Evaluation-driven Disaster Management Exercises: A Collaborative Toolkit

Sebastian Henke
Chair of Information Systems
and Supply Chain Management,
University of Münster
sebastian.henke@ercis.uni-muenster.de

Adam Widera
Chair of Information Systems
and Supply Chain Management,
University of Münster
adam.widera@ercis.uni-muenster.de

Bernd Hellingrath
Chair of Information Systems
and Supply Chain Management,
University of Münster
bernd.hellingrath@ercis.uni-muenster.de

ABSTRACT

Disaster management exercises are a core component of humanitarian organizations' preparedness strategies. They entail diverse purposes, from training capabilities of participants to testing response plans to enhancing collaboration between organizations and many more. However, it is uncertain how much exercises contribute to preparedness. Rigorous evaluation is needed to exploit learning opportunities of an exercise. Therefore, exercises must target evaluable objectives, which is complicated by the socio-technical openness of the exercise system, the heterogeneity of organizational needs, and the scarcity of resources. Many different tools aim to support evaluation but are limited to specific use cases, resulting in a fragmented overview for practitioners. Due to the excessive effort involved, practitioners often consider exercise evaluation to be of secondary importance. This study thus proposes the conceptual design of a combined toolkit that supports the practitioners in a more rigorous but resource-efficient evaluation to make disaster management exercises more evaluation-driven.

Keywords

Disaster Management Exercise, Evaluation, Learning, Collaboration

INTRODUCTION

The number of disasters has been growing globally over the last few years. This trend is expected to continue due to the climate crisis (CRED, 2016). While disasters are still low-probability events, they can have a high adverse impact on the people affected. Disaster management (DM) is thus essential to reduce a disaster's potential impact. One key DM strategy is strengthening the preparedness of responding communities and organizations. It requires a set of diverse capabilities that cannot only be taught theoretically but must be trained actively (Sommer *et al.*, 2013). However, disasters are no appropriate learning environment as they are infrequent, uncontrollable, and high-risk events. Hence, DM organizations perform disaster management exercises (DMEs) that simulate a disaster event in a low-risk environment where the participating exercise players mimic a response situation.

Although it is a common and intuitive belief that DMEs contribute to preparedness, the informative value of their actual contribution is limited (Verheul *et al.*, 2018). How do we know if the *right things* have been exercised? Have they been exercised the *right way*? Misaligned exercises or a biased perception of learning can even increase the gap between perceived preparedness and actual preparedness, creating a false sense of security (Gebbie *et al.*, 2006). A DME needs evident evaluation of the players to assess their capabilities and identify learning potentials. But it also needs validation of its own process and management to align the *right things* in the *right way*, as well as to provide accountability for that alignment. Unfortunately, evaluation is often secondary in practice (Miller *et*

al., 2017; Skryabina *et al.*, 2017). DMEs are conducted because it is the expected norm to conduct them in view of their believed contribution. Consequentially, the ratio between benefit and effort of evaluation is perceived as poor due to several challenges (Beerens & Tehler, 2016):

DM organizations usually operate under resource scarcity (Magnusson *et al.*, 2019). DMEs are thus scheduled within a limited time-window and a limited pool of human resources, experience, and expertise. While small discussion-based DMEs may be implemented quickly, full-scale multi-organization DMEs require significant resource investments (Beerens, 2021). DME design and execution are prioritized because they are prerequisite for DME evaluations. Their errors are directly noticeable by the participants, while evaluation errors may be observed only with a delay (Grunnan & Fridheim, 2017).

DMEs must mirror realistic scenarios and environments to achieve appropriate training or meaningful testing. This conflicts with the methodological rigor of evaluation. The more realistic a DME is, the more complex, dynamic, and susceptible to random factors it is, which makes it more difficult to measure, repeat, or reconstruct (Borodzicz & van Haperen, 2002). Viewing evaluation simply as a separate step after exercise execution overlooks context dependencies in the open socio-technical system. For example, the effectiveness of exercise players is not the same as their performance (Baroutsi, 2018). The expressiveness of isolated quantitative indicators, such as the number of causalities, is limited. It is important to include qualitative information providing context within the evaluation. Currently, this is performed narratively by evaluators, which carries a risk of high subjectivity and a lack of evidence. Evaluators are susceptible to various biases, e.g., the hindsight or the confirmation bias (Comes, 2016; MSB, 2017). To counteract this, their evaluations must therefore be traceable and justified, ideally backed up by supplementary non-narrative data. There are guidelines providing structure to the evaluation process, e.g., handbooks such as the *Homeland Security Exercise and Evaluation Program* (DHS, 2020). However, they do not specify methods that guarantee sufficient evidence. One reason is that DMEs are dynamic processes characterized by unique settings and organizational needs. This results in a heterogeneity of requirements that aggravates methodological standardization (Sheikhbardsiri *et al.*, 2018).

DM organizations often introduce multiple objectives within one DME to save resources, risking incompatibility. For example, they use DMEs for performance measurement of a team while also pursuing learning of exercise players. But performance and learning do not form a clear relationship (Borodzicz & van Haperen, 2002). Learning has the highest potential when there is room for experimentation and decisions are carefully reflected in an error-acceptant culture, which may cause poor performance (Berlin & Carlström, 2014). Additionally, both must be measured with different methods (Seijts & Latham, 2005). DMEs hence must be evaluation-driven: compatible objectives must be defined at an early stage so that the methods for exercise and evaluation can be aligned. The objectives must be SMART, i.e., specific, measurable, achievable, relevant, and time-bound (DHS, 2020). This can be a challenging and resource-intensive task. Current studies on DME evaluation aim at providing more rigor but their methods are usually only applicable to the presented use case, or their design decisions are not sufficiently justified, so the results are not transferable to other contexts. This creates a fragmented overview of support options in which practitioners can hardly navigate. Consequently, DM organizations often stick to traditional, non-evaluation-driven approaches, leading to less useful evaluation products and, thus, a worse perception of benefit (Beerens & Tehler, 2016). More rigorous evaluation is needed to facilitate learning within and through DMEs.

