



How Can We Make Disaster Management Evaluations More Useful? An Empirical Study of Dutch Exercise Evaluations

Ralf Josef Johanna Beerens^{1,2} · Henrik Tehler^{2,3} · Ben Pelzer⁴

Published online: 10 July 2020
© The Author(s) 2020

Abstract The evaluation of simulated disasters (for example, exercises) and real responses are important activities. However, little attention has been paid to how reports documenting such events should be written. A key issue is how to make them as useful as possible to professionals working in disaster risk management. Here, we focus on three aspects of a written evaluation: how the object of the evaluation is described, how the analysis is described, and how the conclusions are described. This empirical experiment, based on real evaluation documents, asked 84 Dutch mayors and crisis management professionals to evaluate the perceived usefulness of the three aspects noted above. The results showed that how evaluations are written does matter. Specifically, the usefulness of an evaluation intended for learning purposes is improved when its analysis and conclusions are clearer. In contrast, evaluations used for accountability purposes are only improved by the clarity of the conclusion. These findings have implications for the way disaster management evaluations should be documented.

Keywords Disaster management evaluation · Evaluation design · Evaluation report · Exercise evaluation · The Netherlands · Usefulness

1 Introduction

Simulated disaster events (for example, exercises) are an important element of disaster risk management (DRM) activities (UNDRR 2020) as they enable the training or testing of people, organizations, and plans in a safe, but realistic environment (Peterson and Perry 1999; Skryabina et al. 2017). Systematic evaluations of both exercises and real responses are important, as they can support the direction of, and investment in, future learning and development, or provide insights into the efficiency and performance of current practice (Borodzicz and Van Haperen 2002; Boin et al. 2017). From a broader perspective, DRM exercises and their evaluation have two overarching purposes: to enhance learning and/or ensure accountability. In both cases, evaluations are a key way to link a specific exercise to broader DRM processes—for example, lessons or insights can be distilled and shared with other organizations.

In practice, many methods and guidelines have been developed to support exercise evaluations (UK Cabinet Office 1998; Bradshaw and Bartenfeld 2009; Swedish Civil Contingencies Agency 2011; Heumüller et al. 2012; ISO 2013; U.S. Department of Homeland Security 2020). However, critics have noted a reliance upon anecdotal evidence, and a lack of systematic studies of effectiveness and/or other forms of validation (Thomas et al. 2005). There is clearly a diversity of opinions regarding their usefulness, and the same holds for related documentation, which is often published in some form of report. To the

✉ Ralf Josef Johanna Beerens
ralf.beerens@ifv.nl

¹ Institute for Safety (IFV), 6801 HA Arnhem, The Netherlands

² Division of Risk Management and Societal Safety, Lund University, 221 00 Lund, Sweden

³ Lund University Centre for Risk Assessment and Management (LUCRAM), Lund University, 221 00 Lund, Sweden

⁴ Department of Sociology, Radboud University Nijmegen, 6500 HE Nijmegen, The Netherlands

best of our knowledge, very few methods or guidelines have been tested to determine the extent to which these reports do actually, for example, improve learning or ensure accountability (Beerens and Tehler 2016; Beerens 2019). More specifically, there seems to be a lack of empirical evidence regarding the usefulness of different ways of conducting and presenting evaluations.

As it is clearly difficult to assess such evaluations using real-life exercises, the current study adopts an experimental approach; we seek to shed light on the question of how to best present the results of an evaluation to support learning and accountability. Our analysis focuses on a set of fictional evaluation descriptions, which draw upon actual, multi-organizational exercise or incident evaluations prepared by Dutch Safety Regions. These descriptions were manipulated in order to investigate whether the manipulation resulted in a more (or less) useful evaluation. Judgements of usefulness were made by 84 professionals working in DRM in the Netherlands.

2 Background: Evaluation of Emergency, Disaster, and Crisis Management Exercises

In broad terms, an evaluation seeks to determine the merit, worth, or value of something (Lincoln and Guba 1980). However, different disciplines use slightly different approaches and methods. Examples include policy evaluation (Bovens et al. 2008; Hertting and Vedung 2012), program evaluation (Frye and Hemmer 2012), and product evaluation (Etkin and Sela 2016).

Our point of departure is the operational disaster response evaluation (Beerens 2019). In the present study, “evaluation” is defined as: (1) a systematic assessment of the value or performance of an operational actor with respect to the intended and actual outcome(s) in a given scenario; or (2) the product (for example, report) of that assessment. Thus, the definition covers both the process of conducting an evaluation, and the product of that process. Here, we focus on the latter, which is often presented in the form of an evaluation report, or a “lessons learned” document. These documents play an important role in recording and transferring information. In theory, this transfer can support individuals and their organizations to, for example, develop relevant meaning structures or interpretative modeling processes. We use the term “evaluation description” to refer to this documented outcome, as a parallel to the concept of the “risk description” in the field of risk assessment (Aven 2010).

2.1 Evaluation Descriptions

Disaster risk management exercise evaluation has received little attention in the scientific literature (Beerens and Tehler 2016). Very few articles propose a structure or framework to support the design of an evaluation process (Heath 1998; Wybo 2008; Abrahamsson et al. 2010; Heumüller et al. 2012; Sinclair et al. 2012; Duarte da Costa et al. 2013; Savoia et al. 2014, 2017) and there is, currently, no coherent body of knowledge to help professionals determine the best way to conduct an evaluation and present the output. We therefore need a conceptual language that mirrors the components found in risk descriptions, namely: events, consequences, uncertainty, and background information (Aven 2010). In the risk domain, a key research question is, for example, how best to describe potential consequences and uncertainty to optimize usefulness (Lin et al. 2015, 2017; Månsson et al. 2019), and similar questions arise in the context of an evaluation description.

Based on Scriven’s “logic of evaluation” (Scriven 1980), work by Calidoni-Lundberg (2006), Wybo (2008), and Beerens (2019), and the definition of the evaluation given above, we propose that there are at least four elements: Purpose (P), Object description (O), Analysis (A), and Conclusions (C). These components provide a logical, sequential foundation for understanding the why–what/who–where/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.

Although this P, O, A, C format is often found when documenting, transferring, and presenting information from (exercise) evaluations, previous research indicates that there is significant variation in how components are described and connected (Beerens 2019). For example, lacking a justification (P) for the evaluation, readers struggle to judge the credibility of the analysis and its conclusions. Similarly, a weak connection between the analysis and conclusions leaves the reader doubting the credibility of the document. Such fragmentation in the logical chain that extends from the purpose (P), via object descriptions (O), to the analysis (A) and conclusions (C) is detrimental to the usefulness of the document.

The present study investigates the influence of this P, O, A, C chain on the usefulness of an evaluation description; the next section briefly outlines each of the four components, and how it influences the overall chain.

