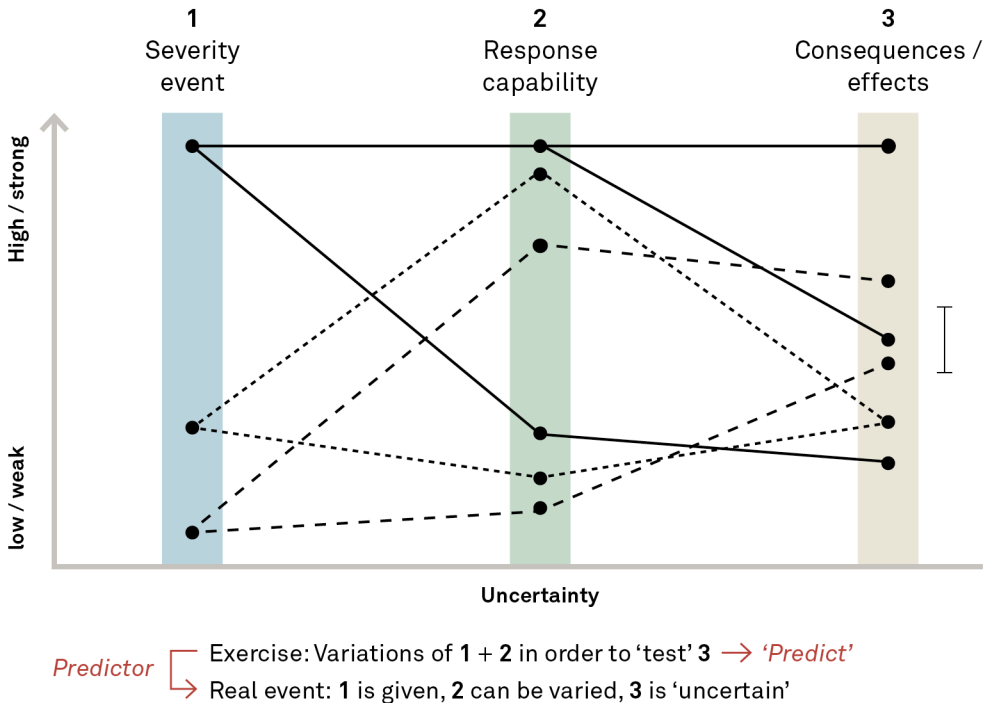


Figure 13: Response variables
This figure illustrates three variables that can be identified when designing and evaluating the response to (simulated) events. They are all interdependent. For example the severity of the event (variable 1) and the response cabapility (variable 2) can have an impact on the consequences or effects (variable 3). More precisely, if the severity of the event is high and the response capacity is weak, the effects can be severe, and this result can be tested in an exercise. But other variatons are possible on the scale from strong to weak. They can all be pre-identified during exercises to get an impression of their relationships. It should be noted that this model does not only apply to exercises, but can also be used for real events. In real events the severity of the event is given. The response capacity can be varied by scaling-up or down, influencing consequences or effects. However, during an event this is uncertain and might be based on trial-and-error. Here, simulated events can help as their evaluation can provide input that reduces uncertainty.

The first independent variable is the scenario or event. In an exercise, a range of choices can be made, such as the scale or magnitude of the event. Is it local, regional, national or international? Is it industrial or natural? These choices have an impact on the severity of the consequences, and determine how challenging the situation is. The response capability is the second independent variable. This, or an element of it, is often the focus of both the exercise and the evaluation, as it is the key artefact or object in managing or mitigating the effects of the event. Again, a variety of choices need to be made. For example, capacity can vary from small to large (i.e. from teams and first-responder groups, to specialist, multi-organisational units). In general, the relationship between the severity of the event and the capabilities that are included determine how difficult the exercise will be. An exercise that tests an event such as a single traffic accident involving two cars, and a system with an abundance of resources is likely to be

perceived as easy. On the other hand, a challenging scenario could include multiple traffic accidents, a terror attack, and a fire, and could test a very limited part of the total response system. Under these conditions, the exercise will most likely be perceived as difficult.

The third variable relates to the consequences or effects of the event/ scenario and the response. It can be seen as dependent upon, or influenced by choices related to the event and capabilities. Maximum value is gained from the exercise when the content and complexity of the scenario matches the capability and size of any proposed response, as this ensures that the elements under evaluation are sufficiently challenged and have a range of response options. Where either the scenario is too complex with respect to the resources available, or the resources available have far greater capability than the scenario requires, the subsequent evaluation will yield little information of value as it does not mirror real-world expectations and actions.

It is often said that scenarios are developed with the last real event in mind (see Paper II). Thus, if it is possible to predict the outcome of an exercise in advance, running it will do little to reduce uncertainty. It can, however, serve as a confirmation of capability tool. In this case, it is best-used during the final stages of a developmental training and education cycle. When a more unpredictable scenario is used, the exercise can test what can be done in a certain situation, and what these actions are likely to achieve. Although this approach has the disadvantage that it is difficult to estimate beforehand how well an event will be managed, the potential for reducing uncertainty about response capabilities in the case of a real incident is greater. It follows from this that if the goal is to reduce uncertainty, exercise design should avoid a significant imbalance between the challenges of the scenario and the skillsets and resources of responders. This applies to exercises at any level within the response organisation in question. An example of how this can be reduced is by getting to know the capability of the response system more accurately.

The ideal scenario acknowledges that if recognised procedures are used in a standard way, there is a general likelihood of success. Moreover, even if procedures are not applied optimally, there is still a likelihood of success, with clear lessons to be learned. This design is ideal for a confirmation of learning-type exercise. A second scenario might introduce unknowns, such as new procedures or emerging risks. In this case, success is less likely, but it is more likely that skills or resource gaps will be identified. An exercise design that gives little thought to the balance will contribute very little, and may even have a negative impact on participants.

Once again, the model presented in Figure 13 is not intended to provide a complete description of all of the factors that influence exercise design and evaluation. It is used

here to illustrate the relationship between the range of variables in the design and evaluation of any exercise or response.

Response evaluation, like risk analysis, takes place in an environment with some degree of uncertainty. The focus is rare events, where true values cannot be determined with certainty. Nevertheless, both exercises and their evaluation can contribute to reducing uncertainty by exploring the range of response options and measuring their effectiveness. It is important that the purpose, function and form of the evaluation corresponds to the purpose and function of the exercise. In the first example given above, expected outcomes are relatively simple to determine when participants have a skillset that is more-or-less adequate for the task. The second example is more a case of organisational learning, as known skills are pitted against novel tasks. In the latter case we are not necessarily looking at the successes, but more at required skills, capability ranges or procedural gaps.

In contrast to exercises, the only variable in a real response is the number and capabilities of responders; the severity of the event and its consequences cannot be controlled and are, at least initially, uncertain. In this case, capabilities can influence the severity of any consequences, and it is possible to evaluate their performance with regard to their actual versus expected influence. In both circumstances, whether an exercise or an emergency, a carefully-considered, targeted formal evaluation will yield valuable data that can support the ongoing development of emergency response capability.

This research makes several contributions to a broader understanding of the role of evaluation. A scoping study established the current level of knowledge regarding this area of evaluation within the scientific community. It provided insight into perceptions of the usefulness of evaluations, based on the novel concept of the evaluation description, and investigated users' expectations. The output is not intended to be used as a set of instructions or protocols, but as an input to the creative and innovative activity of designing structures, processes or interventions. The recognition, on a meta level, that all disasters have broadly similar preparedness, response and recovery requirements, and that they all involve development and validation activities, albeit at differing levels, is a key driver for future research. This observation can be used as a blueprint to shape the design of future models and activities, both in preparedness and response phases.

It is clear that the topic of disaster response evaluation is evolving, with contributions from many perspectives. The present chapter presents a synthesis of several ideas, both from the point of view of the research reported in this thesis and the work of others,

and it may be used as a point of departure for future work. Some of the more promising avenues are presented in the next chapter.

7 Future work

Chapter 4 highlights that research, in general, contributes to greater understanding and the development of knowledge of all aspects of a phenomenon; here, it concerns evaluating responses in a DRM context. This chapter identifies the potential implications of the outcomes of this research for both theory and practice. In doing this it also (partially) addresses the overarching RQ by providing suggestions on how (the usefulness of) disaster response evaluations can be improved in the future.

7.1 Future research

The outcomes presented here are not intended to be used as a set of instructions or fixed protocols, but as input to the creative and innovative activity of designing structures, processes or interventions to evaluate DRM responses during exercises and real events. The main RQ identified specific areas that require in-depth investigation in order to improve the usefulness and, moreover, the impact of disaster response evaluations. Paper I and Paper II showed that the topic of evaluation, within a DRM context, deserves more scientific attention. Both theory and practice would benefit from more research. The sections below set out five proposals.

7.1.1 Cross-discipline transfer

Evaluation is not unique to DRM; it is used as a process tool across a wide range of disciplines. This means that DRM can benefit from approaches and methods that have already been developed and tested in other disciplines. These methods can be examined to see if they meet the needs of the crisis management context and, if so, can be added to the options in disaster response evaluation design. Future research would contribute to this transfer. One example would be to investigate to what extent approaches and methods that are used in areas such as product development and design can be adapted to the crisis management context. Evaluation design could benefit from using a mix of qualitative and quantitative methods, as was the case in this research (Chapter 4). Further research could identify combinations that complement each other.

7.1.2 Methods and processes

Future research could look at methods and processes used in the emergency response community. For example, Paper II showed that in the Netherlands alone, a variety of methods are being used, which are mainly judged on face value. Investigating the strengths and weaknesses of these methods would help professionals to select the most appropriate or reliable option in the design phase. Other innovative techniques, such as technological data collection, could benefit future designs, especially if such tools are implemented rigorously. Experimental research is another promising avenue. For example, it became apparent during this project that actively involving users made a positive contribution to designing products that were seen as useful. The field of DRM offers ample opportunity to exercise and test novel methods, opening up a wealth of possibilities for future research.

7.1.3 A logical approach

A focus on the POAC(D) concept in formal evaluations would be beneficial. POAC(D), as introduced in this thesis, can be seen as a broad and abstract notion. Future research could examine individual components in detail, identify other forms of manipulation, and how they influence the usefulness of evaluations. Studies should put more emphasis on how the collected data is analysed and weighted in order to reach conclusions. It would be especially beneficial if research ended in well-founded guidance for professionals regarding what data should be collected, and how, in order to perform a rigorous and relevant analysis. As an illustration, meta-analyses could compare various evaluation approaches. The focus should go beyond individual cases, and look at the meta-level. This would support a search for a scientifically-proven disaster response evaluation logic or, even better, a generic evaluation theory. Papers III and IV can be seen as a first step towards this, and can be used by researchers as a basis for further, more detailed investigations. Such efforts would improve the quality of evaluations in general, and may help to identify an optimal approach.

7.1.4 The impact of evaluations

Further research could look at enhancing the impact of evaluations. The benefit or impact of the evaluation product, the tangible outcome of an evaluation, could be investigated. Questions include: To what extent are evaluations followed-up? To what extent did it actually achieve its purpose? What did we actually learn from this evaluation? Studies could examine the so-called actionability of evaluations, as highlighted in Paper IV. This would go beyond usefulness, which was the focus of the

present research, and focus on actionability (i.e. whether evaluations motivate users to take positive action and change their behaviour, plans, procedures, etc.).

In this context, studies could investigate how evaluees and stakeholders are contacted and informed after an event, before the final evaluation product is delivered, and how these messages are framed. This would help to identify other interventions that might increase actionability and follow-up. Such issues are difficult to study on a single-case basis, and require a more sustained, multi-case approach. Research could focus on whether lessons identified become lessons learned, and how this transfer can be initiated or improved. An experimental setting could be useful in this respect, as external influences can be controlled and outcomes can be compared. Finally, longer-term challenges, such as the current Covid-19 crisis, might offer interesting research opportunities.