This research takes one step in this direction through the conceptual design of a DME evaluation toolkit. A toolkit has the potential to universally support a more rigorous evaluation process and a successive improvement of DMEs while reducing resource requirements, e.g., by automatically customizing methods, integrating tools, and eliminating redundant work (Sheikhbardsiri *et al.*, 2018). The corresponding research question is: How can such a toolkit be designed conceptually? We provide an overview of the toolkit's essential functionalities, requirements, and potential issues, serving as a blueprint for future research to implement, develop and refine the design.

In the following, we first depict the research methodology. Then, we mention the key points from the identified state of research and practice of DME evaluation, followed by a presentation of the toolkit's requirements and conceptual design. Finally, we critically reflect on the design in the validation and discussion.

METHODOLOGY

The DM domain is characterized by a gap between academics and practitioners (Browne *et al.*, 2018). Thus, our research follows a Design Science Research approach targeting a balance between practical relevance and theoretical rigor. While the designed artifact is a treatment for a practical problem, the design process contributes to the theoretical knowledge base surrounding DME evaluation (Hevner & Chatterjee, 2010). The presentation of the designed artifact is a prescriptive conceptual model of the toolkit's logical architecture associated with its required functions given below.

The applied Design Science Research methodology by Wieringa (2014) encompasses the steps "problem investigation", "treatment design" and "treatment validation". It promotes the frequent integration of practitioners across all steps. Thus, this research incorporates different techniques for the data collection as well as for the evaluation of the resulting requirements and artifact. The "problem investigation" step aims to shed light on the state of research and practice. Two semi-structured, two-hour interviews with DME experts from the field introduced the problem, shaping the subsequent systematic literature review. In parallel, a survey of fifteen practitioners triangulated and supplemented the literature review results. The "treatment design" step started with requirements engineering following Braun *et al.* (2015). It included the identification of stakeholders, requirements elicitation, validation by a semi-structured, two-hour expert interview, and critical reflection. The requirements were used to assess available artifacts. After no sufficient artifact could be identified, a new artifact was designed. Finally, the "treatment validation" step covered two complementary techniques. First, the artifact was demonstrated to four domain experts. The experts were then asked to review the model based on their experience as a first step toward external validity concerning practical relevance. Second, we assessed the internal validity of the artifact checking for insufficient requirement fulfilment to provide a critical review for future refinement.

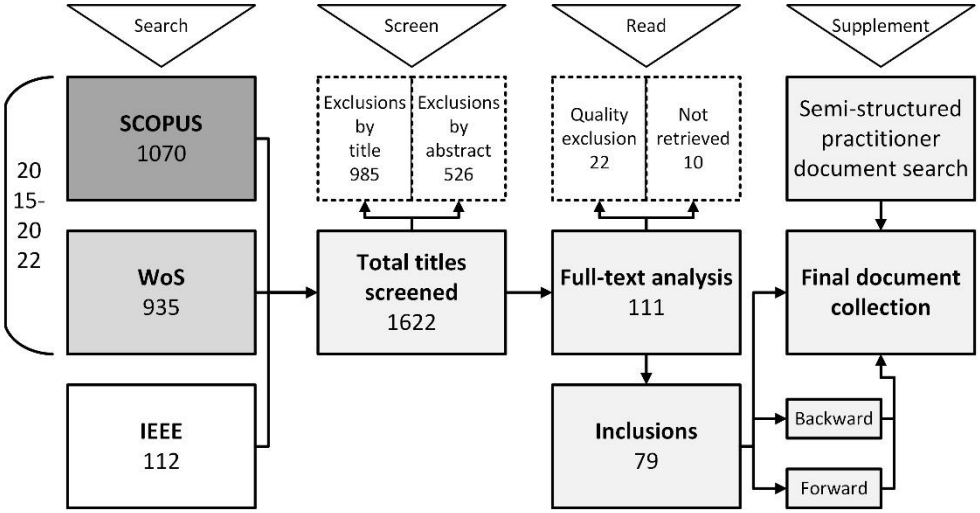

Figure 1 - Systematic Literature Review

The systematic literature review following vom Brocke *et al.* (2015) aimed at finding current processes, concepts, gaps, and challenges of DME and its evaluation, as well as proposed methods, models, theories, and tools from the academic and practitioner literature. The review required broad coverage to reduce the selective bias of the academic researchers. Relevant literature reviews were studied prior to querying databases to familiarize with the subject (e.g., Beerens & Tehler, 2016; Hsu *et al.*, 2004; Miller *et al.*, 2017; Savoia *et al.*, 2017; Sheikhbardsiri *et al.*, 2018; Skryabina *et al.*, 2017; Williams *et al.*, 2008). As the latest comprehensive reviews only covered literature until 2015, a new review was performed to identify the current state of research and practice. The search strategy reconstructed the process of Beerens & Tehler (2016), with "evaluation", "exercise" and "disaster" as core terms, from which they derived a list of relevant synonyms. The results were limited to conference and journal proceedings from 2015 onwards, as the earlier key publications could be identified by backward search. The search was performed on 05-04-2022 using the databases SCOPUS (1070 results), Web of Science (935 results) and IEEE Xplore (112 results), as depicted in *Figure 1*. After de-duplication and exclusion, 79 articles were considered relevant. Backward and forward search was performed on the highest quality articles, based on topic match, citations, and source quality. The backward search also covered practitioner documents, e.g., DME guidance handbooks (e.g., AIDR, 2012; DHS, 2020; MSB, 2011, 2017; SEMC, 2021), supplemented by a semi-structured search for practitioner documents and media.

As mentioned, fifteen practitioners from different European DM backgrounds were involved in the research through a qualitative online survey with a mix of open-ended and closed questions asking for experienced challenges and needs, applied supporting tools, and potential improvements for evaluation. The survey and the interviews were performed using the guidelines for qualitative research of Myers (2011). The thematic analysis following Braun & Clarke (2013) revealed that there is a large number of different focus points that overlap in some places but diverge in others, underlining a heterogeneity of requirements. Common themes were aggregated to derive universal requirements. Due to the limited scope of this conference publication, the following chapter describes only foundational findings for understanding the conceptual design. Further details on theoretical grounding and requirements analysis are to be published in separate publications.