2.1.1 Purpose (P): Why Was the Evaluation Conducted?

A first step in the design, execution, and evaluation of an exercise is to identify why it needs to be performed.

Skryabina et al. (2017) identify five functions: (1) identify or assess, for example, gaps in (capacity) planning or procedures; (2) enhance or improve, for example, visibility, awareness, understanding, and knowledge; (3) test, for example, plans/procedures, tools/equipment, or personnel; (4) validate, for example, training, education, or exercise programs; and (5) measure, for example, performance or confidence.

Here, we distinguish two, overarching purposes: learning (including development and improvement) and accountability (Beerens and Tehler 2016; Beerens 2019). Although an exercise can have many functions, it is possible to relate them to these two, overall purposes. For example, identified gaps (1) can be used for development if, for example, gaps in resources lead to the creation of new resources; they can, however, also be used to hold people accountable for the gap that was created. The same applies to, for example, testing (3); if a person fails a test, the evaluation can support learning by providing them with feedback to improve effectiveness and efficiency. The same process can also provide more normative feedback, and focus on holding an individual accountable for not achieving the required level of performance.

In general, learning provides feedback to participants, exercise designers, or emergency planners, while accountability provides feedback to decision makers, supervisors, or directors, among others. Overall, it is reasonable to assume that why an evaluation is performed, how it will be used, and by whom, are all important influences on its usefulness.

2.1.2 Object (O): What or Who Was Evaluated?

The next step is to, implicitly or explicitly, describe the object of the evaluation (O)—the evaluand or, if it is a person, the evaluatee. Scriven (1991) described the evaluand as “whatever is being evaluated.” This information is needed in order to formulate statements regarding the object’s worth, merit, or value. If the object’s purpose, and how it should achieve it are unclear, it will be impossible to determine whether it performed to the required standard. The context the object was operating in is also important, as this can influence its functioning and helps to determine how well it performed. For example, an emergency response system relies upon various independent services such as fire, police, and ambulance (Uhr et al. 2008). Here, it is important to specify whether the evaluation concerns the system or its independent services, as, without this, it is difficult for the reader to understand what the result applies to.

2.1.3 Analysis (A): What Happened During the Exercise and Why?

The analysis also influences the usefulness of an evaluation description. It addresses the question “what happened during the exercise and why?” and should help to understand why the outcome of the exercise was what it was (Wybo 2008; Abrahamsson et al. 2010). The answer requires collecting data about the performance of the object. For example, decisions made by senior officers can be documented, plans can be studied, and actions can be observed. Evidence can take the form of observations, questionnaires, (group) interviews and discussions, and document reviews (Beerens 2019). However, collection and reporting must be followed by interpretation. Thus, the analysis documents information sources, and transforms evidence into a description of what happened and, possibly, why. Particular attention should be paid to reasoning that precedes outcomes, and care must be taken to ensure that the aim is to find out what actions were undertaken, which should be distinguished from judging (Heath 1998).

2.1.4 Conclusion (C): How (Well) Did the Object of the Evaluation Perform?

How the conclusions of the evaluation are presented is the final aspect of an evaluation description considered here, and P, O, and A should support this. If, for example, the purpose of the exercise is to support learning, conclusions might involve value judgements regarding the performance of the object and, if deemed necessary, recommendations for how it could be improved. On the other hand, if the purpose relates to accountability, conclusions might involve judgements regarding the extent to which the object achieved a predefined performance level, and recommendations regarding how to achieve the required levels.

2.2 Usefulness of Evaluation Descriptions

Although these four aspects are logically connected, it remains difficult to establish their respective degree of influence, and whether one is perceived as more important than another. The relative importance given to each factor is likely to be a function of who is reading the report. The question addressed by this study is, therefore, whether the presentation of the four components (for example, how and what information is presented) influences the usefulness of a report. More specifically, we investigate whether three of the four components (O, A, C) influence usefulness with respect to the first component, purpose (P).

Clearly, leaving one of the components out completely is unlikely to be very useful, and such extreme cases have

little relevance in practice. However, we have reason to believe that the clarity of the description of each component, and the connections between them, might have an impact on the usefulness of the overall description (Beerens 2019). For example, is an unclear conclusion more detrimental to usefulness than an unclear analysis?

Our hypothesis was that variation in the clarity of the four components would influence the overall usefulness of an evaluation description. However, at a more detailed level, we examined whether they are equally important with respect to learning or accountability purposes. Therefore, our final hypothesis was developed as follows: (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 (1) learning and (2) accountability? Learning and accountability (P) were represented as dependent variables. The clarity of the object description (O), analysis (A), and conclusions (C) were modeled as independent variables. The next section outlines the method, and gives a more precise description of what we mean by clarity.

3 Method

This study investigates how the four components of an evaluation description (P, O, A, C) influence its usefulness. Broadly speaking, “usefulness” refers to the extent that the description achieves its purpose. Thus, if the purpose is to support learning, then usefulness is related to the extent that it actually helps actors to learn from the exercise. However, there are several, significant methodological problems associated with measuring learning or accountability. For example, a longitudinal study with broad scope would be hard to conduct for practical reasons. There would need to be a large number of identical exercises that could be grouped depending on the type of evaluation description. In addition, controlling for confounding variables is very difficult.

Therefore, instead of focusing on actual usefulness (that is, the extent to which the presentation of an evaluation description helps, in practice, to achieve its purpose), we focus on perceived usefulness (that is, the extent to which professionals believe that a specific form of evaluation description enhances its ability to achieve its purpose). A key assumption in this study is, thus, that there is a relationship between perceived and actual usefulness.

We study two groups of Dutch professionals who use evaluation descriptions in their day-to-day work. Descriptions are presented in the form of vignettes (Jasso 2006; Atzmüller and Steiner 2010; Auspurg and Hinz 2015). These hypothetical examples are based on actual evaluation reports, with a focus on P, O, A, and C. The latter

components are developed as experimental factors. Participants were asked to rate “perceived usefulness” with respect to: (1) learning, and (2) accountability, while individual components were either presented in a “clear” or “unclear” way.

3.1 Design

Vignettes were constructed around three factors, corresponding to the object description (O), the analysis (A), and the conclusion (C). As noted above, these three factors were varied by presenting them in either a clear (1) or unclear (0) manner. Each factor thus takes one of two values (clear/unclear) and, since there are three factors (independent variables) we had a $2 \times 2 \times 2$ experimental design. All possible combinations were presented (Table 1).

Each respondent was asked to rate up to four vignettes. This resulted in hierarchical data, which is shown in Fig. 1. The figure shows that the ratings of different vignettes are nested within respondents. In order to address this distribution, a multilevel model (MLM) with a random respondent effect was applied (Hox 2002; Raudenbush and Bryk 2002). The model considers intra- and inter-respondent responses, and takes into account the fact that every respondent did not rate four vignettes (due to time constraints, among other factors).