7.1.5 Informal evaluations

While this research focussed on formal evaluations, future investigations of the impact of informal evaluations may have merit. For example, it would be interesting to identify how individuals determine whether performance was good or bad. These informal, individual judgements might support formal evaluation and create a shared evaluation framework. It is also possible to use evaluations as a means to investigate if, for example, exercises are really a useful way to prepare for real events. Past exercises have tested a pandemic scenario. An evaluation could compare these exercises to the real response to the Covid-19 pandemic.

7.2 Practical developments

This research also identified possible avenues for improving the evaluation of (simulated) responses in practice. Paper II and Paper IV provided some practical insights into existing opportunities. This section outlines five suggestions that could be explored and incorporated into (Dutch) evaluation practice.

7.2.1 Evaluation design – guidelines, frameworks and/ or standardisation

Professionals could work towards the standardisation of the evaluation frameworks used within an organisation, or across a range of linked exercises. An organisational or national standard would create a common language to facilitate communication and collaboration. In addition, it would ensure that intended purposes are achieved, quality is guaranteed, and findings can be shared.

A draft standard could provide a starting point for discussions between professionals and researchers. In addition, it would support and guide evaluators towards appropriate solutions and processes. It should not be seen as descriptive guidance, but as a prescriptive norm that stipulates minimum requirements. This would help to develop best practice. Standards provide both evaluators and users with a firm foundation for crafting and defending their evaluation design and its outcomes. They also make the design process more efficient, especially as the added-value of evaluation is sometimes questioned. A standard should provide guidance on the exercise and design processes, and answer questions such as why, who, when, what and how? This would help to ensure that evaluators and their clients communicate effectively, and reach a mutual understanding concerning evaluation criteria.

7.2.2 How good is good enough? Is the gold standard always achievable or necessary?

Professionals should ensure that the evaluations that they perform are not seen as independent but related, and provide performance data that can both inform, and be built upon, by others. More attention should be given to meta-analysis to improve the quality of, and to support the identification of design propositions and generalisations that can be related to the needs of specific user groups. A meta-evaluation can also help professionals to determine how good is good enough? Answers to this question might be *ad-hoc*, with each case requiring an element of consideration or adjudication to determine an acceptable and achievable performance description.

Another route, worthy of consideration, would be to apply an evidence-based approach. In the medical context, Kitson, et al. (1998) define evidence as the combination of research, clinical expertise and patient choice. Applied to crisis management evaluation, it could be defined as a combination of research evidence, crisis management professional consensus and the context in which the response takes place. According to Kitson et al. (1998), research evidence may be unsystematic, anecdotal and descriptive (low value), or a rigorous, systematic (quantitative or qualitative) evaluation (high value). Similarly, professional consensus may be lacking (low value) or very cohesive (high value), and the response context may range from completely overlooked (low value) to a process of systematic feedback and input into decision-making (high value).

Clearly, higher-value evidence is more likely to contribute to successful research, or an evaluation that improves the disaster response system. However, an intervention that is found to be highly effective from a research perspective may be rejected by crisis management professionals and responders. The same applies to evaluation outcomes. Thus, when assessing the nature and strength of evaluation outcomes and their

potential for implementation, a combination of three dimensions—research, crisis management experience and the response context— need to be considered. Together, they can answer the question of 'how good is good enough?' and support the development and implementation of achievable, effective and acceptable improvements. The latter, in turn, will enhance abilities, systems or processes across responder groups at all levels.

This research showed that evaluation professionals should pay attention to describing or justifying data collection and/ or evaluation methods. In other words, why a specific method was chosen, and demonstrate how this method relates to the purpose of the exercise, the object, etc. Professionals should be able to draw upon guidance regarding what is considered 'good' as this would support a rigorous analysis and robust conclusions.

7.2.3 Evaluation products

Usefulness and actionability are key elements of successful evaluations. Currently, most evaluation products are not considered actionable, and recent technological developments might have changed how information is consumed by its users. Alternative media, such as factsheets, videos or infographics could be used to meet different purposes and learning styles of users. Knowledge must be presented in an accessible, concise and user-friendly format. Professionals should, where appropriate, consider these alternative evaluation products. This would maximise the likelihood of outcomes being understood, considered useful and acted upon.

7.2.4 Cyclic development

Evaluations can identify lessons to be learned from a particular (simulated) event, such as an exercise. If this event is one of a series, outcomes can be used to verify performance and, where performance does not meet the specified level, can provide insights into areas that require further development. This process can be followed with further learning or development, and re-evaluated in future exercises to see if the lesson has, in fact, been learned. For this to happen, there needs to be a continuous cycle of evaluations that follow-up and build upon each other. Professionals should, therefore, be encouraged to take this into consideration when setting up exercise programs. For example, in the Netherlands, an initial improvement would be to better-link the various types of evaluation, exercises, systemic tests and real responses. The same argument can be applied to emergencies that happen more frequently.

Exercises often look to the past, and scenarios are based on historical events. Designers should be encouraged to look to the future. This point is emphasised in Figure 12, which illustrates the relationship between emergency evaluation and the evaluation of exercises. It is important that there is a connection between exercise cycles and evaluations of emergencies as this would help to create a (local) knowledge base.

7.2.5 A clear purpose and an ethical approach

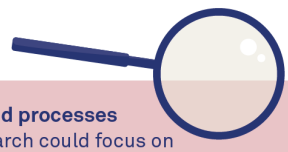
The two main points of departure in an evaluation are its purpose and its users or stakeholders. As highlighted in Paper I and Paper II, the real purpose of an evaluation can be ambiguous. Often learning is given as the primary purpose, while, in practice, accountability is also considered. These purposes are distinct and, as noted by the respondents in Paper IV, holding people accountable or being considered responsible for poor performance can be detrimental. Clearly stating the purpose not only has an impact on the design and selection of methods, it can also influence the impact of the outcome. Therefore, it is important that users and stakeholders are scrupulously transparent about this point and act accordingly.

Finally, it should be noted that all of the proposals put forward in this section and illustrated in Figure 14 consider the active involvement of users in both practice and research. In particular, evaluators need to actively engage with users to ensure that potential outcomes are meaningful, the design is appropriate, measurements are accurate and relevant, and the outcome has practical utility. At the same time, credibility is important, and both researchers and evaluators need to ensure that evaluations are accurate, honest and rigorous. Finally, Kitson et al. (1998) argue that the successful transfer of research into practice is a function of the interplay of: i) the level and nature of the evidence; ii) the context or environment into which the research is to be placed; and iii) the method or manner in which the process is facilitated. These three keys require further consideration to support the successful implementation of the suggestions made here. Evaluators can play an important role in this respect.

As highlighted throughout this thesis, the evaluation should be considered as a tool that helps users achieve their purpose, namely improving preparedness for future disasters. Without their active involvement in quantifying and qualifying perceptions of preparedness, it will remain difficult to ensure that evaluations have an impact and are a useful way to improve DRM. Any improvement will only come about with the investment of time and money. The effective use of well-designed evaluations will ensure that these resources, which are often in limited supply, focus on areas that actually require development, rather than those that are perceived as requiring it, whether in good faith or based on political expediency.

Future Research

Cross-discipline transfer
Future research could focus on examining approaches and methods that have already been developed and tested in other disciplines. They can be examined to see if they, for example, can be combined and meet the needs of evaluation in a DRM context.



Methods and processes
Future research could focus on investigating the strengths and weaknesses of methods (currently in use), which would help professionals in the design phase. The field of DRM offers ample opportunity to exercise, experiment with and test such (novel) methods.

A logical approach
Future research should put more emphasis on POAC(D) components, for example, on how the collected data is analysed and weighted in order to reach conclusions and how this influences the usefulness of evaluations. It would be especially beneficial if such research ended in well-founded guidance for professionals to create useful evaluations. In addition, the focus should go beyond individual cases, and look at the meta-level.

The impact of evaluations
Future research, in particular experimental studies, could focus on investigating the benefits or impacts of the evaluation product, and the tangible outcome of an evaluation. This goes beyond usefulness, and would help to identify other interventions that might increase actionability and follow-up.

Informal evaluations
Future research could focus on how informal, individual judgements might support formal evaluation and create a shared evaluation framework. In addition, it could also focus on investigating if exercises are really a useful way to prepare for real events.

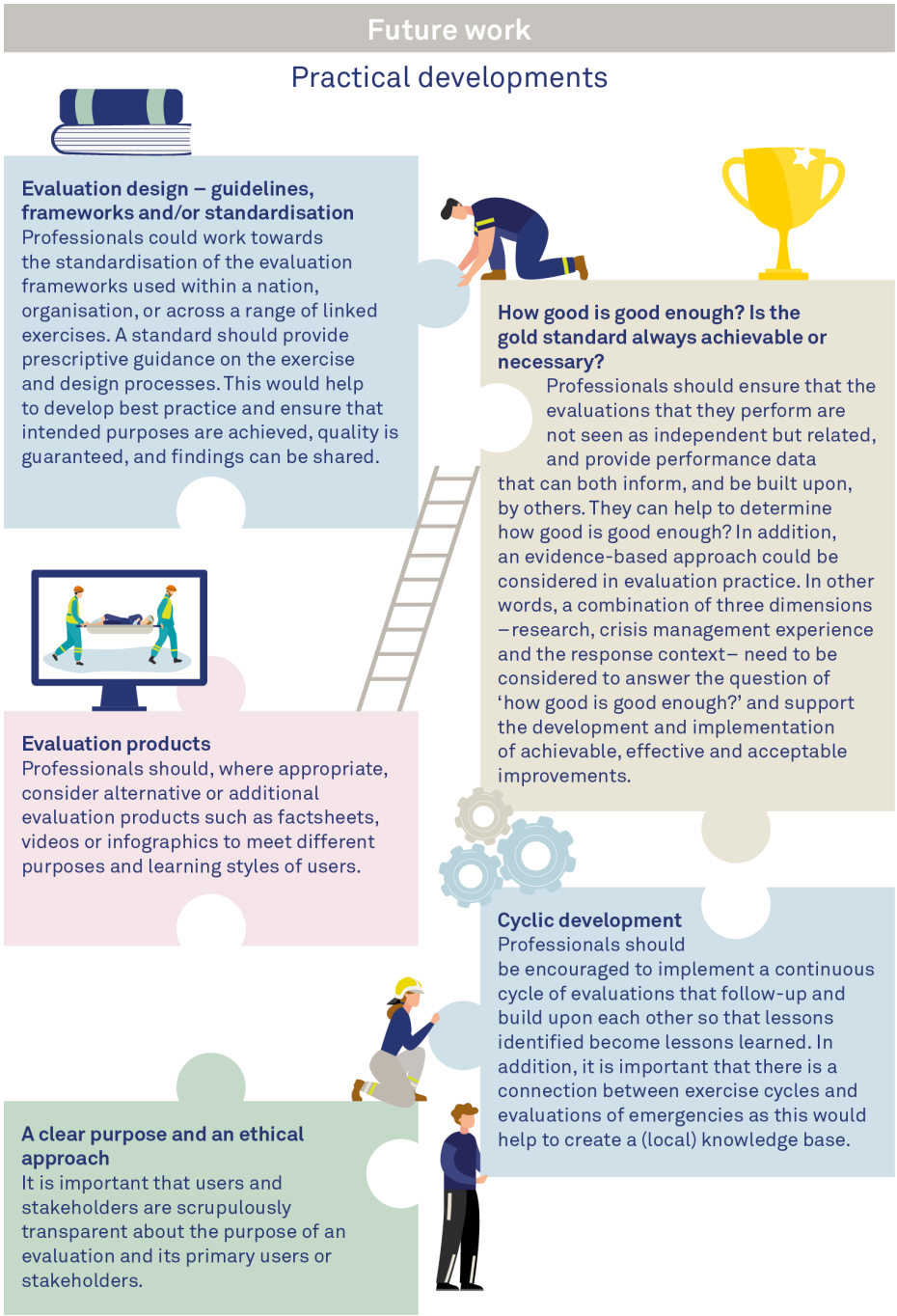


Figure 14: Visual presentation of future work
The figure presents some suggestions for future work. The aim is to encourage its use, in line with the proposals put forward in this section (mainly 7.2.3).