STATE OF RESEARCH AND PRACTICE

The academic interest in evaluation of DMEs is growing. However, current literature reviews state that there are several gaps in research and practice that remain to be bridged. Beerens & Tehler (2016) plead for more explicit clarity in the presentation of evaluation methodologies and their results to foster successive development of DMEs and their evaluations. Savoia *et al.* (2017) argue that actionable knowledge for improving DMEs exists in the literature, but it is short of evidence to transfer it from one system to the other. Sheikhbardsiri *et al.* (2018) identify different approaches to evaluation but state that these evaluations are always vulnerable to biases and thus require meta-evaluation, i.e., the evaluation of evaluations. Moreover, the authors could not identify a supporting tool that is universally applicable. They propose the development of a flexible toolkit.

Recent document analyses of evaluation reports find little priority on evaluation in practice. The authors call for more standardization of evaluation reports, especially considering structural principles that lead to actionable lessons learned and recommendations (Beerens, 2019; Copper *et al.*, 2020; Nordström & Johansson, 2019). Beerens (2021) concludes these findings with a design framework for more useful evaluation products by enhancing the clarity of user aspects, being an evaluation's *purpose* ("Why do we evaluate?"), *object* ("What or who is evaluated?"), *analysis* ("What happened? How did it happen? Why did it happen?"), *conclusion* ("How did the object perform?"), as well as a sophisticated *design* of the evaluation products ("How should the evaluation be presented?"). The author argues that the weakest points of evaluation products are unclear justifications of applied methods, as well as too little involvement of stakeholders to adapt to their needs.

Justification of methods is necessary due to the interdependencies of the DME and evaluation process. A DME process consists of a design and plan, execution, evaluation, and improvement planning phase (DHS, 2020). However, evaluation is not only a chronological phase but a parallel process to the DME process. It is the systematic assessment of an object's value (Thielsch & Hadzihalilovic, 2020). Therefore, the object must exist within the exercise. An evaluation-driven DME must entail objectives based on the objectives of the evaluation, oriented on the goals of the stakeholders involved (Darin-Mattsson & Hallberg, 2019). It is a crucial task of the evaluation designers to find appropriate and aligned metrics, data collection, analysis, synthesis, and tracking methods before the DME begins (van Niekerk *et al.*, 2015). A sound alignment allows for logical reconstruction and reasoning leading to greater clarity in the final evaluation product (Beerens, 2021).

Nevertheless, not only the evaluation product but also the process contributes to learning. The evaluators, e.g., observers, actors, trainers, facilitators, or even the players themselves, should have time to perform so-called double-loop learning, i.e., reflecting and modifying their assumptions and values (Olsén *et al.*, 2019; Pilemalm *et al.*, 2018). At the same time, the evaluation process should not be intrusive to the people fulfilling a task in the DME (Lapierre *et al.*, 2018). This requires well-planned coordination of roles, tasks, and schedules. A key instrument of DME coordination is the Master Scenario Event List (MSEL) – a timeline including injects that are communicated to the DME players to trigger expected actions (DHS, 2020). Based on this, evaluators are equipped with Exercise Evaluation Guidelines (EEGs) that comprise specific questionnaires or checklists for data collection to match the evaluation objectives and reduce potential bias (DHS, 2020; MSB, 2017). The collected data from the EEGs is then synthesized to evaluation products by the report creators.

Since DMEs are no closed systems, the evaluation methods must consider context to identify potential noise. DMEs rarely work exactly as planned. A recalibration of exercise and evaluation methods is often necessary to respond to contingencies (Heumüller *et al.*, 2013). The dynamic environment requires continuous monitoring of the exercise process and documentation of changes without disrupting the exercise process. Validation and evaluation should not be confounded. This illustrates the need to distinguish and consider interrelations between the exercise dimension ("Have the *right things* been expected? Have the settings been adjusted to the expectation in the *right way*?") and the players' dimension ("Have the expected actions been performed? How did the players perform them?"). Errors in the exercise dimension, e.g., a delay of injects, must be considered in the evaluation within the players' dimension (Widera *et al.*, 2018). This differentiated view is also necessary for embedding the DME in a preparedness program efficiently. An evaluation-driven DME provides actionable recommendations to improve and adapt subsequent measures to the identified gaps and requirements in the right dimensions (Beerens *et al.*, 2012; Siman-Tov *et al.*, 2020).

REQUIREMENTS AND CONCEPTUAL DESIGN

The applied requirements engineering approach begins with the identification of stakeholders following Alexander (2005). All primary stakeholders operate the system by providing inputs to the evaluation process and are thus called users in the remainder. Secondary stakeholders consume outputs of the toolkit, such as finalized static reports, but do not interact with it. Therefore, they are not described in further detail here.

The users of the toolkit are the DME evaluators, designers, and report creators, as well as the active recipients of

evaluation products. "Active" means that the role gets involved in the evaluation process, e.g., by specifying their needs for customized evaluation feedback. The users share information reciprocally within the toolkit. The roles are not mutually exclusive and not dependent on the operational role during the DME. The perspective of the identified user roles promoted the generation of respective use cases and user interfaces that illustrated how the toolkit supports the evaluation process. Based on these use cases and interfaces, requirements could be derived by combining them with the knowledge from the literature review. The validated requirements and the identified promising solutions from existing artifacts formed the input for the conceptual design of the toolkit. The design is captured in a description of necessary functions, non-functional requirements, and a corresponding object-oriented model of the logical architecture.