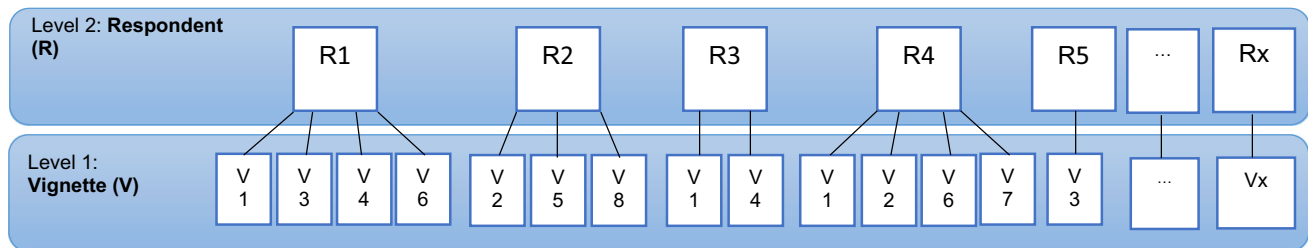
3.2 Vignettes

In a preliminary step, we developed several evaluation descriptions based on a realistic exercise scenario, with the help of crisis management experts. Like respondents, these experts were from the Netherlands. Descriptions were based on a large-scale crash between a passenger train and a truck on a railway crossing. This scenario was selected as we had access to real-life evaluation reports from similar incidents on railway crossings, along with evaluation reports from crisis management exercises that used a similar scenario (Beerens 2019). The aim was to help respondents to develop a clear understanding of the scenario that formed the basis for the vignette.

A key challenge was to establish the respective degree of influence of the three components (O, A, C), in order to understand whether one was perceived as more important than another. As noted above, we use the term “clarity” (Suchan and Dulek 1990; Hartley et al. 2004) to describe the variation that is seen in real documents. Here, textual clarity (Rathjens 1985) is manipulated, to measure how readily the inherent meaning of the text can be understood. Documents and findings from our previous research (Beerens 2019) were used to operationalize clarity and construct the vignettes. Thus, some documents contained

Table 1 Combinations and coding of vignettes

	Clear object description (O_1)		Unclear object description (O_0)	
	Clear conclusion (C_1)	Unclear conclusion (C_0)	Clear conclusion (C_1)	Unclear conclusion (C_0)
Clear analysis (A_1)	Vignette 1	Vignette 2	Vignette 3	Vignette 4
Unclear analysis (A_0)	Vignette 5	Vignette 6	Vignette 7	Vignette 8

**Fig. 1** Distribution of the data

very clear descriptions of O, A, and C, while others did not address them explicitly.

Table 2 provides a sample of text extracts. Each component is described in a clear or unclear manner and extracts are used randomly (that is, they operate as independent “text blocks”). These text blocks were combined into eight, full-text vignettes. Common words, sentences, and styles of writing found in real evaluation reports were used, but adapted to the intended users’ interests and the scenario.

Pre-testing clarity: All eight vignettes were checked to ensure that they correctly applied the Dutch Coordinated Regional Incident Management Procedure¹ (GRIP). Then, they were pre-tested with a group of students of the Master’s program in Crisis and Public Order Management. These students have a professional background in crisis management, but did not participate in the final experiment. Their responses were used to check if the scenario and evaluative texts presented in the descriptions were (un)clear, understandable, realistic, and representative. As there is no single empirical method to measure clarity, which is understood to be a function of the target audience, they were asked to judge it by comparing the text to the definitions of O, A, and C presented in Table 2. More precisely, they were asked to identify whether these aspects were present or not, and if they were clear or unclear. Their qualitative feedback was used to refine descriptions, which were then combined into vignettes of equal length

¹ GRIP is a nationwide incident 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 and Wijkhuijs 2015).

(450–650 words). This was to mitigate the risk of introducing a confounding factor as a longer description might be seen as more useful than a shorter one. Following another round of (student) pre-testing, the vignettes were updated. The final experiment was developed using the online survey tool Qualtrics (2019). The survey was then tested by employees at the Institute for Safety of the Netherlands, who were asked to provide feedback regarding its functioning, timing, and clarity. This highlighted that it took a substantial amount of time to read the vignettes. As we were concerned that this could overload respondents and influence their ratings, we decided to limit their number to four per respondent (taking around a total 15–20 min to complete the experiment). The tool ensured that vignettes were reasonably balanced among participants. Once the experimental setup was finalized, a link was made available to respondents.

3.3 Respondents and Procedure

Professionals were divided into two groups: mayors and operational leaders. These two groups are clearly identifiable within the Dutch crisis management structure and they have different, but related, strategic roles. Mayors are responsible for command and control during local/regional crises, while operational leaders execute commands and lead the response. Directors of Safety Region and other operational users could also participate.

Participants were contacted via email and online (community) newsletters through their respective national networks (for example, the network of mayors, or the Board of Directors of Safety Regions). Two weeks after the first

Table 2 Brief overview (of manipulations) of clarity

A Component (including general description, see 2.1)	B Clear description (1)	C Unclear description (0)
<p>Object description (O)</p> <p>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 organization, an entire organization, or even multiple organizations operating together. An important aspect is the relationship between the object, and the context and scenario.</p>	<p>A clear object description should clarify—</p> <p>(1) Who or what is the subject of the evaluation: <i>The Operational Leader (OL) within the Regional Operational Team (ROT).</i></p> <p>(2) Why the evaluation object exists, that is, what is its role and responsibilities: <i>The OL is (ultimately) responsible for the process of decision making within the ROT.</i></p> <p>(3) What can be expected (in advance) of the evaluation object, that is, what its function is and what it does: <i>The OL ensures that data are collected and shared, the situation is judged, and a well-founded decision is made, shared, and documented.</i></p>	<p>(1) <i>The object (OL) was not specifically mentioned as being the focus of the exercise evaluation, and other actors were also introduced.</i></p> <p>(2) <i>No details regarding the role and responsibilities of the OL were provided and only a generic description of the response organization was provided.</i></p> <p>(3) <i>No details were provided regarding what to expect from the OL and a generic description of the scenario was repeated in the evaluation description.</i></p>
<p>Analysis (A)</p> <p>The analysis supports the arguments put forward, that is, what happened during the exercise and why?</p>	<p>A clear analysis should indicate—</p> <p>(1) How information has been collected: <i>(Observation) notes and reports of the evaluator, meeting reports, and other documents and data.</i></p> <p>(2) Which (value-free) results this has yielded: <i>A (f)actual description of the actual decision-making process was detailed.</i></p> <p>(3) The performed action(s) of the evaluation object: <i>Specifically the OL has taken a number of decisions with regards to the incident and informed the mayor.</i></p>	<p>(1) <i>No data collection methods were described and a general description of the response to the incident was provided.</i></p> <p>(2) <i>There were no results (of O) presented and, again, a generic description of the response to the incident was provided.</i></p> <p>(3) <i>No specific actions (regarding O) were mentioned, only a generic description of the incident was provided.</i></p>
<p>Conclusion (C)</p> <p>The conclusion determines value or performance, that is, how (well) did the object perform? It is the logical consequence of the P, O, A chain, and includes how a judgement is reached. It thus integrates information from P, O, and A.</p>	<p>A clear conclusion should—</p> <p>(1) Give an opinion on the functioning of the evaluation object: <i>The OL has correctly implemented the three-phase decision-making process.</i></p> <p>(2) Judge whether the evaluation object has fulfilled what it had to do (see also object description) in the described context: <i>Although the process was correctly executed some incorrect decisions were made.</i></p>	<p>(1) <i>No opinion was given, only a generic description of the exercise setup and an explanation of the response was provided.</i></p> <p>(2) <i>No judgement was formulated, for example, good or bad and the emergency response process was only vaguely presented.</i></p>