8 Conclusion

This thesis extends our understanding of the use of evaluation as a tool to support the development and improvement of DRM. The overall project was divided into sequential steps: 1) the scientific literature and documents from Dutch practice were analysed. This material provided a starting point from which to explore and describe the current status of the formal evaluation of simulated, coordinated, multi-organisational responses to emergencies, disasters and crises; 2) crisis management professionals were consulted to investigate the usefulness of evaluation products; and 3) insights from research and practice were combined to propose future developments to improve and structure the evaluation process, and increase the usefulness of the evaluation product for the end user, in line with the overarching RQ.

The main conclusions are presented below, based on the research questions outlined in Chapter 1:

Conclusion I: *There is a lack of coherent, cohesive and systematic scientific attention given to building a solid knowledge base that could support best practice in the design of effective and useful evaluations for both simulated and real disaster events. This has a direct impact on the structured improvement of the DRM process.*

Evaluation is only briefly touched upon in the broader disaster preparedness context. This literature does not provide in-depth information regarding the selected methods, or link them to the overall purpose of a specific crisis management exercise. Studies do not address questions such as why evaluations are being performed? What methods or frameworks are being used? Or when and how they are successful? Consequently, professionals lack prescriptive, reliable and valid guidance to support them in designing exercise evaluations. Furthermore, the field of (simulated) disaster response evaluation science is fragmented, and lacks coherence and depth. There is limited scientific development in this area, with few insights and little progress that could support practice (RQ1).

This research showed that, in the Netherlands, a variety of evaluation designs are used on a regional and local level. Here again, there is limited guidance to justify the selection

of a particular method. In particular the analysis of contextual data to reach conclusions lacks methodological transparency, justification or reasoning. Despite the cyclic nature of the various types of evaluations (exercises, tests and real responses), they are seen as independent activities (RQ2) and do not build upon each other. This has implications for nationwide learning and the ongoing development of crisis management. The analysis also revealed that the reviewed reports did not clearly indicate a target audience, and did not make clear how the evaluation's design (why, what and how) made it relevant and useful for its audience. At this point, it remains unclear how effective, or useful, current evaluations are in achieving their purpose, and how they contribute to the development of disaster preparedness (RQ1 and RQ2).

Guided by this outcome, the focus shifted from the literature/ documents to the real world. Here, crisis management professionals participated in survey experiments in order to investigate and enhance usefulness.

Conclusion II: *This research showed that the way evaluations of (simulated) disaster responses are documented and presented to users influences their perceived usefulness. This perception can be enhanced by the use of a user-focused, clear and rigorous approach to documentation, the presentation of the analysis, and/ or the actionability of the conclusions.*

The next step of the research built on earlier outcomes. More specifically, it used real-world examples, along with input from crisis management professionals, to investigate what aspects of usefulness are influential. Paper III introduced the notion of an evaluation description, which encompasses four components that are assumed to influence the usefulness of an evaluation: purpose (P), object description (O), analysis (A) and conclusion (C) (RQ3a, b). These components are similar to the themes that were identified in Paper IV through the thematic analysis of expectations amongst professionals in the Dutch crisis management system.

The survey experiment (Paper III) demonstrated that the way evaluations of emergency exercises are documented influences their usefulness. In particular, the analysis and/ or conclusions appear to have a significant effect on perceptions (RQ3a). More precisely, it was found that the clarity of conclusions was particularly important, as this finding held for both operational and governing users, and for both learning and accountability purposes. This outcome was supported by the datasets that were analysed in Paper IV. Here, crisis management professionals highlighted that a rigorous analysis should go beyond the object of the evaluation, and take into account its context. Furthermore, they noted that the evaluation should provide them with actionable conclusions in order to support future learning (RQ3b).

Both Papers III and IV demonstrated the importance of the evaluation product. These insights were used to propose a framework for the design and execution of exercise evaluation.

Conclusion III: *There is a need to develop models, frameworks or even generic standards that contain clear components, or minimum requirements, that support evaluation designers. These tools should form the basis for the collection of evidence-based feedback on the outcomes of the operational response. This would support a cyclic connection between evaluation outcomes, preparedness training and the optimisation of resources.*

The final step is documented in Chapters 6 and 7 of this thesis. The results from previous studies are used to develop models or frameworks that guide the design, execution and reporting of evaluations (RQ4). These tools can help professionals to design and execute an evaluation and related exercise, and deliver products that have clear conclusions. Furthermore, they can be used prescriptively to standardise the process. Here, it is important to acknowledge that evaluations are not an end in themselves, but are a useful research and analysis tool that can deliver evidence-based capability and gap analyses, at all levels of the responder network, in a simulated or real-world emergency.

Ideally, the evaluation should be a tool that supports DRM preparedness and response, founded on evidence-based recommendations that focus resources on learning and/ or accountability. This research emphasises the importance of the systematic and rigorous evaluation of operational responses, whether in a simulated or actual situation. Although evaluation should not be seen as a holy grail, a well-designed strategy that draws upon naturally-occurring evidence will contribute to improved DRM preparedness and response. Well-structured evaluations can identify areas that meet existing standards and those that require improvement, thereby helping organisations to know where to focus finite time, material, human and financial resources.

Overall conclusion: *Evaluation should be seen as a tool with great potential for the DRM community. As a means to an end, both theory and practice should work together to improve perceptions of usefulness. This important step forward would help to deliver and support evidence-based learning that, in turn, would help responders to learn from the past, and better-prepare for the future.*

Finally, the insights, conclusions and recommendations presented here can influence the impact of evaluation at various levels. This research explored operational disaster response evaluation, primarily during simulated events, but also actual emergencies. It

focused on professionals and first-responders: individuals, teams, organisations or systems. The outcomes provide a clear point of departure for both theory and practice. The findings will enhance the usefulness of evaluation within emergency, disaster and crisis management, ensuring that all responders and other stakeholders can learn from their experience, and apply this learning to better-prepare for the next event.

References

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), 14–25. https://doi.org/10.1111/j.1468-5973.2009.00601.x

Agboola, F., McCarthy, T., & Biddinger, P. D. (2013). Impact of Emergency Preparedness Exercise on Performance. *Journal of Public Health Management and Practice, 19*(5), S77–S83. https://doi.org/10.1097/PHH.0b013e31828ecd84

Alexander, D. E. (2000). Scenario methodology for teaching principles of emergency management. *Disaster Prevention and Management: An International Journal, 9*(2), 89–97. https://doi.org/10.1108/09653560010326969

Alexander, D. E. (2002). *Principles of emergency planning and management.* Oxford University Press.

Alexander, D. E. (2015). Evaluation of civil protection programmes, with a case study from Mexico. *Disaster Prevention and Management: An International Journal, 24*(2), 263–283. https://doi.org/10.1108/DPM-12-2014-0268

Ansell, C., Boin, A., & Keller, A. (2010). Managing Transboundary Crises: Identifying the Building Blocks of an Effective Response System. *Journal of Contingencies and Crisis Management, 18*(4), 195–207. https://doi.org/10.1111/j.1468-5973.2010.00620.x

Argyris, C. (1976). Single-Loop and Double-Loop Models in Research on Decision Making. *Administrative Science Quarterly, 21*(3), 363-375. https://doi.org/10.2307/2391848

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology, 6*(3), 128–138. https://doi.org/10.1027/1614-2241/a000014

Auspurg, K., & Hinz, T. (2015). *Factorial Survey Experiments.* SAGE Publications, Inc. https://doi.org/10.4135/9781483398075

Aven, T. (2010). On how to define, understand and describe risk. *Reliability Engineering & System Safety, 95*(6), 623–631. https://doi.org/10.1016/j.ress.2010.01.011

Aven, T. (2011). On Some Recent Definitions and Analysis Frameworks for Risk, Vulnerability, and Resilience. *Risk Analysis, 31*(4), 515–522. https://doi.org/10.1111/j.1539-6924.2010.01528.x

Aven, T. (2012). The risk concept—historical and recent development trends. *Reliability Engineering & System Safety, 99*(0951), 33–44. https://doi.org/10.1016/j.ress.2011.11.006

Aven, T., & Renn, O. (2009). On risk defined as an event where the outcome is uncertain. *Journal of Risk Research, 12*(1), 1–11. https://doi.org/10.1080/13669870802488883

Aven, T., Renn, O., & Rosa, E. A. (2011). On the ontological status of the concept of risk. *Safety Science, 49*(8–9), 1074–1079. https://doi.org/10.1016/j.ssci.2011.04.015

Baskerville, R., & Pries-Heje, J. (2010). Explanatory Design Theory. *Business & Information Systems Engineering, 2*(5), 271–282. https://doi.org/10.1007/s12599-010-0118-4

Baxter, P., & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report, 13*(4), 544–559. https://doi.org/10.46743/2160-3715/2008.1573

Beerens, R., Abraham, P., Glerum, P., & Kolen, B. (2014). Flood Preparedness Training and Exercises. In J. J. L. M. Bierens (Ed.), *Drowning* (2nd ed., pp. 1009–1016). Springer-Verlag. https://doi.org/10.1007/978-3-642-04253-9_154

Beerens, R. J. J. (2019). Does the means achieve an end? A document analysis providing an overview of emergency and crisis management evaluation practice in the Netherlands. *International Journal of Emergency Management, 15*(3), 221–254. https://doi.org/10.1504/IJEM.2019.102310

Beerens, R. J. J., Abraham, P., & Braakhekke, E. (2012). Maximise your returns in crisis management preparedness: A cyclic approach to training and exercises. *Proceedings of the 4th International Disaster and Risk Conference: Integrative Risk Management in a Changing World - Pathways to a Resilient Society, IDRC Davos 2012*, 62–66. https://doi.org/10.13140/2.1.1138.0482

Beerens, R. J. J., & Haverhoek-Mieremet, K. (2021). What do practitioners expect from an evaluation report? A qualitative analysis of Dutch crisis management professionals' expectations. *International Journal of Emergency Services, 10*(1), 1–25. https://doi.org/10.1108/IJES-12-2019-0063

Beerens, R. J. J., Kolen, B., & Helsloot, I. (2010). EU FloodEx 2009: An analysis of testing international assistance during a worst credible flood scenario in the North Sea area. *WIT Transactions on Ecology and the Environment, 133*, 241–255. https://doi.org/10.2495/FRIAR100211

Beerens, R. J. J., & Tehler, H. (2016). Scoping the field of disaster exercise evaluation - A literature overview and analysis. *International Journal of Disaster Risk Reduction, 19*, 413–446. https://doi.org/10.1016/j.ijdrr.2016.09.001

Beerens, R. J. J., Tehler, H., & Pelzer, B. (2020). How Can We Make Disaster Management Evaluations More Useful? An Empirical Study of Dutch Exercise Evaluations. *International Journal of Disaster Risk Science, 11*(5), 578–591. https://doi.org/10.1007/s13753-020-00286-7

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open, 2*, 8–14. https://doi.org/10.1016/j.npls.2016.01.001

Berlin, J. M., & Carlström, E. D. (2014). Collaboration Exercises—The Lack of Collaborative Benefits. *International Journal of Disaster Risk Science, 5*(3), 192–205. https://doi.org/10.1007/s13753-014-0025-2

Biddinger, P. D., Cadigan, R. O., Auerbach, B. S., Burstein, J. L., Savoia, E., Stoto, M. A., & Koh, H. K. (2008). On Linkages Using Exercises to Identify Systems-Level Preparedness Challenges. *Public Health Reports, 123*(1), 96–101. https://doi.org/10.1177/003335490812300116

Biddinger, P. D., Savoia, E., Massin-Short, S. B., Preston, J., & Stoto, M. A. (2010). Public Health Emergency Preparedness Exercises: Lessons Learned. *Public Health Reports, 125*(5_suppl), 100–106. https://doi.org/10.1177/00333549101250S514

Birkland, T. A. (2009). Disasters, Lessons Learned, and Fantasy Documents. *Journal of Contingencies and Crisis Management, 17*(3), 146–156. https://doi.org/10.1111/j.1468-5973.2009.00575.x