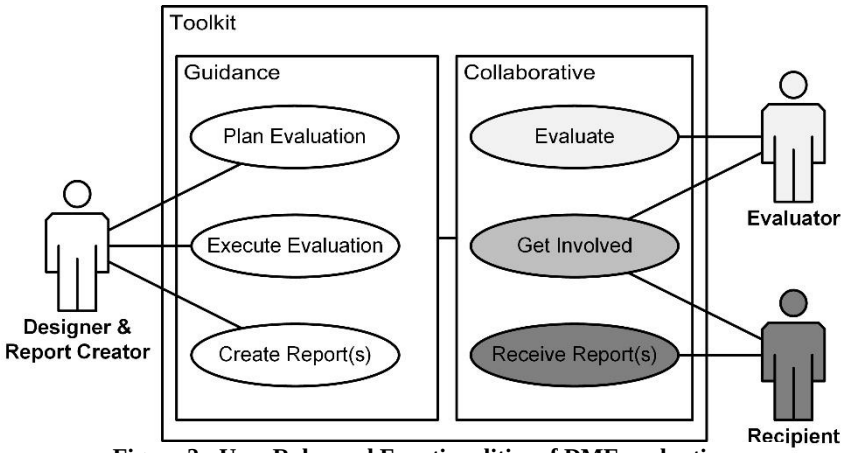

Figure 2 - User Roles and Functionalities of DME evaluation

As depicted in *Figure 2*, the toolkit offers two key functionalities: First, *evaluation guidance* for designers facilitates a more rigorous but feasible evaluation process. It includes the planning and execution of the evaluation, as well as the report creation. Second, *collaboration support* between DME and evaluation participants targets the automation of tasks, elimination of redundant work, involvement of participants in the DME and evaluation process, triangulation of observations, as well as the creation of a sustainable (cross-)organizational memory storing and sharing of data.

The *evaluation guidance* and *collaboration support* functionalities are subject to several requirements. Both must be integrated across the whole DME process – from the early planning to the final evaluation product – to facilitate compatibility between DME and evaluation, as well as consistent and traceable evaluation decisions. Frequent information exchange between roles is necessary, e.g., when filling out EEGs. Thus, the functionalities draw from the same database. The toolkit shall be a cross-organizational system so that DMEs with multiple organizations can be supported. Additionally, this enables the sharing of experiences and successive improvement of evaluation practices beyond individual organizations. Hence, the toolkit is designed as a publicly accessible web application.

User roles are established through a role-based access. Every user owns a global account where they store a history of activities and profile information. The roles for a DME are specified per DME instance. This allows reusability in varying constellations and provides a low barrier to entry. To enable transparent and easily traceable communication between users, the toolkit shall incorporate digital communication functions, such as private direct and group messages, public forums, as well as commentaries linked to specific elements. Remote communication options are important for cross-organizational DMEs where participants are located far from each other pre- and post-exercise (see also Magnusson *et al.*, 2019).

As the toolkit relies on the users' willingness to provide data, it requires their trust. Therefore, it shall incorporate anonymization and control mechanisms for user data (Evans *et al.*, 2017). There is a generally sceptical attitude toward operational change and technological investments across humanitarian practitioners (Magnusson *et al.*, 2019). The toolkit must find a balance between perceived benefit and effort for the users. Thus, it shall ensure usability and provide digital tools that enhance the efficiency of the DME and evaluation process, e.g., real-time data exchange as proposed by Lapierre *et al.* (2018). However, the toolkit shall also acknowledge the experience of practitioners to avoid alienating experts. Since each DME has unique requirements, the toolkit shall be assistive and flexible instead of commanding and static. It shall allow the integration of traditional methods, such as paper-based forms or spreadsheets, if necessary. This guarantees applicability to diverse contexts and allows for change management progressively leading to mature evaluation processes.

The toolkit shall be applicable to all discussion-based and operation-based exercises. Integrated tools, i.e., tools that use inputs and send outputs to other tools, should also work independently to make the toolkit more

universally applicable. For example, the digital sharing of EEGs can be used when not using the guidance. Designers shall be able to include or exclude a tool. A careful selection of tools based on the evaluation strategy keeps the evaluation effort low. Therefore, the structure of the toolkit must be modular. In addition, its modularity facilitates updates and expansions for future development.

The object-oriented model in *Figure 3* is a simplified representation of modules summarizing the toolkit's functions. The components are depicted in three tiers with arrows indicating the main direction of information flow. The presentation tier includes the user interface with shades defining the associated user roles. The business tier shows the internal services performing the calculations on the server side. Finally, the data tier presents the distinct types of data stored in a shared database and feeding into assorted services. The respective functions of the modules are presented in chronological order of a DME evaluation process in the following.