Text in italic (column B and C) refers to the actual text used in the experiment

invitation, a reminder was sent out. In addition, mayors were contacted directly via a personal letter. Operational personnel were provided with further information on the Institute for Safety's website. All correspondence stated the purpose of the research and included a link to the survey.

Participants who visited the study's web page were presented with an explanation of the survey's structure, and what they were asked to do. The first questions

characterized the respondents, and familiarized them with the tool. Then they were introduced to the scenario, and asked questions about its realism and importance. Next, each of the four full vignettes (the combination of individual O, A, and C texts) was presented separately, followed by questions about their perceived usefulness for either learning or accountability. Part 5 was optional and allowed respondents to provide additional, qualitative

Table 3 Overview of respondents' roles

Role	No.	%	Group	No.	%
Regional operational leaders (ROL)	39	46.4	Operational	50	59.5
Other	11	13.1	(= ROL + Other)		
Mayors	28	33.3	Governing	34	40.5
Directors of safety regions	6	7.1	(= Mayors + Directors)		
Total	84	100	Total	84	100

Table 4 Respondent-level descriptive statistics ($N = 84$)

Variable	Min	Max	Mean	SD
Experience				
Years of experience in the current role ($N = 83$, data for 1 Governing respondent are missing)	1	20	8.05	5.056
Current use of evaluations				
Accountability	1	7	4.23	1.623
Learning	1	7	5.50	1.167
Scenario				
Realism	1	7	5.31	1.119
Importance	1	7	5.31	1.481

comments. Finally, they were thanked for their participation and given details about follow-up.

4 Data and Measurements

The first part of the survey generated data that made it possible to characterize individual respondents, their current use of evaluations, and how the scenario was perceived. These data are presented here to provide an overview of the research population. This section also introduces the measurements that were used in vignettes, namely, independent variables and their manipulation, along with the dependent variable and how it was measured.

4.1 Respondents

The first part of the survey aimed to characterize respondents. More precisely, they were asked to provide information about their role in the Netherlands' crisis management structure, their experience, and their geographical location (safety region). These data were used to identify each respondent. They were also asked about how he/she currently uses evaluations. Here, the aim was to investigate attitudes to evaluations, which was used as a control variable. Finally, participants were asked about the realism of the scenario, in order to determine if this could have an influence on the data.

4.1.1 Background

The two main groups of participants were mayors and regional operational leaders. However, as two other groups (Safety Region Directors and "others") could also participate, respondents were asked to indicate their role in the Netherlands' crisis management structure. Table 3 presents an overview. For analysis purposes, the four roles were merged into two groups. The first consisted of people holding a governing or supervisory position: mayors and directors ($N = 34$). The second merged (regional) operational leaders and the "other" group² ($N = 50$). As we anticipated that the respondent's background could affect the outcome, these data were used as a control variable.

Respondents gave details regarding their experience in their current role. Table 4 shows that both groups had average experience of approximately eight years, ranging from 1–20 years in both cases. Finally, they were asked to indicate which of the 25 Dutch Safety Regions they worked in. This found that 22 of the 25 regions (88%) were represented (Fig. 2).

² A qualitative analysis of the "other" group showed that their function was comparable to that of an operational leader. Examples are (municipal) crisis management advisors, incident commanders, preparedness experts, emergency planners, and crisis coordinators.