Boin, A. (2009). The New World of Crises and Crisis Management: Implications for Policymaking and Research. *Review of Policy Research, 26*(4), 367–377. https://doi.org/10.1111/j.1541-1338.2009.00389.x

Boin, A., & 't Hart, P. (2003). Public Leadership in Times of Crisis: Mission Impossible? *Public Administration Review, 63*(5), 544–553. https://doi.org/10.1111/1540-6210.00318

Boin, A., 't Hart, P., Stern, E., & Sundelius, B. (2017). *The Politics of Crisis Management* (2nd Revise). Cambridge University Press. https://doi.org/10.1017/9781316339756

Boin, A., Ekengren, M., & Rhinard, M. (2020). Hiding in Plain Sight: Conceptualizing the Creeping Crisis. *Risk, Hazards & Crisis in Public Policy, 11*(2), 116–138. https://doi.org/10.1002/rhc3.12193

Boin, A., McConnell, A., & 't Hart, P. (2008). Governing after crisis. *Governing after Crisis: The Politics of Investigation, Accountability and Learning, July 2017*, 3–30. https://doi.org/10.1017/CBO9780511756122.001

Borell, J., & Eriksson, K. (2008). Improving emergency response capability: an approach for strengthening learning from emergency response evaluations. *International Journal of Emergency Management, 5*(3/4), 324-337. https://doi.org/10.1504/IJEM.2008.025101

Borodzicz, E., & Van Haperen, K. (2002). Individual and Group Learning in Crisis Simulations. *Journal of Contingencies and Crisis Management, 10*(3), 139–147. https://doi.org/10.1111/1468-5973.00190

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal, 13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bovens, M. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics, 33*(5), 946–967. https://doi.org/10.1080/01402382.2010.486119

Bovens, M., 't Hart, P., & Kuipers, S. (2008). The Politics of Policy Evaluation. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford Handbook of Public Policy* (Issue June 2020). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199548453.003.0015

Bowen, A. A. (2008). Are We Really Ready? The Need for National Emergency Preparedness Standards and the Creation of the Cycle of Emergency Planning 1. *Politics & Policy, 36*(5), 834–853. https://doi.org/10.1111/j.1747-1346.2008.00137.x

Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Brainich von Brainich Felth, E. T. (2004). *Het systeem van crisisbeheersing.* Boom Juridische uitgevers.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica, 81*(3), 211–241. https://doi.org/10.1016/0001-6918(92)90019-A

Brehmer, B. (2007). Understanding the Functions of C2 Is the Key to Progress. *The International C2 Journal, 1*(1), 211–232.

Broekema, W., Porth, J., Steen, T., & Torenvlied, R. (2019). Public leaders' organizational learning orientations in the wake of a crisis and the role of public service motivation. *Safety Science, 113*(March 2018), 200–209. https://doi.org/10.1016/j.ssci.2018.11.002

Bryman, A. (2016). *Social Research Methods* (5th ed.). Oxford University Press.

Callan, T. (2009). So, you want to run an exercise? *Australian Journal of Emergency Management, 24*(2), 59–62. https://ajem.infoservices.com.au/items/AJEM-24-02-10

Cassidy, S. (2004). Learning Styles: An overview of theories, models, and measures. *Educational Psychology, 24*(4), 419–444. https://doi.org/10.1080/0144341042000228834

Clarke, L. (1999). *Mission Improbable: Using Fantasy Documents to Tame Disaster*. University of Chicago Press.

Clarke, L. (2005). *Worst Cases - Terror and Catastrophe in the Popular Imagination*. University of Chicago Press.

Coetzee, C., & Van Niekerk, D. (2012). Tracking the evolution of the disaster management cycle: A general system theory approach. *Jàmbá: Journal of Disaster Risk Studies, 4*(1), 9 pages. https://doi.org/10.4102/jamba.v4i1.54

Cranmer, H., Chan, J. L., Kayden, S., Musani, A., Gasquet, P. E., Walker, P., Burkle, F. M., & Johnson, K. (2014). Development of an Evaluation Framework Suitable for Assessing Humanitarian Workforce Competencies During Crisis Simulation Exercises. *Prehospital and Disaster Medicine, 29*(1), 69–74. https://doi.org/10.1017/S1049023X13009217

Creswell, J. W. (2003). *Research Design - Qualitative, Quantitative, and Mixed Methods Approaches* (2nd ed.). SAGE Publications Inc.

Crotty, M. (1998). *The foundations of social research*. Allen & Unwin.

Daudt, H. M., Van Mossel, C., & Scott, S. J. (2013). Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Medical Research Methodology, 13*(1), 48. https://doi.org/10.1186/1471-2288-13-48

Dausey, D. J., & Moore, M. (2014). Using exercises to improve public health preparedness in Asia, the Middle East and Africa. *BMC Research Notes, 7*(1), 474. https://doi.org/10.1186/1756-0500-7-474

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319-340. https://doi.org/10.2307/249008

Denyer, D., Tranfield, D., & van Aken, J. E. (2008). Developing Design Propositions through Research Synthesis. *Organization Studies, 29*(3), 393–413. https://doi.org/10.1177/0170840607088020

Deverell, E. (2012). Is best practice always the best? Learning to become better crisis managers. *Journal of Critical Incident Analysis, 3*(1), 26–40. http://jcia.aciajj.org/files/2012/12/Deverell-2-Final.pdf

Djalali, A., Carenzo, L., Ragazzoni, L., Azzaretto, M., Petrino, R., Della Corte, F., & Ingrassia, P. L. (2014). Does Hospital Disaster Preparedness Predict Response Performance During a Full-scale Exercise? A Pilot Study. *Prehospital and Disaster Medicine, 29*(5), 441–447. https://doi.org/10.1017/S1049023X1400082X

Downe-Wamboldt, B. (1992). Content analysis: Method, applications, and issues. *Health Care for Women International, 13*(3), 313–321. https://doi.org/10.1080/07399339209516006

Driscoll, M. P. (1994). *Psychology of learning for instruction*. Allyn & Bacon.

Easton, G. (2010). Critical realism in case study research. *Industrial Marketing Management, 39*(1), 118–128. https://doi.org/10.1016/j.indmarman.2008.06.004

Eriksson, K. (2010). *Preparing for Preparedness - Shaping Crisis Planning Processes in Local Authorities*. [Lund University]. https://lup.lub.lu.se/search/publication/dfd3d8dc-39f8-4557-8e91-081fa889b9a0

Forss, K., Rebien, C. C., & Carlsson, J. (2002). Process Use of Evaluations. *Evaluation, 8*(1), 29–45. https://doi.org/10.1177/1358902002008001515

Franz, C. R., & Robey, D. (1986). Organizational context, user involvement, and the usefulness of information systems. *Decision Sciences, 17*(3), 329–356. https://doi.org/10.1111/j.1540-5915.1986.tb00230.x

Gebbie, K. M., Valas, J., Merrill, J., & Morse, S. (2006). Role of Exercises and Drills in the Evaluation of Public Health in Emergency Response. *Prehospital and Disaster Medicine, 21*(3), 173–182. https://doi.org/10.1017/S1049023X00003642

Gov.uk. (2021). *Emergency Planning and preparedness: exercises and training*. https://www.gov.uk/guidance/emergency-planning-and-preparedness-exercises-and-training

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth Generation Evaluation*. SAGE Publications Inc.

Guha-Sapir, D. (2020). *EM-DAT: The Emergency Events Database - Université catholique de Louvain (UCL) - CRED*. Brussels, Belgium. www.emdat.be

Haimes, Y. Y. (2009). On the Complex Definition of Risk: A Systems-Based Approach. *Risk Analysis, 29*(12), 1647–1654. https://doi.org/10.1111/j.1539-6924.2009.01310.x

Hammersley, M., & Atkinson, P. (2019). *Ethnography: Principles in Practice*. (4th ed.). Routledge Taylor & Francis Group. https://doi.org/10.4324/9781315146027

Hansen, M. B., & Vedung, E. (2010). Theory-Based Stakeholder Evaluation. *American Journal of Evaluation, 31*(3), 295–313. https://doi.org/10.1177/1098214010366174

Hansson, S. O., & Aven, T. (2014). Is Risk Analysis Scientific? *Risk Analysis, 34*(7), 1173–1183. https://doi.org/10.1111/risa.12230

Heath, R. (1998). Looking for answers: suggestions for improving how we evaluate crisis management. *Safety Science, 30*(1–2), 151–163. https://doi.org/10.1016/S0925-7535(98)00043-5

Hendrickson, A. R., Massey, P. D., & Cronan, T. P. (1993). On the Test-Retest Reliability of Perceived Usefulness and Perceived Ease of Use Scales. *MIS Quarterly, 17*(2), 227-230. https://doi.org/10.2307/249803

Hertting, N., & Vedung, E. (2012). Purposes and criteria in network governance evaluation: How far does standard evaluation vocabulary takes us? *Evaluation, 18*(1), 27–46. https://doi.org/10.1177/1356389011431021

Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems* (Vol. 22). Springer US. https://doi.org/10.1007/978-1-4419-5653-8

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly, 28*(1), 75. https://doi.org/10.2307/25148625

Hodgkinson, G. P., & Rousseau, D. M. (2009). Bridging the Rigour-Relevance Gap in Management Research: It's Already Happening! *Journal of Management Studies, 46*(3), 534–546. https://doi.org/10.1111/j.1467-6486.2009.00832.x

Hsu, E. B., Jenckes, M. W., Catlett, C. L., Robinson, K. A., Feuerstein, C., Cosgrove, S. E., Green, G. B., & Bass, E. B. (2004). Effectiveness of Hospital Staff Mass-Casualty

Incident Training Methods: A Systematic Literature Review. *Prehospital and Disaster Medicine, 19*(3), 191–199. https://doi.org/10.1017/S1049023X00001771

Hughes, R. T. (1996). Expert judgement as an estimating method. *Information and Software Technology, 38*(2), 67–75. https://doi.org/10.1016/0950-5849(95)01045-9

Hunter, J. C., Yang, J. E., Petrie, M., & Aragón, T. J. (2012). Integrating a framework for conducting public health systems research into statewide operations-based exercises to improve emergency preparedness. *BMC Public Health, 12*(1), 680. https://doi.org/10.1186/1471-2458-12-680

Hurteau, M., Houle, S., & Mongiat, S. (2009). How Legitimate and Justified are Judgments in Program Evaluation? *Evaluation, 15*(3), 307–319. https://doi.org/10.1177/1356389009105883

Hutchinson, B., Dekker, S., & Rae, A. (2018). Fantasy planning: the gap between systems of safety and safety of systems. *Australian System Safety Conference*.

IBM Corp. (2019). *SPSS* (No. 25).

Ingrassia, P. L., Prato, F., Geddo, A., Colombo, D., Tengattini, M., Calligaro, S., La Mura, F., Michael Franc, J., & Della Corte, F. (2010). Evaluation of Medical Management During a Mass Casualty Incident Exercise: An Objective Assessment Tool to Enhance Direct Observation. *The Journal of Emergency Medicine, 39*(5), 629–636. https://doi.org/10.1016/j.jemermed.2009.03.029

Johnson, P., & Duberley, J. (2000). *Understanding Management Research*. SAGE Publications Ltd. https://doi.org/10.4135/9780857020185

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher, 33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Joint Committee on Standards for Educational Evaluation (1994). *The program evaluation standards* (2nd ed.). Corwin Press.

Jongejan, R. B., Helsloot, I., Beerens, R. J. J., & Vrijling, J. K. (2011). How prepared is prepared enough? *Disasters, 35*(1), 130–142. https://doi.org/10.1111/j.1467-7717.2010.01196.x

Kaji, A. H., & Lewis, R. J. (2008). Assessment of the Reliability of the Johns Hopkins / Agency for Healthcare Research and Quality Hospital Disaster Drill Evaluation Tool. *Annals of Emergency Medicine, 52*(3), 204–210, 210.e1-8. https://doi.org/10.1016/j.annemergmed.2007.10.026

Kapucu, N., & Van Wart, M. (2006). The Evolving Role of the Public Sector in Managing Catastrophic Disasters. *Administration & Society, 38*(3), 279–308. https://doi.org/10.1177/0095399706289718

Karahanna, E., & Straub, D. W. (1999). The psychological origins of perceived usefulness and ease-of-use. *Information & Management, 35*(4), 237–250. https://doi.org/10.1016/S0378-7206(98)00096-2

Kelly, C. (1995). A framework for improving operational effectiveness and costefficiency in emergency planning and response. *Disaster Prevention and Management: An International Journal, 4*(3), 25–31. https://doi.org/10.1108/09653569510088041

Khan, H., Vasilescu, L. G., & Khan, A. (2008). Disaster Management Cycle – a Theoretical Approach. *Management & Marketing, 6*(1), 43–50.