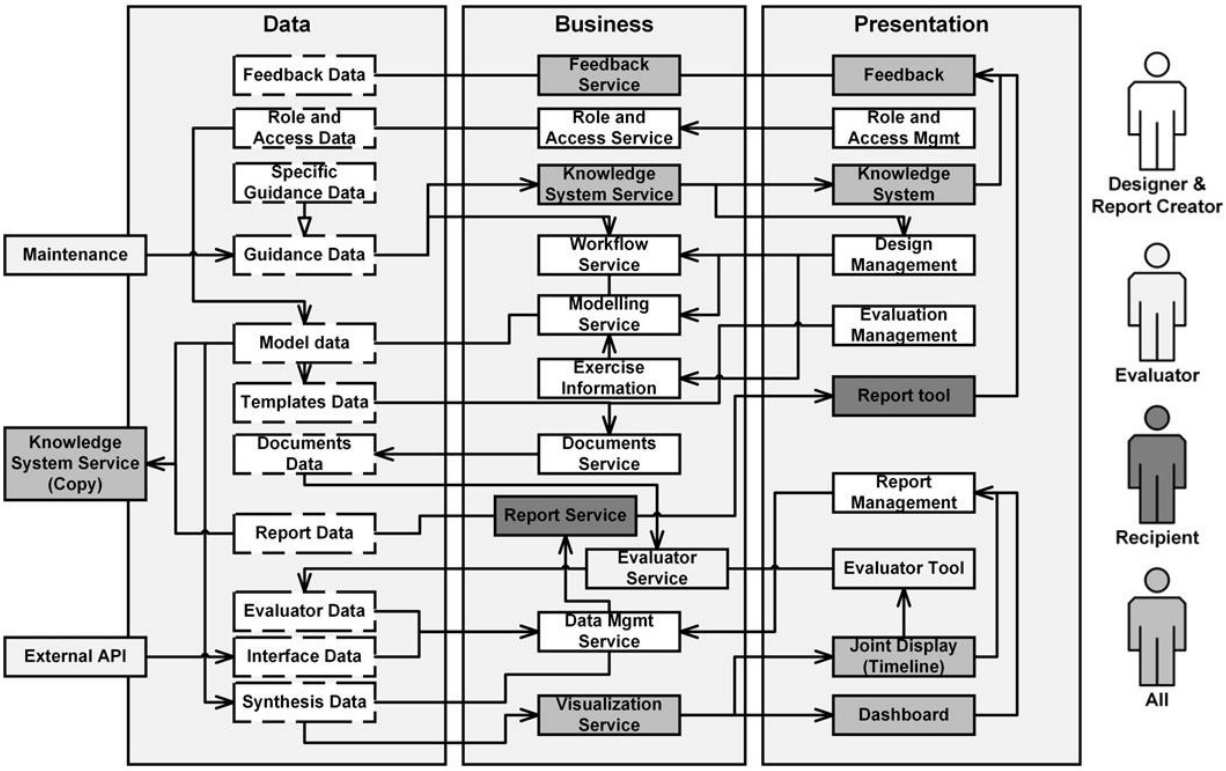

Figure 3 - Logical Architecture of the Toolkit

Before a DME, a user creates a new DME instance in the *design management* module. They are automatically the admin designer of this instance who can invite participants via email. The *role and access management* module is used to assign roles, responsibilities, and access rights to the respective accounts. The profile information may help designers to assess which responsibilities and tasks match to whom and helps to group the access rights.

Afterwards, the *workflow service* starts supported through the integrated *guidance data*. The workflow is a sequence asking for the key exercise and evaluation design decisions following the established guidance handbooks. All decisions are automatically documented. A digital review assistant suggests elements based on previous decisions, which eliminates redundant work, and checks for missing elements, ambiguities, and errors. This fosters consistency and justification of design decisions.

However, the digital assistant should not make users feel patronized. It should allow the necessary flexibility and make suggestions in a subtle way. The final decision always lies with the user. The default workflow can be replaced by proprietary solutions, which may be required for accountability reasons. As such, the service offers the versatile applicability of the handbooks while providing a simpler overview and fostering a more precise planning structure. It allows the integration of complex *specific guidance data* from research and practice. For example, if the tool recognizes through the comparison of annotations that a similar exercise specification has been developed, validated, and shared before, it can suggest its design decisions, e.g., a collection of matched indicators as proposed by Agboola *et al.* (2015), to the designers for inspiration. Moreover, the *knowledge system service* incorporates data from real incidents serving as input for realistic scenario development. The *knowledge system* can also be accessed independently from the *workflow service* as a searchable catalog, so users can browse through guidelines, reports, and examples without having to design a DME instance. Since the toolkit is used by

multiple organizations with different terminologies, it shall use natural language processing to annotate the design decisions and reports, so they can be categorized and searched more easily.

As mentioned before, it is vital to recognize the interdependencies of the exercise and evaluation process to justify decisions and make the evaluation product as well as the process retraceable. However, interdependencies can become complex. Therefore, the toolkit offers a *modelling service* that enables the visualization and simplification of the model. For example, the tool can connect scenario injects with expected actions to a specific capability. For this, the evaluation model includes the MSEL connected to the evaluation plan. The model is structured in layers, so its elements can be expanded and collapsed to foster a clear overview while also being able to display necessary details. The resulting model can be simplified to create a "dry run" testing whether the evaluation elements are complete and work as envisioned. Besides information on the evaluation, the model can entail *exercise information* which may be of interest for participants, e.g., location details, time schedules, and a respective allocation of tasks.

The *model data* is transformed into *templates data* to create information documents for the participants, e.g., documents describing the purpose and object of the exercise and evaluation, evaluation plans and links to the knowledge system, or EEGs. The toolkit uses the created model as master data to populate the different templates. Designers then tailor the documents based on the individual needs of recipients in the *evaluation management* module using the *documents service.* The use of templates means less redundant work for the designers to inform all participants adequately. This tailored approach is also used later in the report creation. Customization can be role- or group-based, e.g., observers receive more DME information than players, or it can be on an individual level, e.g., based on the location of the evaluator. The tailored *documents data* is sent to the recipient via the *evaluator service* at scheduled points in time. The digital documents can also be customized by the recipients through the creation of individual views in the *evaluator tool.*

The EEGs are filled out by the evaluators and sent back to the *data management service,* where their statements add to the *synthesis data*. The respective evaluator is linked to their provided statements to foster traceability. Since some evaluators are working in the field with limited access to the internet, the toolkit must be accessible via a mobile device, either as a dedicated app or as responsive webpage that is able to process offline information. The reduced mobile display should focus on the essential functions for the respective roles. As the EEGs are based on expected actions and should thus be matched with the MSEL, contingency injects in the MSEL require re-configurable EEGs that are updated automatically on the devices of the evaluators. The evaluators may supplement the EEGs with self-recorded media, like pictures, videos, or audio, or with additional notes to provide qualitative context. They shall also be able to provide unstructured or unplanned observations and comments on situations not covered in an EEG. Highly experienced evaluators shall have the option to create individual EEGs as evaluation template. Additionally, the toolkit shall enable self- and peer evaluation between players, cross-checking of statements between evaluators, input and linking of external statements, e.g., from interviews with players, printing of EEGs and mobile scanning of completed sheets, and scheduling of EEGs, e.g., to request contextual information before each DME episode.

The data is sent in real-time or after the exercise execution. This depends on the permissible level of intrusiveness, e.g., players should not fill out an EEG during the execution if it interferes with the performance of their operational tasks or breaks their immersion. Besides narrative *evaluator data*, the toolkit shall incorporate *interface data* via external data collection systems, e.g., for audio logs of radio calls or videos. This allows for more triangulation and offers context (Dold *et al.*, 2020). The *visualization service* integrates the provided data in a visualized localization timeline on the *joint display* based on Holdsworth & Zagorecki (2020), so the report creators can inspect and arrange the whole DME in a common operational picture, which allows better logical reconstruction (Pilemalm *et al.*, 2018). This option can be enabled for evaluators to embed their statements in the specific context or to decrease their subjective bias on a situation. The real-time communication of statements is important for the activation of contingency injects and reconfiguration of EEGs. Apart from the *joint display* of the timeline, report creators are provided with a *dashboard* that displays first automatically aggregated results and performance indicators based on the received EEGs.