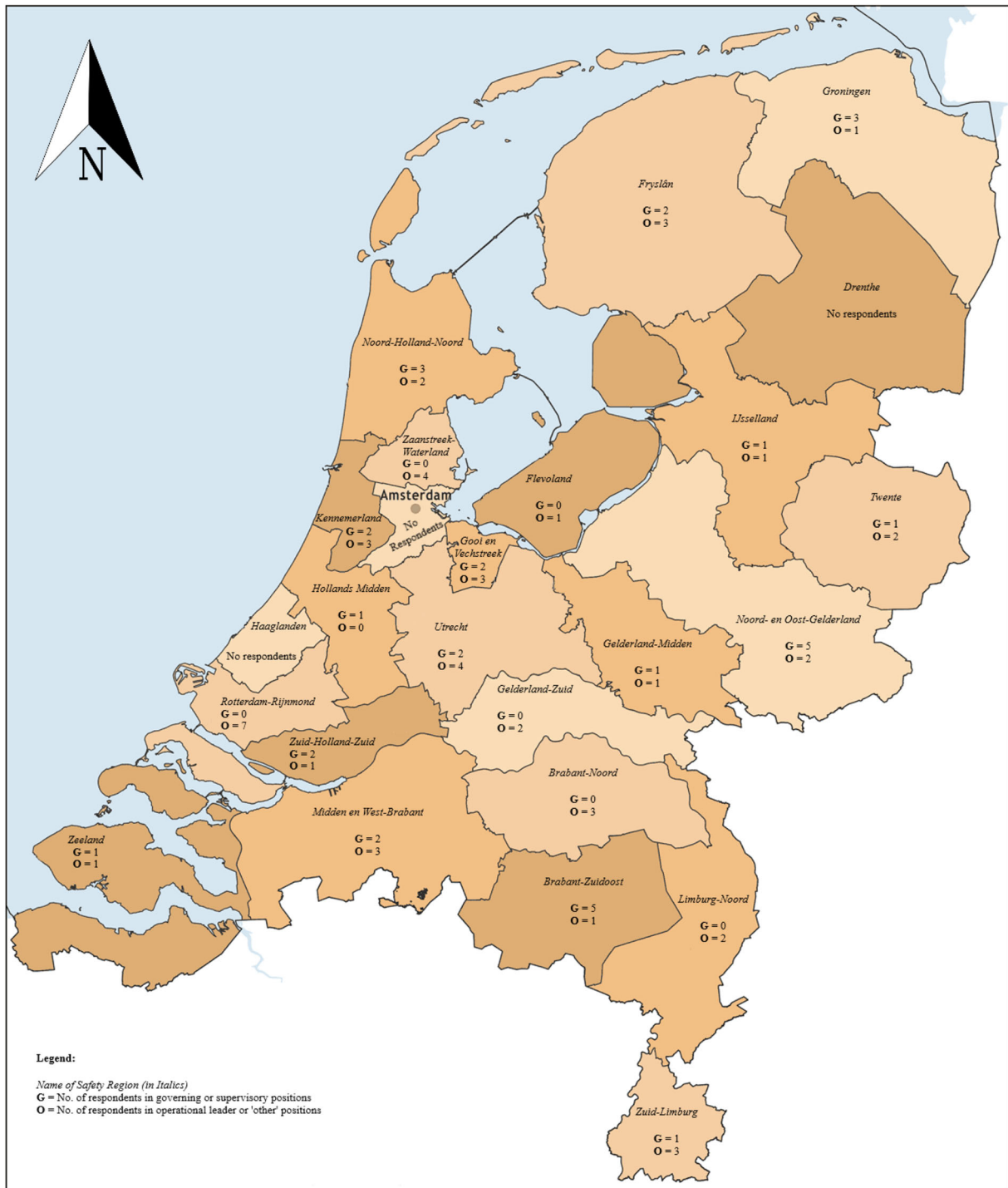


Fig. 2 Geographical distribution of respondents. Source Adapted from Map 25 Safety Regions (Imergis.nl 2020)

4.1.2 Use of Evaluations

The first question concerned how respondents used evaluations to justify their own actions and/or performance, which is related to accountability. The second concerned how they used evaluations to learn from previous actions.

The two statements were rated on a 7-point Likert-type scale with options ranging from “not at all” (1) to “extensive use” (7). Here again, we anticipated that these data might affect outcomes, and it was therefore used as a control variable in the analysis.

4.1.3 Realism and Importance of the Scenario

Respondents were asked how realistic and important they perceived the scenario. Answers were rated on a 7-point Likert-type scale ranging from “not at all” (1) to “very” (7). Table 4 presents descriptive statistics. Overall ($N = 84$), the mean rating of realism was 5.31 (SD 1.119), suggesting that the scenario was somewhat realistic. Responses for importance were similar ($M = 5.31$; SD 1.481). Differences were found between groups. The Governing group rated the scenario as “important” ($M = 5.82$; SD 1.114) while the Operational group rated it as “somewhat important” ($M = 4.96$; SD 1.603). An independent samples t test found a significant difference between these ratings ($p = 0.005$).

4.2 Vignettes

The eight vignettes reflected the contents of current evaluation reports and were developed to consistently reflect the key experimental factors. Three factors were tested as independent variables: the description of the object (O), the analysis (A), and the conclusion (C). The fourth component, purpose (P) was implemented as the dependent variable.

4.2.1 Independent Variables

The three independent variables were manipulated by varying their clarity in the following ways: O (0 = unclear, 1 = clear), A (0 = unclear, 1 = clear), and C (0 = unclear, 1 = clear). Each factor corresponded to a text block, and a complete vignette contained three such blocks, one for each factor. The manipulation is illustrated in Table 2.

4.2.2 Dependent Variables

The experiment explored how differences in O, A, and C influenced the perceived usefulness of a vignette. In this context, “perceived usefulness” indicates the extent to which the respondent believes that the contents of a specific vignette help to achieve its purpose (based on Davis 1989). Thus, if the purpose is to support learning, then usefulness is related to the extent that the description helps actors to learn. Purpose was measured by nine statements, rated on a 7-point Likert-type scale, with answers ranging from “strongly disagree” (1) to “strongly agree” (7).

A principal axis factor analysis with Oblimin with Kaiser normalization was run. This revealed two factors: Y1 was associated with six statements (usefulness with respect to learning), while Y2 was associated with three statements (usefulness with respect to accountability). The

correlation between them was 0.661, indicating that a high score on learning was correlated with a high score on accountability. This two-factor solution met the interpretability criterion. Internal consistency was adequate (Cronbach’s $\alpha \geq 0.95$), notably Y1 was 0.963, compared to 0.944 for Y2. We therefore conclude that the nine items were suitable measures of usability for learning and accountability. In total, 277 vignettes were rated by the 84 respondents. Mean Y1 was 3.788 (SD 1.488), compared to 3.439 (SD 1.474) for Y2. The correlation between these two means was 0.642, again indicating that a high score on learning is correlated with a high score on accountability.

Table 5 provides an overview of means for each dependent variable as a function of the three independent variables. This shows that, with the exception of O, means are higher if the component is described clearly—for both learning and accountability.

5 Analysis and Results

Analyses were run using IBM SPSS Statistics for Windows, version 25 (IBM Corp. 2019). Given the nested structure of our data, we estimated various MLMs. One-tailed tests were run for O, A, and C. Models were constructed for Y1 and Y2, starting with an unconditional null model. As we expected to find a positive relation with prior use of evaluation descriptions, we controlled for fixed effects. The respondent’s background (Operational or Governing) was also expected to affect ratings, and was thus also added as a control variable.

5.1 Usability for Learning (Y1)

The first step was to create a null model to determine the intraclass correlation. Inter-respondent variance was 0.948, compared to intra-respondent variance of 1.305. The intraclass correlation coefficient (ICC) was calculated as 0.421, indicating that 42.1% of the difference in usability ratings was due to inter-respondent variance.

The full model (Table 6) shows that both A and C have significant, positive effects on usability for learning purposes. However, O was not significant. The conclusion has a particularly strong effect—a clear conclusion is associated with a 1.146 higher rating on a 7-point Likert-type scale, compared to an unclear conclusion. Similarly, a clear analysis is associated with a 0.266 higher rating on a 7-point Likert-type scale, indicating that a clear conclusion is perceived as “very important,” and a clear analysis is “somewhat important.”