Kieser, A., & Leiner, L. (2009). Why the Rigour-Relevance Gap in Management Research Is Unbridgeable. *Journal of Management Studies, 46*(3), 516–533. https://doi.org/10.1111/j.1467-6486.2009.00831.x

Kim, H. (2013). Improving simulation exercises in Korea for disaster preparedness. *Disaster Prevention and Management: An International Journal, 22*(1), 38–47. https://doi.org/10.1108/09653561311301961

Kirschenbaum, A. (2003). *Chaos Organization and Disaster Management*. Routledge.

Kitson, A., Harvey, G., & McCormack, B. (1998). Enabling the implementation of evidence based practice: a conceptual framework. *Quality and Safety in Health Care, 7*(3), 149–158. https://doi.org/10.1136/qshc.7.3.149

Kitzinger, J. (1995). Qualitative Research: Introducing focus groups. *BMJ, 311*, 299–302. https://doi.org/10.1136/bmj.311.7000.299

Klein, K. R., Brandenburg, D. C., Atas, J. G., & Maher, A. (2005). The Use of Trained Observers as an Evaluation Tool for a Multi-Hospital Bioterrorism Exercise. *Prehospital and Disaster Medicine, 20*(3), 159–163. https://doi.org/10.1017/S1049023X00002387

Kothari, C. R. (2004). *Research Methodology - Methods and Techniques* (2nd ed.). New Age International Publishers.

Lachman, S. J. (1997). Learning is a Process: Toward an Improved Definition of Learning. *The Journal of Psychology, 131*(5), 477–480. https://doi.org/10.1080/00223989709603535

Latiers, M., & Jacques, J. M. (2009). Emergency and crisis exercises: methodology for understanding safety dimensions. *International Journal of Emergency Management, 6*(1), 73-84. https://doi.org/10.1504/IJEM.2009.025174

Lawrence, J. E. S., & Cook, T. J. (1982). Designing useful evaluations: The stakeholder survey. *Evaluation and Program Planning,* 5(4), 327–336. https://doi.org/10.1016/0149-7189(82)90005-2

Lecy, J. D., & Beatty, K. E. (2012). Representative Literature Reviews Using Constrained Snowball Sampling and Citation Network Analysis. *SSRN Electronic Journal*, 1–15. https://doi.org/10.2139/ssrn.1992601

Letourneau, C. U. (1962). Evaluation of a hospital disaster plan. 2. A general critique of a disaster drill. *Hospital Management, 94*, 52–55.

Lindbom, H. (2020). *Improving capability assessments for disaster risk management* [Lund University]. https://lup.lub.lu.se/record/b31abdce-8185-472f-9904-783a173c1c7b

Lindbom, H., Tehler, H., Eriksson, K., & Aven, T. (2015). The capability concept – On how to define and describe capability in relation to risk, vulnerability and resilience. *Reliability Engineering & System Safety, 135*, 45–54. https://doi.org/10.1016/j.ress.2014.11.007

Lindell, M. K. (2013a). Emergency management. In P. T. Bobrowsky (Ed.), *Encyclopedia of Natural Hazards* (pp. 263–271). Springer. https://doi.org/10.1007/978-1-4020-4399-4

Lindell, M. K. (2013b). Recovery and reconstruction after disaster. In P. T. Bobrowsky (Ed.), *Encyclopedia of Natural Hazards* (pp. 812–824). Springer. https://doi.org/10.1007/978-1-4020-4399-4_285

Lonka, H., & Wybo, J. L. (2005). Sharing of experiences: a method to improve usefulness of emergency exercises. *International Journal of Emergency Management, 2*(3), 189-202. https://doi.org/10.1504/IJEM.2005.007359

MacInnis, D. J. (2011). A Framework for Conceptual Contributions in Marketing. *Journal of Marketing, 75*(4), 136–154. https://doi.org/10.1509/jmkg.75.4.136

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

McConnell, A., & Drennan, L. (2006). Mission Impossible? Planning and Preparing for Crisis. *Journal of Contingencies and Crisis Management, 14*(2), 59–70. https://doi.org/10.1111/j.1468-5973.2006.00482.x

McEntire, D. A. (2007). *Disaster Response and Recovery*. John Wiley & Sons, Inc.

Mearns, S. L. (2011). Pragmatic critical realism: Could this methodological approach expand our understanding of employment relations? *Work, 38*(4), 359–367. https://doi.org/10.3233/WOR-2011-1139

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2004). *Beleidsplan Crisisbeheersing 2004-2007*. https://zoek.officielebekendmakingen.nl/kst-29668-1.html

Ministry of Security and Justice (2013). *Safety Regions Act.* https://www.government.nl/documents/decrees/2010/12/17/dutch-security-regions-act-part-i

Morris, J. G., Greenspan, A., Howell, K., Gargano, L. M., Mitchell, J., Jones, J. L., Potter, M., Isakov, A., Woods, C., & Hughes, J. M. (2012). Southeastern Center for Emerging Biologic Threats Tabletop Exercise: Foodborne Toxoplasmosis Outbreak on College Campuses. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science, 10*(1), 89–97. https://doi.org/10.1089/bsp.2011.0040

Nilsson, H., & Rüter, A. (2008). Management of resources at major incidents and disasters in relation to patient outcome: a pilot study of an educational model. European *Journal of Emergency Medicine, 15*(3), 162–165. https://doi.org/10.1097/MEJ.0b013e3282f4d14b

Nolan, P., & Walsh, J. (1995). The Structure of the Economy and Labour Market. In P. Edwards (Ed.), *Industrial Relations: Theory and Practice in Britain*. Blackwell Publishing.

Overheid.nl (2017). *Besluit Veiligheidsregio's.* http://wetten.overheid.nl/BWBR0027844/2017-12-01

Overheid.nl (2021). *Rijkswet Onderzoeksraad voor veiligheid.* https://wetten.overheid.nl/BWBR0017613/2021-01-01

Payne, C. F. (1999). Contingency plan exercises. *Disaster Prevention and Management: An International Journal, 8*(2), 111–117. https://doi.org/10.1108/09653569910266157

Perrow, C. (1994). The Limits of Safety: The Enhancement of a Theory of Accidents. *Journal of Contingencies and Crisis Management, 2*(4), 212–220. https://doi.org/10.1111/j.1468-5973.1994.tb00046.x

Perrow, C. (1999). Organizing to Reduce the Vulnerabilities of Complexity. *Journal of Contingencies and Crisis Management, 7*(3), 150–155. https://doi.org/10.1111/1468-5973.00108

Perry, R. W. (2004). Disaster Exercise Outcomes for Professional Emergency Personnel and Citizen Volunteers. *Journal of Contingencies and Crisis Management, 12*(2), 64–75. https://doi.org/10.1111/j.0966-0879.2004.00436.x

Perry, R. W., & Lindell, M. K. (2007). *Emergency Planning*. John Wiley & Sons, Inc.

Peterson, D. M., & Perry, R. W. (1999). The impacts of disaster exercises on participants. *Disaster Prevention and Management: An International Journal, 8*(4), 241–255. https://doi.org/10.1108/09653569910283879

Pettigrew, A. M. (2001). Management Research After Modernism. *British Journal of Management, 12*(s1), S61–S70. https://doi.org/10.1111/1467-8551.12.s1.8

Powell, R. A., & Single, H. M. (1996). Focus Groups. *International Journal for Quality in Health Care, 8*(5), 499–504. https://doi.org/10.1093/intqhc/8.5.499

Powell, R. R. (2017). Evaluation Research: An Overview. *Library Trends, 55*(1), 102–120. https://doi.org/10.1353/lib.2006.0050

Rådestad, M., Nilsson, H., Castrén, M., Svensson, L., Rüter, A., & Gryth, D. (2012). Combining performance and outcome indicators can be used in a standardized way: a pilot study of two multidisciplinary, full-scale major aircraft exercises. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 20*(1), 58. https://doi.org/10.1186/1757-7241-20-58

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decisionmaking and system management. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15*(2), 234–243. https://doi.org/10.1109/TSMC.1985.6313353

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety Science, 27*(2–3), 183–213. https://doi.org/10.1016/S0925-7535(97)00052-0

Remenyi, D., Williams, B., Money, A., & Swartz, E. (1998). *Doing Research in Business and management: An introduction to Process and Method*. Sage.

Rimbo-Gilde.nl (2021). *Het stelsel van veiligheidsmanagement: referentiemodellen*. https://rimbo-gilde.nl/wp-content/uploads/2021/04/REFERENTIEMODELLEN-VEILIGHEIDSMANAGEMENT-APRIL-2021.pdf

Ritchie, L. A., & MacDonald, W. (2010). Evaluation of disaster and emergency management: Do no harm, but do better. *New Directions for Evaluation, 2010*(126), 107–111. https://doi.org/10.1002/ev.333

Romme, A. G. L. (2003). Making a Difference: Organization as Design. *Organization Science, 14*(5), 558–573. https://doi.org/10.1287/orsc.14.5.558.16769

Runeson, P., Host, M., Rainer, A., & Regnell, B. (2012). *Case Study Research in Software Engineering: Guidelines and Examples*. John Wiley & Sons, Inc.

Rüter, A., Nilsson, H., & Vilkström, T. (2006). Performance Indicators as Quality Control for Testing and Evaluating Hospital Management Groups: A Pilot Study. *Prehospital and Disaster Medicine, 21*(6), 423–426. https://doi.org/10.1017/S1049023X00004131

Savoia, E., Agboola, F., & Biddinger, P. (2014). A Conceptual Framework to Measure Systems' Performance during Emergency Preparedness Exercises. *International Journal of Environmental Research and Public Health, 11*(9), 9712–9722. https://doi.org/10.3390/ijerph110909712

Savoia, E., Biddinger, P. D., Burstein, J., & Stoto, M. A. (2010). Inter-agency communication and operations capabilities during a hospital functional exercise: reliability and validity of a measurement tool. *Prehospital and Disaster Medicine, 25*(1), 52–58. https://doi.org/10.1017/S1049023X00007664

Savoia, E., Preston, J., & Biddinger, P. D. (2013). A Consensus Process on the Use of Exercises and After Action Reports to Assess and Improve Public Health Emergency Preparedness and Response. *Prehospital and Disaster Medicine, 28*(3), 305–308. https://doi.org/10.1017/S1049023X13000289

Savoia, E., Testa, M. A., Biddinger, P. D., Cadigan, R. O., Koh, H., Campbell, P., & Stoto, M. A. (2009). Assessing Public Health Capabilities during Emergency Preparedness Tabletop Exercises: Reliability and Validity of a Measurement Tool. *Public Health Reports, 124*(1), 138–148. https://doi.org/10.1177/003335490912400117

Sayer, A. (2000). *Realism and Social Science*. SAGE Publications Ltd.

Scholtens, A. (2008). Controlled Collaboration in Disaster and Crisis Management in the Netherlands, History and Practice of an Overestimated and Underestimated Concept. *Journal of Contingencies and Crisis Management, 16*(4), 195–207. https://doi.org/10.1111/j.1468-5973.2008.00550.x

Scientific Software Development (2019). *Atlas.ti* (version 8.2.32). https://atlasti.com/

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Rand McNally.

Scriven, M. (1980). The logic of evaluation. Edgepress.

Scriven, M. (1991). *Evaluation thesaurus*. SAGE Publications.