The report creators use the *report management module* to generate customized evaluation products, e.g., for different player teams, as proposed by Beerens & Haverhoek-Mieremet (2021). These evaluation products can be sent to trainers, players, or other recipients via the *report service* and *tool*. First evaluation products can be displayed and discussed at hotwash debriefings to exploit learning potentials. The final report can be created after the exercise is finished, also incorporating additional input that was collected post-exercise, e.g., during the hotwash debriefings. The *report data* is added to the *knowledge system service* and can be shared. Since evaluation reports must be actionable and therefore comprehensible to recipients, the report creation should follow a standardized structure considering the user-aspects of evaluations presented by Beerens (2021). An important requirement for the reports is that the toolkit adheres to privacy regulations, e.g., the European GDPR. As mentioned, it should automatically anonymize data wherever possible and equip the user with transparent control

over their data. Digital reports allow the inclusion of alternative media, such as video or audio, which can facilitate comprehensibility (Sheikhbardsiri *et al.*, 2018). They are customizable regarding access rights but also different views. In this way, there can be multiple versions of a report, e.g., a public report and a report only available to participating or selected organizations. The reports can be commented by the users to identify and discuss potential gaps, needs, or context interpretations. Similarly, the DMEs can be evaluated remotely by the participants with a delay offering new perspectives.

Users shall be able to share evaluation designs. This allows easier collaboration and development of standards between organizations for more comparability. It is also possible to store an evaluation design for later reuse, which makes it easier to embed the DME in a preparedness program. A key advantage of the toolkit is the long-term storage of all evaluation data, which enables better collaboration between academics and practitioners with cross-fertilizing effects, e.g., researchers could train the natural language processing model and maintain the content of the *knowledge system* by adding guidelines from new studies or checking process design submissions of organizations. In exchange, they can access detailed data of shared reports to perform meta-evaluations.

VALIDATION AND DISCUSSION

The conceptual design of the toolkit was presented to four experts from different European DM backgrounds. The experts have extensive experience with DMEs in leading and advisory positions, e.g., from the *EU-MODEX* (see also Beerens *et al.*, 2012) or the *Resilience Advisors Network* (RAN, 2023). The toolkit was presented via a textual walkthrough of functions. Afterwards, we performed semi-structured, one-hour interviews focusing on relevance, usefulness, feasibility, and potential issues of the toolkit.

All interviewees considered the design relevant, useful, and feasible. It was highlighted that the major added value of the conceptual design is its well-founded analysis of components that results in a valuable combination and integration of functions. Although the separate functions are mostly not novel by themselves, the respective systems in practice do not deliver an integrated approach that fully utilizes the combination of functions, e.g., by automatically linking injects from the MSEL with evaluator statements. Current solutions had mostly been created ad-hoc to close single gaps, e.g., with Google Spreadsheet templates for EEG management. Even the Exercise Control Tool (EU-MODEX, 2022), which is considered one of the most mature tools in practice, does not yet offer satisfactorily integrated evaluation functions. The resulting redundant work occupies resources that could be utilized elsewhere. It was therefore concluded that the conceptual design provides a sound basis for further development that should be assessed early in the field.

The toolkit's functions were estimated to work on local levels, as well as on full-scale DMEs. It was argued that the system might be too elaborate for small discussion-based DMEs if the designers are experienced. This stresses the need for modular applicability of functions, as well as for an assistive, non-binding functionality. In contrast, it was again confirmed that especially small teams without senior evaluation experts often conduct exercises without systematic evaluation. Thus, the toolkit can be a helpful reminder for integrated planning with low resource investments.

One feature that was particularly well received was the ease of reusing elements of previous DMEs and evaluations to embed them in programs, e.g., by copying the MSEL, objectives, and data collection techniques to foster comparability, even though it is not possible to repeat a DME exactly. The sharing of DME and evaluation designs was appreciated for inspirations. The essential assumption that DME and evaluation processes cannot be separated was confirmed. Often, attempts are made to involve external evaluators who were not involved in the planning of the DME, which usually fails. The toolkit allows for better remote planning of the DME and evaluation by facilitating involvement in planning, e.g., by using the modelling tool.

Nevertheless, it was noted that practitioners might expect the toolkit to offer additional exercise design and control functions, e.g., venue management and logistics coordination, which were not in focus of this research. Otherwise, they have to use multiple systems, which might cause redundant work if no easily deployable interface is available and hampers acceptance. Another critique was made about the reporting function of the toolkit. Complete standardization of reports is not considered feasible because of the socio-technical dynamics of DMEs. Usually, many things go differently than planned, whether in players' actions or exercise organization. Managing the data appropriately during the DME execution is beyond most capacities. Hence, the master data is not updated after the exercise. This was already anticipated in the design, as it allows customization, re-configuration, or export of the report design to spreadsheets. However, some reports might require so much re-configuration that the re-imported reports are not comparable anymore without tremendous effort. This issue should be given more focus to find a way to easily incorporate changes without adding to the workload of the busy staff.

Additionally, the significance of the hotwash debriefing function was emphasized, i.e., some evaluation results shall be accessible directly after the exercise to reflect and discuss. This is a further reason why the re-

configurability of necessary results must be simple; otherwise, only flexible but less integrated tools like spreadsheets or paper will be used. Likewise, the distinction between local and system evaluation was mentioned. While the system perspective is essential, the individual perspective for immediate feedback should not be neglected. The needs of the players should be in the foreground, followed by the system evaluation. The toolkit must enable small and simple cycles of local evaluation, e.g., trainers giving feedback to trainees on their team performance, to be carried out within the overall system evaluation process.

Furthermore, evaluators often switch their mode during the DME, e.g., between training and passive observation. The toolkit should therefore consider not only the given sub-roles of the evaluator in their statements but also their current mode. The joint display tool appears more useful for contextualization in analysis, especially for report creators who have not observed the scene, than for the input process by the evaluators. It could overwhelm untrained evaluators as they have no familiar analog equivalent for it. Thus, it would require a high maturity of the toolkit and of the evaluator training, which calls for appropriate change management.