The full model also controlled for the user’s background. This found a significant effect (0.045), suggesting that users with an operational background rate evaluation

Table 5 Means for each dependent variable

Aspect	Clarity	No.	Min	Max	Mean learning (Y1)	SD	Mean accountability (Y2)	SD
Description of the object (O)	Clear (1)	140	1	7	3.737	1.466	3.405	1.501
	Unclear (0)	137	1	7	3.841	1.514	3.475	1.450
Analysis (A)	Clear (1)	142	1	7	3.889	1.443	3.531	1.471
	Unclear (0)	135	1	7	3.683	1.532	3.343	1.476
Conclusion (C)	Clear (1)	138	1	7	4.318	1.401	3.749	1.484
	Unclear (0)	139	1	7	3.263	1.386	3.132	1.403

Table 6 Multilevel model summary, main effects

Regression effects and standard errors	Y1: Usability for learning		Y2: Usability for accountability	
	Estimate	SE	Estimate	SE
Intercept	1.648**	0.703	1.955***	0.403
Object description (O)	-0.127*	0.124	-0.139*	0.141
Analysis (A)	0.266**	0.121	0.206*	0.139
Conclusion (C)	1.146***	0.123	0.722***	0.141
<i>Background (operational vs. governing)</i>	<i>0.539**</i>	<i>0.264</i>	<i>0.570**</i>	<i>0.250</i>
<i>Current use = Accountability</i>			0.196**	0.078
<i>Current use = Learning</i>	0.227**	0.117		
Intra-respondent variance	0.888***	0.091	1.195***	0.121
Inter-respondent variance	1.043***	0.217	0.814***	0.192
-2 restricted log likelihood	888.454		938.061	

Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$

2-tailed tests are presented in italics

descriptions more highly than users with a governing background. We also controlled for respondents' current use of evaluations for learning purposes. This also found a significant effect on usefulness ratings (0.028), which is comparable in size to the previous result. Notably, respondents who are already using evaluations for learning purposes rated them more highly with regard to their usefulness.

5.2 Usability for Accountability (Y2)

Here again, the first step was to create a null model. Inter-respondent variance was found to be 0.814, compared to 1.379 for intra-respondent variance. The ICC was calculated as 0.371. Thus, 37.1% of the difference in usability for accountability was due to inter-respondent variance. The accountability results shown in Table 6 are similar to the full model for learning. Here again, O is not significant, while C is (0.000). In this case, conclusions have a substantial effect: a clear conclusion is associated with a 0.722 higher rating on a 7-point Likert-type scale, compared to an

unclear conclusion. On the other hand, the effect of A is marginal (0.07). It appears that a good conclusion is more important than a good analysis when using an evaluation description for accountability.

As before, the full model controlled for the user's background and, here again, this was found to have a significant effect (0.025). Users with an operational background rate evaluations more highly than users with a governing background. Finally, we controlled for current use of evaluations for accountability, which also found a significant effect (0.007) on usefulness. Like learning, people already using evaluations for accountability purposes rate them more highly with regards to their usefulness.

5.3 Interactions

We also investigated whether particular combinations of (un)clear O, A, and C are consistent with a significant increase (decrease) in usefulness. To this end, we added

two- and three-way interactions to the MLMs. However, as these were insignificant, they are not reported.

6 Discussion

This experiment showed that the clarity of conclusions (C) had a significant, positive effect on the perceived usefulness of an evaluation description, for both learning and accountability. In addition, the clarity of the analysis (A) had a significant effect for learning. These results might appear obvious, but we do not agree. First, although it might seem obvious that an evaluation description without an analysis or conclusion is not very useful, that was not the case here—we did not remove the analysis or the conclusion altogether. Instead, we focused on two ways of writing: clearly and unclearly. It should be noted that our examples mirrored text that was taken from current, real-life evaluation descriptions, and it is somewhat surprising that unclear versions are still used in practice. More specifically, our results indicate that how evaluation descriptions are presented does matter and support standardization, or at least providing guidelines to improve the clarity of (at a minimum) conclusions. Other results came as more of a surprise.

There is a clear logical connection between the components of an evaluation description: without a clear purpose it is difficult to arrive at focused conclusions; without a clear object description it is difficult to present an analysis; and without a clear analysis it is difficult to reach robust conclusions. We expected each of these individual components to influence usefulness—hence our investigation of whether particular combinations of (un)clear O, A, and C were consistent with a significant increase (decrease) in usefulness. However, our results suggest that participants did not consider relationships between components as essential, as neither learning nor accountability usefulness increased (or decreased) for particular combinations of O, A, and C.

Although there were differences between vignettes that contained (un)clear object descriptions, these results were not statistically significant. Instead, judgements appear to be based primarily on the analysis and conclusions. This observation is in line with research that examined learning preferences among employees in the fire and rescue services in the Netherlands (Instituut Fysieke Veiligheid 2017). The latter study revealed that respondents learn by: (1) sharing experiences; (2) being provided with sufficient background information and reasoning with regards to findings or insights; and (3) avoiding mistakes. Our findings are consistent with (2), as the analysis provides information about what happened, how, and why. Information presented in this section can be used in order to

(better) understand and learn, for example, why the mechanisms behind certain activities worked (or not). Despite the lack of a significant result, we believe that O is an important component of real evaluation descriptions. It is the point of departure for understanding A and C, as it provides useful information regarding, for example, tasks, responsibilities, or activities.

We recognize that there is a difference between reality and what is recorded on paper. As Heath (1998) notes, it is likely that hindsight bias, or time and organizational distortion occur when analysing and evaluating DRM documentation. Birkland (2009) claimed that such documents simply reflect the preferred social construction of a problem by a group and its target populations; therefore, reports should not necessarily be seen as detailed or precise accounts of what happened. They are merely reflections or outcomes of a systematic process that transfers data into information or knowledge, and seek to reliably record the most important elements in order to discuss and share them.

Nevertheless, Beerens (2019) concluded that there is no common framework to ensure that the intended purposes are achieved. In particular, data collection lacks transparency and the underlying reasoning is often unclear. Our new concept of the “evaluation description” offers practitioners a way to make evaluations more useful for their users. Although there is no detailed prescription, it provides some basics that can be further investigated. Similar to the concept of the risk description, it can support the creation of a common knowledge base. This knowledge base can, for example, support root cause analyses, provide details of why (and when) a system does not seem to be ready to respond, and provide guidance on what can be changed.

Finally, the benefits of an exercise evaluation extend far beyond the physical product. Personnel are likely to learn lessons even without a report (Perry 2004; Borell and Eriksson 2008; Birkland 2009; Nohrstedt et al. 2018). Although hands-on experience is important, it must be recognized that, as in a real disaster, an exercise is an opportunity for collective learning (Klein et al. 2005; Gebbie et al. 2006). Biddinger et al. (2008) note that exercises can significantly improve preparedness on two levels: (1) at the individual level they present an opportunity to educate personnel on disaster plans and procedures through hands-on practice; and (2) at the institutional and/or system level, they can reveal gaps and weaknesses, and clarify specific roles and responsibilities. Exercises and their evaluation can, thus, generate insights that extend far beyond the persons involved—notably, to exercise designers, facilitators, and evaluators (Borodzicz and Van Haperen 2002).