Scriven, M. (1993). Author's notes. *New Directions for Program Evaluation, 1993*(58), 1–4. https://doi.org/10.1002/ev.1639

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action Design Research. *MIS Quarterly, 35*(1), 37-56. https://doi.org/10.2307/23043488

Shadish Jr., W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Sage Publications, Inc.

Simon, H. A. (1969). *The sciences of the artificial*. MIT Press.

Sinclair, H., Doyle, E. E., Johnston, D. M., & Paton, D. (2012). Assessing emergency management training and exercises. *Disaster Prevention and Management: An International Journal, 21*(4), 507–521. https://doi.org/10.1108/09653561211256198

Skryabina, E., Reedy, G., Amlôt, R., Jaye, P., & Riley, P. (2017). What is the value of health emergency preparedness exercises? A scoping review study. *International Journal of Disaster Risk Reduction, 21*, 274–283. https://doi.org/10.1016/j.ijdrr.2016.12.010

Skryabina, E., Riley, P., Reedy, G., & Amlôt, R. (2018). A scoping review of evaluation methods for health emergency preparedness exercises. *American Journal of Disaster Medicine, 13*(2), 107–127. https://doi.org/10.5055/ajdm.2018.0292

Stake, R. E. E. (1974). *Evaluating the arts in education: A responsive approach*. Merrill.

Stemler, S. E. (2001). An Introduction to Content Analysis. *ERIC Digest*, 1–7. www.eric.ed.gov

Stufflebeam, D. L. (2003). The CIPP Model for Evaluation. In *International Handbook of Educational Evaluation* (pp. 31–62). Springer Netherlands. https://doi.org/10.1007/978-94-010-0309-4_4

Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, and applications* (2nd ed.). Jossey-Bass.

Swedish Civil Contingencies Agency (MSB) (2017). *Exercise guidance: basic manual - an introduction to the fundamentals of exercise planning*. Swedish Civil Contingencies Agency (MSB). https://www.msb.se/sv/publikationer/exercise-guidance--basic-manual--an-introduction-to-the-fundamentals-of-exercise-planning/

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation, 27*(2), 237–246. https://doi.org/10.1177/1098214005283748

Thomas, T. L., Hsu, E. B., Kim, H. K., Colli, S., Arana, G., & Green, G. B. (2005). The Incident Command System in Disasters: Evaluation Methods for a Hospital-based Exercise. *Prehospital and Disaster Medicine, 20*(1), 14–23. https://doi.org/10.1017/S1049023X00002090

Tierney, K. J., Lindell, M. K., & Perry, R. W. (2001). *Facing the Unexpected: Disaster Preparedness and Response in the United States*. Joseph Henry Press. https://doi.org/10.17226/9834

Uhr, C. (2009). *Multi-organizational Emergency Response Management – A Framework for Further Development for Further Development* [Lund University]. https://lup.lub.lu.se/search/publication/65b442e6-55b1-4f02-a42b-008da2910332

Uhr, C., Johansson, H., & Fredholm, L. (2008). Analysing Emergency Response Systems. *Journal of Contingencies and Crisis Management, 16*(2), 80–90. https://doi.org/10.1111/j.1468-5973.2008.00536.x

UN Secretary-General (2016). *Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction* (Vol. 21184, Issue December). https://www.undrr.org/publication/report-open-ended-intergovernmental-expert-working-group-indicators-and-terminology

United Nations (2010). *Hyogo Framework for Action 2005–2015: Building the Resilience of Nations and Communities to Disasters*. https://www.undrr.org/publication/hyogo-framework-action-2005-2015-building-resilience-nations-and-communities-disasters

United Nations (2015). *Sendai Framework for Disaster Risk Reduction 2015-2030*. https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030

United Nations International Strategy for Disaster Reduction (UNISDR) (2009). *2009 UNISDR Terminology on Disaster Risk Reduction*. https://www.unisdr.org/files/7817_UNISDRTerminologyEnglish.pdf

United Nations Office for Disaster Risk Reduction (UNDRR) (2021). *UNDRR Terminology - Disaster Risk Management (DRM).* https://www.undrr.org/terminology/disaster-risk-management#:~:text=Disaster risk management is the,and reduction of disaster losses.

United States Department of Homeland Security (2021). *COURSE: IS-870 - Dams Sector: Crisis Management Overview Course Lesson 4: Crisis Management Programs: Exercises*. https://emilms.fema.gov/IS870/DCM0104summary.htm#:~:text=HSEEP exercise types can be,are the operations-based exercises.

United States Departement of Homeland Security (2020). *Homeland Security Exercise and Evaluation Program (HSEEP)* (Issue January). https://www.fema.gov/media-library-data/1582669862650-94efb02c8373e28cadf57413ef293ac6/Homeland-Security-Exercise-and-Evaluation-Program-Doctrine-2020-Revision-2-2-25.pdf

United States Government (2021). *Ready.gov*. https://www.ready.gov/exercises

Vaishnavi, V., Kuechler, B., & Petter, S. (Eds.) (2004). "*Design Science Research in Information Systems*" January 20, 2004 (created in 2004 and updated until 2015 by Vaishnavi, V. and Kuechler, W.); last updated (by Vaishnavi, V. and Petter, S.), June 30, 2019. http://www.desrist.org/design-research-in-information-systems/

Van Aken, J. E. (2004). Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. *Journal of Management Studies, 41*(2), 219–246. https://doi.org/10.1111/j.1467-6486.2004.00430.x

Van Aken, J. E. (2005). Management Research as a Design Science: Articulating the Research Products of Mode 2 Knowledge Production in Management. *British Journal of Management, 16*(1), 19–36. https://doi.org/10.1111/j.1467-8551.2005.00437.x

Van Aken, J. E., & Romme, A. G. L. (2012). A Design Science Approach to Evidence-Based Management. *The Oxford Handbook of Evidence-Based Management, May 2014*. https://doi.org/10.1093/oxfordhb/9780199763986.013.0003

Van Asselt, M. B. A., & Renn, O. (2011). Risk governance. *Journal of Risk Research, 14*(4), 431–449. https://doi.org/10.1080/13669877.2011.553730

Van Duin, M., & Wijkhuijs, V. (2015). *De flexibiliteit van GRIP*. https://www.ifv.nl/advieseninnovatie/Documents/201504-IFV-De-flexibiliteit-van-GRIP.pdf

Van Eck, N. J., & Waltman, L. (2014). Visualizing Bibliometric Networks. In *Measuring Scholarly Impact* (pp. 285–320). Springer International Publishing. https://doi.org/10.1007/978-3-319-10377-8_13

Van Hoozer, M. (2008). *Moments: Making your life count for what matters most*. Insight Publishing.

Vedung, E. (1997). *Public Policy and Program Evaluation*. Taylor and Francis A.S.

Vedung, E. (2010). Four Waves of Evaluation Diffusion. *Evaluation, 16*(3), 263–277. https://doi.org/10.1177/1356389010372452

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Verheul, M. L. M. I., Dückers, M. L. A., Visser, B. B., Beerens, R. J. J., & Bierens, J. J. L. M. (2018). Disaster-exercises to prepare hospitals for Mass Casualty Incidents . Does it contribute to preparedness or is it ritualism ? *Prehospital and Disaster Medicine, 33*(4), 287–393. https://doi.org/10.1017/S1049023X18000584

Vlek, C. A. J. (1996). A multi-level, multi-stage and multi-attribute perspective on risk assessment, decision-making and risk control. *Risk Decision and Policy, 1*(1), 9–31.

Weber, R. (1990). *Basic Content Analysis* (2nd ed.). SAGE Publications Inc. https://doi.org/10.4135/9781412983488

Wikipedia (2021a). *El Al Flight 1862*. https://en.wikipedia.org/wiki/El_Al_Flight_1862

Wikipedia (2021b). *Enschede fireworks disaster*. https://en.wikipedia.org/wiki/Enschede_fireworks_disaster

Wikipedia (2021c). *Volendam New Year's fire*. https://en.wikipedia.org/wiki/Volendam_New_Year%27s_fire

Wildavsky, A. (1988). *Searching for Safety*. Transaction.

Wybo, J.-L. (2008). The Role of Simulation Exercises in the Assessment of Robustness and Resilience of Private or Public Organizations. In H. J. Pasman & I. A. Kirillov (Eds.),

Resilience of Cities to Terrorist and Other Threats (pp. 491–507). https://doi.org/10.1007/978-1-4020-8489-8_23

Yin, R. K. (2003). *Case study research: design and methods*. SAGE Publications Inc.

Annex A: Swedish and Dutch translations of the summary

Sammanfattning

Det är svårt att förutsäga framtida olyckor och kriser. Men, det är möjligt att lära från tidigare inträffade händelser och därigenom hela tiden förbättra förmågan att hantera dem. Att utvärdera hur man har hanterat en olycka eller kris är därför en mycket viktig aktivitet för att exempelvis räddningstjänsten, sjukvården och polisen hela tiden skall kunna förbättra sin förmåga att hantera olika typer av händelser.

Utvärdering kan inte bara ske efter olyckor och kriser utan också efter övningar där man kan simulera olika typer av händelseförlopp. Oavsett om man utvärderar en inträffad olycka, kris eller en övning är det centralt att utvärderingsprocessen och det slutliga resultatet, vilket ofta sammanfattas i någon typ av rapport, upplevs som användbart. Tyvärr förefaller detta ofta inte vara fallet och i värsta fall anses utvärderingen och dess slutsatser vara en onödig byråkratisk procedur som inte tillför mycket värde. För att öka värdet av utvärderingar är det därför nödvändigt att ta reda på vad det är som gör dem användbara för de tänkta användarna. Därför är också den övergripande frågan som den aktuella avhandlingen försöker besvara – hur kan utvärderingar av olycks- och krishantering förbättras?

Första delen av avhandlingsarbetet fokuserade på att kartlägga kunskapsfronten gällande utvärdering av olycks- och krishanteringsövningar, både när det gäller forskningslitteratur och praktik. Andra delen handlade om att finna sätt att förbättra utvärderingarnas användbarhet. Resultaten från den första delen visar att trots ett ökat vetenskapligt intresse för området under senare år finns det förhållandevis få bidrag inom området. Dessutom förefaller dessa bidrag i låg utsträckning bygga vidare på varandra, vilket delvis kan förklara varför den vetenskapliga kunskapsbasen är splittrad. Bristen på transparens och spårbarhet när det gäller de metoder som föreslås för utvärdering gör att professionella inte har tillförlitliga och validerade procedurer för att designa, genomföra och utvärdera olycks- och krishanteringsövningar. Den övergripande slutsatsen rörande den första delen av avhandlingsarbetet är därför att det

är svårt att veta huruvida de utvärderingar som genomförs i praktiken leder till att de mål och syften som finns med arbetet faktiskt uppnås. Därmed är det också oklart hur utvärderingarna bidrar till det övergripande arbetet med olycks- och krishantering.

Avhandlingens andra del bygger vidare på slutsatserna från den första delen och innebär ett fokus på att försöka klarlägga vilka faktorer, relaterat till utvärdering av olycks- och krishanteringsövningar, som påverkar hur användbara dessa blir i det övergripande krisberedskapsarbetet. Arbetet bestod delvis i att lägga en konceptuell grund för en analys av utvärderingar, vilken innebär att kopplingen mellan utvärdering och andra viktiga aktiviteter såsom exempelvis riskanalys och förmågeanalys blir tydligare. Och, det inkluderade också empiriska studier av både kvalitativ och kvantitativ karaktär. Exempelvis genomfördes experiment där professionella som arbetar med olycks- och krishantering fick svara på frågor om hur de uppfattar olika typer av utvärderingsbeskrivningar. Beskrivningarna var baserade på verkliga utvärderingsrapporter. Resultaten pekar bland annat på att vissa delar av en utvärdering av en olycks- och krishanteringsövning är viktigare än andra för att den skall vara användbar. Exempelvis bör analysdelen vara tydlig och kopplad till slutsatserna från utvärderingen för att utvärderingen skall vara användbar. I korthet innebär det att man måste kunna härleda slutsatserna från analysens olika delar. Detta är dock inte alltid fallet i praktiken och resultaten från avhandlingen tydliggör vikten av systematiska, evidensbaserade, och transparenta utvärderingar.