Finally, it is important to consider the toolkit as one of many means to an end. It enables rather than solves. The toolkit offers the technical infrastructure for improving DME and evaluation design but is only useful if it is applied in sound processes. For example, the organizational structure of the exercise organization team in terms of how they make decisions must be efficient for the toolkit to be efficient. Moreover, the success of the toolkit depends on the quality of content in the knowledge base, e.g., the quality of the guidelines, on the usability of the implementation, e.g., the user experience of the modelling tool, as well as on future development of disaster management evaluation in general, e.g., theories on how learning can be measured or how performance in real disasters can be attributed to DMEs. Therefore, the toolkit does not produce evidence by itself but paves the way in the right direction.

CONCLUSION

In this paper, we presented and discussed the conceptual design for a toolkit supporting evaluation of disaster management exercises. Our findings show that exercises must become more evaluation-driven to exploit their learning potential. Only rigorous evaluation can generate evaluation products that are perceived as useful and actionable by practitioners, which again can enhance their willingness to invest into evaluation. The toolkit provides the infrastructure for a resource-efficient, integrated, and collaborative evaluation process between designers, evaluators, report creators, and recipients of evaluation products.

For future research, we propose to use the presented conceptual design and discussion as a blueprint for implementing a modular prototype and evaluating it with practitioners in the field to refine and extend it successively. Moreover, additional research on evaluation in disaster management is necessary in general, e.g., considering the validation of evaluation methods or the long-term impact of preparedness programs. An implementation of the proposed toolkit offers new research opportunities by offering a collaborative platform for disseminating specialized tools and methods among practitioners, as well as by facilitating access to comprehensive practitioner data from exercises and educational programs.

REFERENCES

Agboola, F., Bernard, D., Savoia, E. & Biddinger, P. D. (2015) Development of an Online Toolkit for Measuring Performance in Health Emergency Response Exercises. *Prehospital and disaster medicine*, **30** (5), 503–508.

AIDR (2012) *Australian Disaster Resilience Handbook 3: Managing Exercises* [WWW document]. URL https://knowledge.aidr.org.au/media/3547/handbook-3-managing-exercises.pdf, accessed 30 June 2022.

Alexander, I. F. (2005) A Taxonomy of Stakeholders. *International Journal of Technology and Human Interaction,* **1** (1), 23–59.

Baroutsi, N. (2018) A practitioners guide for C2 evaluations: Quantitative measurements of performance and effectiveness. In: *Proceedings of the 15th ISCRAM Conference*, pp. 170–189.

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. *Int. J. Emergency Management*, **15** (3), 221–254.

Beerens, R. J. J. (2021) *Improving disaster response evaluations: Supporting advances in disaster risk management through the enhancement of response evaluation usefulness*. Dissertation, Lund, Sweden.

Beerens, R. J. J., Abraham, P. & Braakhekke, E. (2012) *Maximise your returns in Crisis Management preparedness: A Cyclic Approach to training and exercises*. Unpublished.

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.

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.

Berlin, J. M. & Carlström, E. D. (2014) Collaboration Exercises—The Lack of Collaborative Benefits. *International Journal of Disaster Risk Science*, **5** (3), 192–205.

Borodzicz, E. & van Haperen, K. (2002) Individual and Group Learning in Crisis Simulations. *Journal of Contingencies and Crisis Management*, **10** (3), 139–147.

Braun, R., Benedict, M., Wendler, H. & Esswein, W. (2015) Proposal for Requirements Driven Design Science Research. In: *New Horizons in Design Science: Broadening the Research Agenda*. Donnellan, B., Helfert, M., Kenneally, J., VanderMeer, D., Rothenberger, M., Winter, R. (eds.), pp. 135–151. Springer International Publishing, Cham.

Braun, V. & Clarke, V. (2013) *Successful qualitative research: A practical guide for beginners*. sage.

Browne, K. E., O'Connell, C. & Yoder, L. M. (2018) Journey Through the Groan Zone with Academics and Practitioners: Bridging Conflict and Difference to Strengthen Disaster Risk Reduction and Recovery Work. *International Journal of Disaster Risk Science*, **9** (3), 421–428.

Comes, T. (2016) Cognitive biases in humanitarian sensemaking and decision-making lessons from field research. In: *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pp. 56–62. IEEE.

Copper, F. A., Mayigane, L. N., Pei, Y., Charles, D., Nguyen, T. N., Vente, C., Chiu de Vázquez, C., Bell, A., Njenge, H. K., Kandel, N., Ho, Z. J. M., Omaar, A., La Rocque, S. de & Chungong, S. (2020) Simulation exercises and after action reviews - analysis of outputs during 2016-2019 to strengthen global health emergency preparedness and response. *Globalization and health*, **16** (1), 115.

CRED (2016) *The human cost of natural disasters: A global perspective* [WWW document]. URL http://repo.floodalliance.net/jspui/handle/44111/1165, accessed 19 November 2021.

Darin-Mattsson, K. K. & Hallberg, N. (2019) Do's and Don'ts in Inter-Organizational Crisis Management Exercises. In: *Proceedings of the 16th ISCRAM Conference*.

DHS (2020) *Homeland Security Exercise and Evaluation Program (HSEEP)* [WWW document]. URL https://www.fema.gov/sites/default/files/2020-04/Homeland-Security-Exercise-and-Evaluation-Program-Doctrine-2020-Revision-2-2-25.pdf, accessed 30 June 2022.

Dold, C., Munschauer, C. & Mudimu, O. A. (2020) Real-Life Exercises as a Tool in Security Research and Civil Protection-Options for Data Collections. In: *Proceedings of the 17th ISCRAM Conference*.

EU-MODEX (2022) *Exercise Control Tool* [WWW document]. URL https://10years.eu-modex.eu/future-of-eu-modex/future/new-exercise-control-tool, accessed 17 July 2022.

Evans, A. B., Hulme, J. M., Nugus, P., Cranmer, H. H., Coutu, M. & Johnson, K. (2017) An Electronic Competency-Based Evaluation Tool for Assessing Humanitarian Competencies in a Simulated Exercise. *Prehospital and disaster medicine*, **32** (3), 253–260.