Optimizing the evaluation and related products requires effectively communicating the findings to relevant actors, especially if used to reflect on the experience of others. We argue that the evaluation description can support learning—in particular the transfer from lessons identified to lessons learned, and the creation of a knowledge base. It not only provides a framework that ensures that findings can be shared, but also offers scope for situational customization. However, the robustness of the concept should be investigated in future (longitudinal) experimental research.

Despite its contributions, this study has some limitations. First, while O, A, and C were manipulated with respect to clarity, no information was completely removed. This was because our study focused on the usefulness of the description as a whole, rather than individual components. Therefore, generic information was always provided. Although not evaluative, or related to a specific component, the respondent might perceive these data as clear and useful. Future research could focus on each component individually, as this would help to establish a more detailed picture of the information that needs to be presented.

Second, previous research has highlighted the importance of a clear link between the report and users' needs (Beerens 2019). Different groups use reports for different purposes and, in turn, require different evaluation products. In practice, a DRM system is comprised of many more (sub)groups than the two represented in this study. In particular, exercise designers, facilitators, and other educational staff are likely to (indirectly) use evaluations in order to create new exercises or develop training programs. As noted above, using the overall evaluation product to develop tailor-made evaluations could be further investigated and operationalized.

Third, it could be argued that respondents themselves were the object of this study. This might have influenced the findings, in particular, with regard to O, as participants may have already understood the tasks and responsibilities. This could be investigated in future research by running the experiment with user groups that are not the object of the evaluation.

7 Conclusion

This study investigated whether the clarity of an evaluation description has a (positive) effect on its perceived usefulness, with respect to learning and accountability. The “evaluation description” concept assumes that several, distinct components contribute to a report's usefulness. The present study looked at: the purpose of the evaluation (P), the object description (O), the analysis (A), and the

conclusions (C). The latter three aspects were manipulated (via an (un)clear description) to investigate their effect on perceived usefulness. The experiment used vignettes and a population of professionals working in the Netherlands.

We conclude that the clarity of conclusions (C) has a significant effect on perceived usefulness—for both learning and accountability purposes, and operational and governing users. The analysis (A) was significant for learning purposes, with a marginally significant effect for accountability. Although no significant effect was found for the object description (O), we still believe that clarity is important from a practical perspective.

Our findings indicate the importance of how emergency exercise evaluations are documented, and underline the need for clear guidelines. We believe that such guidelines should help professionals to improve their work, notably by indicating the criteria used to arrive at any conclusions, and supporting arguments. Otherwise, reports risk being perceived as “fantasy documents” (Birkland 2009), gathering dust on a shelf. Finally, we should not forget that an evaluation is not an end in itself; rather it is a means to achieving a higher goal or purpose.

Acknowledgements We gratefully acknowledge funding from the Institute for Safety (IFV) of the Netherlands and the Swedish Civil Contingencies Agency (MSB). In addition, we would like to thank the volunteer reviewers for their statistical expertise and feedback on the methodology.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abrahamsson, M., H. Hassel, and H. Tehler. 2010. Towards a system-oriented framework for analysing and evaluating emergency response. *Journal of Contingencies and Crisis Management* 18(1): 14–25.
- Atzmüller, C., and P.M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology* 6(3): 128–138.
- Auspurg, K., and T. Hinz. 2015. *Factorial survey experiments*. Thousand Oaks, CA: Sage.
- Aven, T. 2010. On how to define, understand and describe risk. *Reliability Engineering and System Safety* 95(6): 623–631.
- 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): Article 221.
- Beerens, R.J.J., and H. Tehler. 2016. Scoping the field of disaster exercise evaluation—A literature overview and analysis. *International Journal of Disaster Risk Reduction* 19: 413–446.
- Biddinger, P.D., R. Orfaly Cadigan, B.S. Auerbach, J.L. Burstein, E. Savoia, M.A. Stoto, and H.K. Koh. 2008. On linkages: Using exercises to identify systems-level preparedness challenges. *Public Health Reports* 123(1): 96–101.
- Birkland, T.A. 2009. Disasters, lessons learned, and fantasy documents. *Journal of Contingencies and Crisis Management* 17(3): 146–156.
- Boin, A., P. 't Hart, E. Stern, and B. Sundelius. 2017. *The politics of crisis management*, 2nd edn. Cambridge: Cambridge University Press.
- Borell, J., and K. Eriksson. 2008. Improving emergency response capability: An approach for strengthening learning from emergency response evaluations. *International Journal of Emergency Management* 5(3–4): Article 324.
- Borodzicz, E., and K. Van Haperen. 2002. Individual and group learning in crisis simulations. *Journal of Contingencies and Crisis Management* 10(3): 139–147.
- Bovens, M., P. 't Hart, and S. Kuipers. 2008. The politics of policy evaluation. In *The Oxford handbook of public policy*, ed. M. Moran, M. Rein, and R.E. Goddin, 317–333. Oxford: Oxford University Press.
- Bradshaw, C.C., and T.A. Bartenfeld. 2009. Exercise evaluation guides for public health emergency preparedness. *Homeland Security Affairs Journal* 5(3): 1–41.
- Calidoni-Lundberg, F. 2006. Evaluation: Definitions, methods and models: An ITPS framework. Östersund: Swedish Institute for Growth Policy Studies. https://pdfs.semanticscholar.org/6847/9ec9fc396e1a3f1ab859502e2e0cb0267d41.pdf?_ga=2.138291760.1448429144.1592574523-1144041233.1589189182. Accessed 15 Jun 2020.
- Davis, F.D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13(3): Article 319.
- Duarte da Costa, A., J.O. Gomes, M.R. da Silva Borges, and P.V. Rodrigues de Carvalho. 2013. ASC model: A process model for the evaluation of simulated field exercises in the emergency domain. In *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, ed. T. Comes, F. Fiedrich, S. Fortier, J. Geldermann, and T. Müller, 551–555. Baden-Baden: Karlsruher Institut für Technologie (KIT).
- Etkin, J., and A. Sela. 2016. How experience variety shapes postpurchase product evaluation. *Journal of Marketing Research* 53(1): 77–90.
- Frye, A.W., and P.A. Hemmer. 2012. Program evaluation models and related theories: AMEE guide No. 67. *Medical Teacher* 34(5): 288–299.
- Gebbie, K.M., J. Valas, J. Merrill, and S. Morse. 2006. Role of exercises and drills in the evaluation of public health in emergency response. *Prehospital and Disaster Medicine* 21(3): 173–182.
- Hartley, J., E. Sotto, and C. Fox. 2004. Clarity across the disciplines. *Science Communication* 26(2): 188–210.
- Heath, R. 1998. Looking for answers: Suggestions for improving how we evaluate crisis management. *Safety Science* 30(1–2): 151–163.
- Herting, N., and E. Vedung. 2012. Purposes and criteria in network governance evaluation: How far does standard evaluation vocabulary take us? *Evaluation* 18(1): 27–46.
- Heumüller, E., S. Richter, and U. Lechner. 2012. Towards a framework for command post exercises. In *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, ed. L. Rothkrantz, J. Ristvej, and Z. Franco, 1–6. Vancouver: Simon Fraser University.
- Hox, J. 2002. *Multilevel analysis: Techniques and applications*. Mahmah, NJ: Lawrence Erlbaum Associates Inc.
- IBM Corp. 2019. *SPSS*. Armonk, NY: IBM Corp.
- Imergis.nl. 2020. Map of 25 safety regions (*Fysieke Veiligheid—25 Veiligheidsregio's 2020*). Zeist: Imergis Organisatiebloe. <https://www.imergis.nl/map/2020/2020-Veilighedsregio-1200.png>. Accessed 25 Jun 2020 (in Dutch).
- Instituut Fysieke Veiligheid. 2017. Learning and the Fire Services: About how people learn—Also within the Fire Services (*Leren and brandweer: Over hoe mensen leren—ook binnen de brandweer*). Arnhem: Instituut Fysieke Veiligheid. <https://www.ifv.nl/kennisplein/Documents/20170316-BA-Leren-en-brandweer.pdf>. Accessed 19 Jun 2020 (in Dutch).
- ISO (International Organization of Standardization). 2013. *International standard, societal security—Guidelines for exercises ISO-22398-2013*. Geneva: ISO.
- Jasso, G. 2006. Factorial survey methods for studying beliefs and judgments. *Sociological Methods and Research* 34(3): 334–423.
- Klein, K.R., D.C. Brandenburg, J.G. Atas, and A. Maher. 2005. The use of trained observers as an evaluation tool for a multi-hospital bioterrorism exercise. *Prehospital and Disaster Medicine* 20(3): 159–163.
- Lin, L., A. Nilsson, J. Sjölin, M. Abrahamsson, and H. Tehler. 2015. On the perceived usefulness of risk descriptions for decision-making in disaster risk management. *Reliability Engineering and System Safety* 142: 48–55.
- Lin, L., C. Rivera, M. Abrahamsson, and H. Tehler. 2017. Communicating risk in disaster risk management systems—Experimental evidence of the perceived usefulness of risk descriptions. *Journal of Risk Research* 20(12): 1534–1553.
- Lincoln, Y.S., and E.G. Guba. 1980. The distinction between merit and worth in evaluation. *Educational Evaluation and Policy Analysis* 2(4): 61–71.
- Månsson, P., M. Abrahamsson, and H. Tehler. 2019. Aggregated risk: An experimental study on combining different ways of presenting risk information. *Journal of Risk Research* 22(4): 497–512.
- Nohrstedt, D., F. Bynander, C. Parker, and P. 't Hart. 2018. Managing crises collaboratively: Prospects and problems—A systematic literature review. *Perspectives on Public Management and Governance* 1(4): 257–271.
- Perry, R.W. 2004. Disaster exercise outcomes for professional emergency personnel and citizen volunteers. *Journal of Contingencies and Crisis Management* 12(2): 64–75.
- Peterson, D.M., and R.W. Perry. 1999. The impacts of disaster exercises on participants. *Disaster Prevention and Management: An International Journal* 8(4): 241–255.
- Qualtrics. 2019. *Qualtrics*. Provo, UT: Qualtrics. <https://www.qualtrics.com>. Accessed 15 Jun 2020.
- Rathjens, D. 1985. The seven components of clarity in technical writing. *IEEE Transactions on Professional Communication* PC-28(4): 42–46.
- Raudenbush, S.W., and A.S. Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*, 2nd edn. Newbury Park, CA: Sage.
- Savoia, E., F. Agboola, and P. Biddinger. 2014. A conceptual framework to measure systems' performance during emergency preparedness exercises. *International Journal of Environmental Research and Public Health* 11(9): 9712–9722.
- Savoia, E., L. Lin, and G.M. Gamhewage. 2017. A conceptual framework for the evaluation of emergency risk communications. *American Journal of Public Health* 107(S2): 208–214.
- Scriven, M. 1980. *The logic of evaluation*. Inverness, CA: Edgepress.