Avhandlingen innehåller också förslag på hur utvärderingar av olycks- och krishanteringsövningar kan anpassas för att göra dem mer användbara. Sådana förbättringar bör leda till bättre förutsättningar för att förbereda olika organisationer inför olyckor och kriser. Och, slutligen, tydliggör resultaten från avhandlingen att utvärderingar inte bör ses som ett mål i sig, utan som ett medel för att långsiktigt kunna förbättra förmågan att hantera olyckor- och kriser.

Samenvatting

Hoewel de toekomst moeilijk te voorspellen is, is het mogelijk en noodzakelijk om van het verleden te leren. Recente rampen en crises laten zien dat crisisprofessionals hun vaardigheden, procedures en systemen voortdurend moeten reviewen, onderhouden, aanpassen en (door)ontwikkelen. Hierdoor zijn ze beter geprepareerd, paraat en in staat om effectief en efficiënt te reageren op toekomstige gebeurtenissen.

Evaluatie is een instrument dat dit cyclische proces ondersteunt. Het kan bijdragen aan het geven van antwoorden op vragen van belanghebbenden en crisisprofessionals, en hun organisaties helpen om hun preparatie en respons inzichtelijk te maken, te beoordelen, te ontwikkelen of zelfs te verbeteren. Ervaringen die opgedaan zijn in zowel gesimuleerde gebeurtenissen (oefeningen) als tijdens daadwerkelijke gebeurtenissen (rampen of crises) kunnen worden gebruikt om toekomstige activiteiten te verbeteren. Belangrijke randvoorwaarden hiervoor zijn dat het evaluatieproduct en -proces door de eindgebruiker als nuttig en/of bruikbaar worden ervaren. Helaas lijkt het erop dat dit niet altijd het geval is. Evaluaties en hun producten worden zelden volledig gebruikt of, in extreme gevallen, slechts gezien als een symbolische exercitie zonder impact. Om hier verandering in te brengen is het van cruciaal belang om te identificeren wat eindgebruikers als belangrijk of nuttig beschouwen. Daarom is de vraag die aan dit onderzoek ten grondslag ligt hoe evaluaties (of hun perceptie) kunnen worden verbeterd. Het antwoord helpt crisisprofessionals om hun preparatie, paraatheid en respons te verbeteren door (beter) gebruik te maken van evaluaties en de inzichten die hieruit voortkomen.

Dit onderzoek is gebaseerd op voortschrijdend inzicht: bevindingen uit eerdere (deel-) onderzoeken zijn gebruikt om nieuwe onderzoeksvragen te formuleren en aan te passen. Hierbij is, vanwege consistentie en vergelijkbaarheid, het Nederlandse systeem van crisismanagement als case-study gebruikt. Het eerste deel van dit onderzoek richt zich op het in kaart brengen van de huidige staat van evalueren van crisisoefeningen, zowel in theorie als in de praktijk. Het tweede deel geeft inzicht in manieren om het nut en de bruikbaarheid van evaluaties te vergroten. De uiteindelijke onderzoeksstrategie combineert zorgvuldig geselecteerde kwantitatieve (survey-experimenten) en kwalitatieve (documentanalyses en expert judgement) methoden in een mixed-methods-aanpak. De bevindingen uit dit onderzoek zijn nuttig voor individuen, teams, organisaties of systemen (crisisbeheersingsprofessionals of hulpverleners) die beter voorbereid willen zijn op toekomstige rampen.

De eerste resultaten laten zien dat, ondanks de toegenomen wetenschappelijke aandacht, weinig wetenschappelijke onderzoeken zich focussen op het evalueren van

crisisoefeningen. De literatuur beperkt zich vaak tot één specifieke gebeurtenis en/of evaluatiesoort. Zowel onderzoek als praktijk hebben bovendien de neiging om evaluaties als op zichzelf staande gevallen te bekijken en bijvoorbeeld geen lessen te trekken uit eerdere evaluaties. Hierdoor ontstaat een onsamenhangend en gefragmenteerd beeld dat tevens diepgang mist. Het gebrek aan wetenschappelijke onderbouwing betekent ook dat professionals geen betrouwbare, valide basis hebben voor het ontwerpen van (oefen)evaluaties. Bovendien wordt in de meeste documentatie de gekozen (evaluatie-) methode niet nader toegelicht of verantwoord, waardoor het niet duidelijk is waarom een bepaalde aanpak gekozen is en hoe deze bijdraagt aan het bereiken van het gewenste doel. Ook de context wordt hierbij onvoldoende meegenomen. Verder is er een gebrek aan transparantie over de manier waarop de verzamelde gegevens worden geanalyseerd en gebruikt worden om conclusies te trekken en aanbevelingen te doen. In deze eerste fase van het onderzoek is dan ook gebleken dat het moeilijk is om te weten of de huidige evaluaties effectief en/of nuttig zijn en hoe ze uiteindelijk bijdragen aan een (betere) voorbereiding op rampen en crises.

Het tweede deel van het onderzoek bouwt voort op de resultaten uit het eerste deel zoals hierboven beschreven. Crisisprofessionals zijn gevraagd om gemanipuleerde evaluaties uit de praktijk te beoordelen met als doel om te onderzoeken welke aspecten van evaluaties van invloed zijn op hun nut voor de gebruiker. Om dit mogelijk te maken is het concept van evaluatiebeschrijving geïntroduceerd. Dit concept bevat vier componenten die zeer waarschijnlijk belangrijk zijn in een evaluatie: het doel, de objectbeschrijving, de analyse en de conclusie. De resultaten geven aan dat de manier waarop de analyse en/of conclusies worden gedocumenteerd, van invloed is op de bruikbaarheid van evaluaties. Ook is duidelijk dat de verschillende componenten in meer of mindere mate van invloed zijn en dat dit afhankelijk is van het doel dat de evaluatie tracht te bereiken (leren of verantwoorden). Crisisprofessionals benadrukken daarnaast ook dat een grondige analyse in een evaluatie verder moet gaan dan slechts het doel van de evaluatie en rekening moet houden met de context ervan. Bovendien zijn zij van mening dat evaluaties evidence-based, actiegerichte conclusies moet opleveren.

De resultaten van dit onderzoek onderstrepen het belang van systematische, valide, betrouwbare en evidence-based evaluaties. Er zijn verschillende problemen geïdentificeerd en mogelijke oplossingen aangedragen voor de wijze waarop een evaluatie verbeterd kan worden. Hierdoor zorgt dit onderzoek ervoor dat de evaluatie beter aansluit bij de wensen en verwachtingen van de gebruiker, waardoor toepasbaarheid wordt vergroot. Dit heeft op zijn beurt waarschijnlijk een positieve invloed op de preparatie en respons, waardoor we beter in staat zijn om toekomstige rampen en crises te managen. Tenslotte laat het onderzoek zien dat het niet de evaluatie

zélf is die tot verbetering leidt, maar dat het gaat om het gebruik ervan. Evaluatie moet daarom niet gezien worden als een doel op zich, maar als een middel om een doel te bereiken.

Annex B: Abstracts of related publications

Disaster exercises to prepare hospitals for Mass-Casualty Incidents: Does it contribute to preparedness or is it ritualism?

Verheul, M.L.M.I., Dückers, M.L.A., Visser, B.B., **Beerens, R.J.J.**, & Bierens, J.J.L.M. (2018)

Paper published in: Prehospital and Disaster Medicine 33 (4), 387–393.

Online available: https://doi.org/10.1017/S1049023X1

Abstract

The central question this study sought to answer was whether the team members of Strategic Crisis Teams (SCTs) participating in mass-casualty incident (MCI) exercises in the Netherlands learn from their participation.

Methods: Evaluation reports of exercises that took place at two different times were collected and analysed against a theoretical model with several dimensions, looking at both the quality of the evaluation methodology (three criteria: objectives described, link between objective and items for improvement, and data-collection method) and the learning effect of the exercise (one criterion: the change in number of items for improvement).

Results: Of all 32 evaluation reports, 81% described exercise objectives; 30% of the items for improvement in the reports were linked to these objectives, and 22% of the 32 evaluation reports used a structured template to describe the items for improvement. In six evaluation categories, the number of items for improvement increased between the first (T1) and the last (T2) evaluation report submitted by hospitals. The number of items remained equal for two evaluation categories and decreased in six evaluation categories.

Conclusion: The evaluation reports do not support the ideal-typical disaster exercise process. The authors could not establish that team members participating in MCI exercises in the Netherlands learn from their participation. More time and effort must

be spent on the development of a validated evaluation system for these simulations, and more research into the role of the evaluator is needed.

Flood Preparedness Training and Exercises

Beerens, R.J.J., Abraham, P.J., Glerum, P. And Kolen, B. (2014)

Chapter published in: Bierens, J.L.M. (Ed.) Handbook on Drowning (2nd edition). Berlin: Springer-Verlag.

Online available: https://doi.org/10.1007/978-3-642-04253-9_154

Abstract

Preparedness is the key to provide an effective response to drowning, whether it is a single person at risk of drowning through unplanned entry into water or a larger number through a disaster such as flooding. Preparedness is achieved by training and exercise. There is increased awareness that water search and rescue response needs to become more effective as worldwide flood risks grow. Teams identified as water incident responders need training and exercises to acquire and consolidate the knowledge, skills and behaviour that are needed to deliver a safe, effective response. This should be a cyclic and holistic process that allows to approach to identify gaps and barriers that reduce the effectiveness. Training and exercises are not independent activities, and they should be part of a larger process of disaster management preparedness that has a specific risk-based context. This chapter describes how training and exercise may help water search and rescue teams to be prepared for a response to a flood disaster.

Maximise your returns in Crisis Management preparedness: A Cyclic Approach to training and exercises.

Beerens, R.J.J., Abraham P. and Braakhekke, E. (2012)

Conference proceeding prepared for: International Disaster Risk Conference (IDRC) in Davos.

Online available: https://doi.org/10.13140/2.1.1138.0482

Abstract

Training and exercises programmes are not independent activities, forming part of a larger, risk-based, process of disaster management preparedness. In order to have an impact on an individual's skills knowledge or behaviours, or to influence organizational

learning or procedures, the programmes needs a cyclic and holistic approach. It should focus on clearly identified outcomes that are designed to meet the demands of identified gaps and emerging threats. This will support meaningful evaluations against clear indicators. Without having clear outcomes, standards or values, it is not possible to evaluate the effectiveness of a programme. These outcomes form measurable performance indicators around which a detailed programme can be designed. Following delivery, analysis of the evaluation observations will identify critical gaps in knowledge, skills, behaviour or policy. This analysis allows clear, structured recommendations to be formulated that provide guidance as to the content of the continuing training programme cycle, prioritising key needs and ensuring maximum efficiency and utilisation of resources at all levels. By analysing and comparing six European Modules exercises (EU ModEx 2010-2011) and their outcomes we can demonstrate the benefits of this approach. We end this paper with recommendations that would potentially increase the learning outcomes in any future training or exercise programme.

How prepared is prepared enough?

Jongejan, R.B., Helsloot, I., **Beerens, R.J.J.** en Vrijling, J.K. (2011)

Paper published in: Disasters, the Journal of Disaster Studies, Policy and Management.

Online available: https://doi.org/10.1111/j.1467-7717.2010.01196.x

Abstract

Decisions about disaster preparedness are rarely informed by cost-benefit analyses. This paper presents an economic model to address the thorny question, 'how prepared is prepared enough?' Difficulties related to the use of cost-benefit analysis in the field of disaster management concern the tension between the large number of high-probability events that can be handled by a single emergency response unit and the small number of low-probability events that must be handled by a large number of them. A further special feature of disaster management concerns the opportunity for cooperation between different emergency response units. To account for these issues, we introduce a portfolio approach. Our analysis shows that it would be useful to define disaster preparedness not in terms of capacities, but in terms of the frequency with which response capacity is expected to fall short.

EU FloodEx 2009: An analysis of testing international assistance during a worst credible flood scenario in the North Sea Area

Beerens, R.J.J., Kolen, B., and Helsloot, I. (2010)

Conference proceeding prepared for: FRIAR 2010. United Kingdom: University of Wessex.