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.

Grunnan, T. & Fridheim, H. (2017) Planning and conducting crisis management exercises for decision-making: the do's and don'ts. *EURO Journal on Decision Processes*, **5** (1-4), 79–95.

Heumüller, E., Richter, S. & Lechner, U. (2013) Training, test and experimentation: A classification of command post exercises. In: *Proceedings of the 10th ISCRAM Conference*.

Hevner, A. R. & Chatterjee, S. (2010) Design Science Research in Information Systems. In: *Design Research in Information Systems*. Hevner, A. R., Chatterjee, S. (eds.), pp. 9–22. Springer US, Boston, MA.

Holdsworth, D. & Zagorecki, A. (2020) The SERIES model: development of a practitioner focused emergency response evaluation system. *International Journal of Emergency Services*, **9** (3), 313–337.

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.

Lapierre, D., Tena-Chollet, F., Tixier, J., Bony-Dandrieux, A. & Weiss, K. (2018) How Can We Evaluate the Participants of a Crisis Management Training Exercise? In: *Decision-making in Crisis Situations*. Sauvagnargues, S. (ed.), pp. 103–124. John Wiley & Sons, Inc, Hoboken, NJ, USA.

Magnusson, M., Pettersson, J. S., Bellström, P. & Andersson, H. (2019) Developing Crisis Training Software for Local Governments—From User Needs to Generic Requirements. In: *Advances in Information Systems Development*. Andersson, B., Johansson, B., Barry, C., Lang, M., Linger, H., Schneider, C. (eds.), pp. 79–96. Springer International Publishing, Cham.

Miller, A. N., Sellnow, T., Neuberger, L., Todd, A., Freihaut, R., Noyes, J., Allen, T., Alexander, N., Vanderford, M. & Gamhewage, G. (2017) A Systematic Review of Literature on Effectiveness of Training in Emergency Risk Communication. *Journal of health communication*, **22** (7), 612–629.

MSB (2011) *Handbook evaluation of exercises* [WWW document]. URL https://www.msb.se/siteassets/dokument/publikationer/english-publications/evaluation-of-exercises.pdf,

accessed 30 June 2022.
MSB (2017) *Exercise Guidance: Method Booklet – Exercise Evaluation* [WWW document]. URL https://www.msb.se/siteassets/dokument/publikationer/english-publications/exercise-guidance-method-booklet--exercise-evaluation.pdf, accessed 30 June 2022.
Myers, M. D. (2011) *Qualitative Research in Business & Management*. SAGE Publications Ltd, London.
Nordström, J. & Johansson, B. (2019) Inter-Organisational Learning-A Review of Knowledge Sharing in Post-Exercise Reports. In: *Proceedings of the 16th ISCRAM Conference*.
Olsén, M., Hallberg, N. & Mattsson, K. D. (2019) Who Learns from Crisis Management Exercises: An Explorative Study. In: *Proceedings of the 16th ISCRAM Conference*.
Pilemalm, S., Radianti, J., Munkvold, B. E., Majchrzak, T. A. & Steen-Tveit, K. (2018) Turning Common Operational Picture Data into Double-loop Learning from Crises – can Vision Meet Reality? In: *Proceedings of the 15th ISCRAM Conference*, pp. 417–430.
RAN (2023) *Resilience Advisors Network* [WWW document]. URL https://www.resilienceadvisors.eu/index.html, accessed 28 February 2023.
Savoia, E., Lin, L., Bernard, D., Klein, N., James, L. P. & Guicciardi, S. (2017) Public Health System Research in Public Health Emergency Preparedness in the United States (2009-2015): Actionable Knowledge Base. *American journal of public health*, **107** (S2), e1-e6.
Seijts, G. H. & Latham, G. P. (2005) Learning versus performance goals: When should each be used? *Academy of Management Perspectives*, **19** (1), 124–131.
SEMC (2021) *Western Australia Managing Exercises Guideline* [WWW document]. URL https://semc.wa.gov.au/capability-and-preparedness/exercising/Pages/Templates-and-Resources.aspx, accessed 30 June 2022.
Sheikhbardsiri, H., Yarmohammadian, M. H., Khankeh, H. R., Nekoei-Moghadam, M. & Raeisi, A. R. (2018) Meta-evaluation of published studies on evaluation of health disaster preparedness exercises through a systematic review. *Journal of education and health promotion*, **7**, 15.
Siman-Tov, M., Davidson, B. & Adini, B. (2020) Maintaining Preparedness to Severe Though Infrequent Threats-Can It Be Done? *International journal of environmental research and public health*, **17** (7).
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.
Sommer, M., Braut, G. S. & Njå, O. (2013) A model for learning in emergency response work. *International Journal of Emergency Management*, **9** (2), 151.
Thielsch, M. T. & Hadzihalilovic, D. (2020) Evaluation of Fire Service Command Unit Trainings. *International Journal of Disaster Risk Science*, **11** (3), 300–315.
van Niekerk, D., Coetzee, C., Botha, D., Murphree, M. J., Fourie, K., Le Roux, T., Wentink, G., Kruger, L., Shoroma, L., Genade, K., Meyer, S. & Annandale, E. (2015) Planning and Executing Scenario Based Simulation Exercises: Methodological Lessons. *Journal of Homeland Security and Emergency Management*, **12** (1).
Verheul, M. L., La Dückers, M., Visser, B. B., Beerens, R. J. J. & Bierens, J. J. (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.
vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R. & Cleven, A. (2015) Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, **37**.
Widera, A., Fonio, C., Lechtenberg, S. & Hellingrath, B. (2018) Measuring Innovations in Crisis Management. In: *Proceedings of the 15th ISCRAM Conference*.
Wieringa, R. J. (2014) *Design Science Methodology for Information Systems and Software Engineering*, 1st ed. 2014. Springer Berlin Heidelberg; Imprint: Springer, Berlin Heidelberg.
Williams, J., Nocera, M. & Casteel, C. (2008) The effectiveness of disaster training for health care workers: a systematic review. *Annals of emergency medicine*, **52** (3), 211-22, 222.e1-2.