- Scriven, M. 1991. *Evaluation thesaurus*. Beverly Hills, CA: Sage.
- Sinclair, H., E.E. Doyle, D.M. Johnston, and D. Paton. 2012. Assessing emergency management training and exercises. *Disaster Prevention and Management: An International Journal* 21(4): 507–521.
- Skryabina, E., G. Reedy, R. Amlôt, P. Jaye, and P. Riley. 2017. What is the value of health emergency preparedness exercises? A scoping review study. *International Journal of Disaster Risk Reduction* 21: 274–283.
- Suchan, J., and R. Dulek. 1990. A reassessment of clarity in written managerial communications. *Management Communication Quarterly* 4(1): 87–99.
- Swedish Civil Contingencies Agency. 2011. *Handbook evaluation of exercises*. Stockholm: Swedish Civil Contingencies Agency. <https://www.msb.se/RibData/Filer/pdf/25885.pdf>. Accessed 13 Jun 2020.
- Thomas, T.L., E.B. Hsu, H.K. Kim, S. Colli, G. Arana, and G.B. Green. 2005. The incident command system in disasters: Evaluation methods for a hospital-based exercise. *Prehospital and Disaster Medicine* 20(1): 14–23.
- Uhr, C., H. Johansson, and L. Fredholm. 2008. Analysing emergency response systems. *Journal of Contingencies and Crisis Management* 16(2): 80–90.
- UK Cabinet Office. 1998. *The exercise planners guide*. London: UK Cabinet Office. <http://www.cabinetoffice.gov.uk/resource-library/exercise-planners-guide>. Accessed 19 Jun 2020.
- UNDRR (United Nations Office for Disaster Risk Reduction). 2020. *Words into action—Design and conduct of simulation exercises—SIMEX*. Geneva: UNDRR.
- U.S. Department of Homeland Security. 2020. *Homeland Security Exercise and Evaluation Program (HSEEP)*. Washington, DC: U.S. Department of Homeland Security.
- van Duin, M., and V. Wijkhuijs. 2015. The flexibility of GRIP (*De flexibiliteit van GRIP*). <https://www.ifv.nl/adviesennovatie/Documents/201504-IFV-De-flexibiliteit-van-GRIP.pdf>. Accessed 15 Jun 2020 (in Dutch).
- Wybo, J.-L. 2008. The role of simulation exercises in the assessment of robustness and resilience of private or public organizations. In *Resilience of cities to terrorist and other threats*, ed. H.J. Pasman, and I.A. Kirillov, 491–507. New York: Springer.

International Journal of Disaster Risk Science is a copyright of Springer, 2020. All Rights Reserved.