Online available: https://doi.org/10.2495/FRIAR100211

Abstract

This paper discusses a case study example of testing international disaster response assistance within the European Union during a worst credible flood scenario in the North Sea area. It describes and evaluates the processes of requesting and receiving international assistance and the field operations with responding international teams during an exercise for large scale flooding ('EU FloodEx 2009'). It also discusses some of the issues identified during this exercise in the Netherlands. Additionally the characteristics of an (inter)national response in case of flooding are related to various processes and the effectiveness after initiating them. For initiating and planning of these processes, the results of the exercise are reflected to availability of information during a threat or flood with regards to warning, decision making and response in case of uncertainty. The paper also introduces the structures, mechanisms and teams at the disposal of the Dutch and EU flood response community. It ends by discussing some experiences of 'EU FloodEx 2009' to improve the design of the EU response system and future exercises by implementing the lessons identified.

Annex C: Components often found in (crisis management) evaluation frameworks

COMPONENT	RELATED ASPECTS IN THE LITERATURE
Purpose (P)	*Why do we perform an exercise/evaluation?*
	Sinclair et al. (2012): the objectives of the exercise.
	Duarte et al. (2013) an element of the process where the assessment objectives are identified.
	Savoia et al. (2014): the purpose of the measurement, i.e. accountability or quality improvement.
	Stufflebeam & Coryn (2014) distinguish four main uses: improvement, accountability, dissemination and enlightenment.
Users (U) or Stakeholders (S)	*How will the evaluation be used, and by whom?* The intended audience.
Object (O) (Artefact, Evaluee, Evaluand or Intervention)	*What is being evaluated?* The object of the evaluation. This can be a person, a program, project, policy, proposal, product, equipment, services, concepts and theories, data and other types of information, individuals, or organisations. It is described by identifying the specific aspects or conditions that are being evaluated (e.g. objectives: elements that should be achieved by the object).
	Heath (1998): Interdependent components of an evaluation can be analysed by looking at the structures, systems, processes and people involved.
	Alexander (2015): the (stated) goals of the (civil protection) system.
Context (C) & Scenario (S) A factual account of events or descriptive information	*What is the context (or environment) in which the object (O) is operating and what was the scenario?*
	Abrahamsson et al. (2010): the conditions for the evaluation in terms of describing the events that led to the initiation of the response, the preconditions under which the emergency response system operated and establishing the objectives of the emergency response operation at the highest system level (i.e. for the total emergency response system).
	Heath (1998): the crisis environment, which includes the physical space and any structures or processes involved in (or contributing to) the crisis incident.
	Wybo et al. (2008): structures (what is prescribed by the organisation, objectivable and measurable) and relations (the roles played by the different actors).
	Heath (1998): the crisis incident covers the precipitating event – the earthquake, storm, explosion, or 'thing gone wrong' that triggers the crisis situation.
	Savoia et al. (2014): the type of exercise, quality of the exercise (scenario, participants, etc.).

COMPONENT	RELATED ASPECTS IN THE LITERATURE
Measures (M) or evidence (indicators, instruments or evaluation criteria)	*What are the (minimum) requirements or objectives for the object under evaluation?* A measure of success or a normative judgement.
	Heath (1998): for structures, systems and processes three primary assessment tasks are design and construction, safety and security, and function and output or outcomes. For people, assessment tasks are selection and training, skills and readiness, and behaviour and actions.
	Savoia et al. (2014): the instrument is defined as the combination of performance measures used during a given exercise. Multiple types of measures need to be included (checklists, scores, open ended questions) when using exercises to measure performance
	Wybo et al. (2008): Action: decisions made and actions carried out.
Evidence/ data collection method (M)	*How is data or evidence gathered or collected?* Evaluators/observers. Evaluator: the person responsible for conducting the evaluation. Information related to the issues in question. Quantitative or qualitative.
Analysis (A), Reasoning (R) or Justification	*What happened during the exercise and why?* Answering such a question requires the collection of evidence with regard to the performance of the object in the context and scenario that served as the basis for the exercise. Thus, the analysis should help to understand why the outcome of the exercise was what it was. An explanation or justification that logically demonstrates that the evidence collected and analysed leads to evaluation outcomes.
	Abrahamsson et al. (2010): generating an understanding of how the emergency response system performed, and why the outcome was what it was.
	Wybo et al. (2008): analysis is how people perceived (on the spot) the situation and its evolution, and the hypotheses that were considered.
	Savoia et al. (2014): during an exercise data are gathered by the observation of individuals' key actions and decisions forming the bulk of the source of measurement. As a result, there is frequently a discrepancy between the level of data collection (most often individual) and the level of data analysis (organisational or system) when the focus of the evaluation is at the public health system level. As a result of such observations a third concept was identified in the development of the conceptual model: the unit.
Judgement (J) or Conclusion (C)	*How (well) did the object of the evaluation perform?* Good and bad features. Evaluation is the process of determining the merit, worth and value of things.
	Heath (1998): finding out what happened is not the same as judging. Biases in memory and time distortion can colour evidential recall.
	Scriven (1993): evaluations are not value-free. This is also related to evaluation's root term 'value', which denotes that evaluations essentially involve making value judgements. Values may include: effectiveness, efficiency, usability, cost, safety, legality
	Abrahamsson et al. (2010): whether the performance of the emergency response system during the emergency (or exercise) was acceptable or how it could be improved.
	Stufflebeam and Coryn (2014): intrinsic vs extrensic value. Many evaluations carry a need to draw a definitive conclusion or make a definite decision on quality, safety or some other variable.
Impact (I), Application (A) or use (U)	Applying the outcomes of the evaluation in order to achieve its purpose. How evaluative information is used, by whom, for what purposes, and how legitimate use can be increased.

Annex D: Overview of research contributions

The table provides a concise, but complete overview of the individual research contributions.

PAPER	RESEARCH QUESTION(S) AND AIM	RESEARCH METHODS AND EMPIRICAL DATA	CONTRIBUTIONS
I	RQ 1: What is known about the evaluation of disaster management exercises in scientific literature? To provide a comprehensive overview of the literature regarding the evaluation of disaster management exercises.	Scoping study and in-depth analyses 246 articles (overall analysis) 43 articles (in-depth analysis)	There appears to be a lack of scientific literature that focuses on the evaluation of disaster management exercises *per se*, or within disaster management in general. Little is known about the evaluation of disaster management exercises.Most interest in the topic is in the medical domain, followed by the social sciences and engineering. Scientific contributions are descriptive and focus on single/ stand- alone cases from various perspectives. There appears to be a lack of interest in building a solid knowledge base. There are signs of the creation of a more cohesive corpus on the evaluation of disaster exercises as it was possible to identify four research clusters with distinct foci. Evaluative approaches dominate and are applied using a variety of data collection methods. It is unclear how specific evaluation methods are linked to their overall purpose in order to determine usefulness and/ or effectiveness.
II	RQ 2: How are disaster management exercise and real-life response evaluations documented in the Netherlands? To increase knowledge relating to how emergency, disaster or crisis evaluations are performed and reported in practice, and whether they actually meet their intended purpose.	Content analysis Expert judgement group session 62 documents (18 exercise evaluations, 23 systemic test evaluations, 21 real emergency response evaluations) 55 participants	The Dutch disaster and crisis management system is decentralised and managed locally by the 25 Safety Regions The evaluation of incidents and exercises are seen as independent activities, and are documented as such. They do not naturally build upon each other, with implications for national learning and the ongoing development of crisis management. Unlike systemic tests, there is no common framework that would ensure that the intended purposes are achieved and findings can be shared (or meta-analyses can be performed to identify systemic issues). There are a variety of data-collection methods and evaluative approaches. However they lack transparency and the underlying reasoning and/ or justification is unclear. A holistic knowledge base, on a national level, is lacking. Evaluation design and the relationship with its purpose and usefulness for stakeholders requires further attention. Group sessions highlighted that it remains unclear how evaluations achieve their purpose, and contribute to being better-prepared for the next crisis or disaster.

PAPER	RESEARCH QUESTION(S) AND AIM	RESEARCH METHODS AND EMPIRICAL DATA	CONTRIBUTIONS
	RQ 3:What makes evaluations (descriptions/texts) more or less useful to professionals?		
III	RQ 3a: How does the clarity of the presentation of the object (O), the analysis (A) and/ or the conclusion (C) in an evaluation description influence its perceived usefulness for the purposes of: (i) learning; and (ii) accountability? The usefulness of evaluations for the most common DRM purposes (learning and accountability) is investigated by manipulating some key components.	Survey experiments 84 participants (50 'operational' and 34 'governing')	This paper introduced the notion of the evaluation description, which encompasses four components that are assumed to influence the usefulness of an evaluation: Purpose (P), Object description (O), Analysis (A) and Conclusions (C). Based on a sample of mayors and crisis management professionals (N = 84) working in the Netherlands, the results showed that how evaluations of emergency exercises are documented influences their usefulness. The clarity of conclusions has a significant effect on perceived usefulness. The clarity of the analysis has a significant effect on perceived usefulness for learning purposes, with a marginally significant effect for accountability purposes. The clarity of the object description did not have a significant effect on usefulness.
IV	RQ 3b: What do crisis management professionals expect to find in a useful crisis management evaluation report? To gain a better understanding of what users expect to find in a useful evaluation report.	Thematic analysis of qualitative survey data. 84 participants in key roles	An analysis of individual expectations of crisis management professionals in the Netherlands regarding useful disaster (exercise) evaluations identified five main themes: (I) information regarding why the evaluation is needed or what it is used for (purpose); (II) information about what, or who, is being evaluated (object); (III) information that is needed to reach conclusions, or details of what happened, how and why (analysis); (IV) information that details the outcome, or how well the object of the evaluation performed (conclusions); and (V) details on how data should be presented (design). Despite these common main themes, the findings highlighted that it is difficult to create one evaluation that meets all expectations. Most respondents expect evaluations to support learning and improvement by providing actionable feedback on what should be differently or better. In general, evaluations should avoid blaming, finger-pointing or scapegoating as this is perceived to have a negative impact. Respondents refer to 'objects' at various levels (i.e. the system, the organisation, the team or the individual). Each object might require a different evaluation approach. Respondents noted that the context in which the object is evaluated should be taken into account. Respondents expect analyses/ evaluations to be rigorous, go beyond the individual case/ context, and take into account a broader system perspective. They should also take into account the fact that evaluations are performed *ex-post*. The (meta)analysis should distinguish between case-specific (one time only) and continuous systemic (recurring) failures and provide directions for change.

Annex E: Appended papers

Paper I: Beerens, R.J.J., & Tehler, H., (2016). Scoping the field of disaster exercise evaluation - A literature overview and analysis. *International Journal of Disaster Risk Reduction,* 19, 413–446.

Online available: https://doi.org/10.1016/j.ijdrr.2016.09.001

Paper II: Beerens, R.J.J. (2019). Does the means achieve an end? A document analysis providing an overview of emergency and crisis management evaluation practice in the Netherlands. *International Journal of Emergency Management,* 15 (3), 221–254.

Online available: https://doi.org/10.1504/IJEM.2019.102310

Paper III: Beerens, R.J.J., Tehler, H., & Pelzer, B. (2020). How can we make crisis management evaluations more useful? An empirical study of Dutch evaluation descriptions. *International Journal of Disaster Risk Science,* 11, 578–591.

Online available: https://doi.org/10.1007/s13753-020-00286-7

Paper IV: Beerens, R.J.J., & Haverhoek-Mieremet, K. (2021). What do practitioners expect from an evaluation report? A qualitative analysis of Dutch crisis management professionals' expectations. *International Journal of Emergency Services,* 10 (1), 1–25.

Online available: https://doi.org/10.1108/IJES-12-2019-0063

Printed by Media-Tryck, Lund 2021 NORDIC SWAN ECOLABEL 3041 0903

Lund University
Faculty of Engineering
Division of Risk Management and Societal Safety

ISBN 978-91-7895-923-5