A Practitioners Guide for C2 Evaluations
Quantitative Measurements of Performance and Effectiveness

Nicoletta Baroutsi
Swedish Defense University, Lund's University
nicoletta.baroutsi@fhs.se

ABSTRACT

Quantitative evaluations are valuable in the strive for improvements and asserting quality. However, the field of Command & Control (C2) evaluations are hard to navigate, and it is difficult to find the correct measurement for a specific situation.

A comprehensive Scoping Study was made concerning measurements of C2 performance and effectiveness. A lack of an existing appropriate framework for discussing C2 evaluations led to the development of the Crisis Response Management (CRM) Matrix. This is a new analysis tool that assigns measurements into categories, and each category display unique strengths, weaknesses and trends. The analysis yielded results proving to be too rich for a single article, thusly, this is the first of two articles covering the results.

In this article, the Practitioners Guide focus on results valuable for someone interested in evaluating C2. Each evaluation has specific requirements to be met, for example, whether it is a real response or an exercise, or what competencies the evaluator has. For best result, these requirements ought to be reflected in the chosen measurement.

Keywords

Performance, Effectiveness, Evaluation, Crisis Response Management, Command & Control, Quantitative Measurements

INTRODUCTION

Feedback is key for increasing performance, it enables the receiver to effectively learn from their experiences. Studies have even shown that fields with diffuse or inexistent feedback results in similar performance for novices and experts (Schanteau, 1992). A crisis response is a typical example of a domain with diffuse and inexistent feedback because of the multitude of factors contributing to the effects in the environment. An array of organizations is involved, some collaborating and some not, and the environment itself is even changing as a consequence of time. For teams to understand the consequences of their own actions is, to say the least, problematic. The purpose of the Command & Control (C2) unit is to create direction and coordination for response (Brehmer 2007). They make the strategic decisions to create long term solutions for organizations, personnel and resources. In the absence of naturally occurring feedback, evaluations can serve as a tool to support the C2 unit in their quest to improve performance and effectiveness.

Evaluation is a systematic determination of the quality or value of something. Affirming worth and striving for progress is important in all parts of a crisis response system, and quantitative evaluations serves as a vital compliment to the qualitative. Quantitative evaluations force the evaluator to repeatedly and systematically acknowledge the same aspects, a value not to be underestimated. Just imagine the difficulty in winning a video game that change the scoring systems in each round. Shifting goals is a natural part of C2, but to expect someone to become skilled within a domain while always changing the criteria for what is considered 'good' is unreasonable. A quantitative evaluation is in this article defined as an evaluation offering the result in numbers or on a scale; this includes everything from raw numbers and percentages to more advanced statistical calculations, such as coherence and correlations.

To identify suitable measurements for conducting quantitative evaluations can be a battle. The field is scarce, and it requires a lot of time and effort before realizing whether a measurement is meeting your needs. In this article are the results of a comprehensive Scoping Study, which collects measurements available today for evaluating C2 performance and effectiveness. The measurements are analyzed through a newly developed analysis tool called the CRM Matrix. The final result is the Practitioners Guide for C2 Evaluations, presented in the end of this article. In this guide measurements are assigned into categories, and each category display unique strengths, weaknesses and trends. This presentation allows for easier access in the search for measurements that can meet specific requirements when evaluating C2.

Performance vs. Effectiveness

Performance and effectiveness are two different but related concepts: effectiveness is related to the accomplishment of the set goals, while performance is related to the team's or organizations capacities and processes (Essens, Vogelaar, Mylle, Blendell, Paris, Halpin & Baranski, 2005). A fictive scenario will describe these differences, and how the different concepts relate to each other. Imagine a wartime scenario, two fighting parties are both occupying an area, but important cargo needs to be delivered on route crossing the unsafe region.

- **Team A** initiates by contacting the conflicting parties who both assures that no attacks will take place on the given route for the time of the transportation. Both the cargo and the driver arrive safely at the goal destination. In this scenario the team accomplished the task effectively (the cargo arrived) and showed good performance (how they handled the task).
- **Team B** sends the cargo without taking any precautionary, but the cargo still arrives safely. This team reaches high effectiveness, but the performance was not satisfactory since they did not perform any safety measures.
- **Team C** completes all precautionary measures, however, there is a shooting and the cargo gets hit and the driver is injured. In this situation the team did display good performance since they did everything they could, but the effectiveness was still low since the cargo never arrived.

Two conclusions can be derived from this scenario: Effectiveness is more influenced by external factors than performance, and high performance can increase the probability of higher effectiveness but never guarantee it.

THE CRM MATRIX

The CRM Matrix categorizes measurements based on 2 distinguishing characteristics; the measurements Origin and its Focus (see Figure 1). The Origin dimension represent the logical foundations the measurement is built upon, and it consists of two classes: Expert Knowledge and Theoretical Knowledge. The Focus dimension represent the analytical perspective used for choosing what data to collect, and it consists of three classes: Process, Capability, and Macro-Cognition. The combination of these classes creates 6 categories that inherit their traits from the classes: CEM, CTM, PEM, PTM, MEM and MTM.

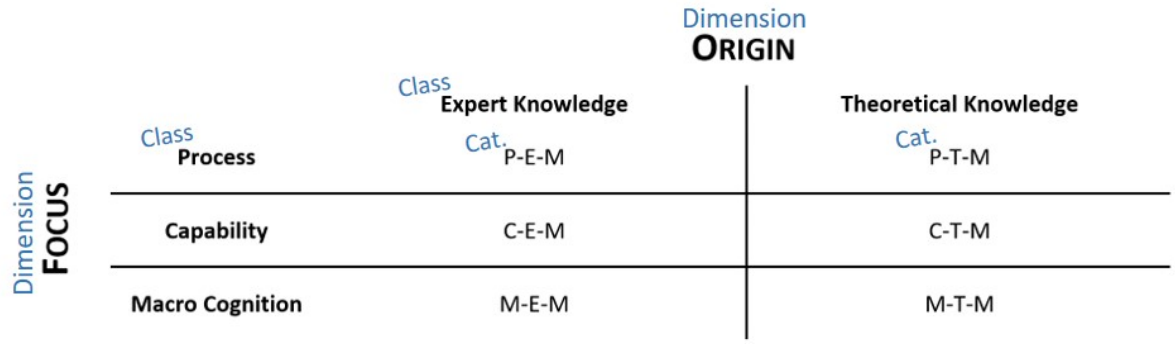

Figure 1. A visual representation of the CRM Matrix.

Each category is named based on an abbreviation of the classes it inherits traits from, then ending with the letter "M" for "Measurement". Hence, if a measurement is focused on Capability and has its origin in Expert Knowledge the abbreviation would be CEM, i.e. Capability – Expert Knowledge – Measurement (see Figure 2).

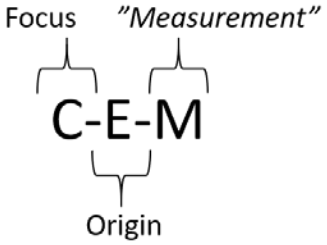

Figure 2: The name of each category is an abbreviation of the classes it belongs to.

The Origin Dimension

The Origin is the logical foundations the measurement is built upon, i.e. the support for why something is considered good versus bad in a specific context. The distinction between them are simple; methods based on Expert Knowledge do not present any theory or model related to the measurement while the methods belonging to Theoretical Knowledge do. One might ask whether the Expert class could be considered a poor scientific standard since it is difficult for a third party to assess the quality of the measurements foundation. However, systematic validity and reliability testing should be seen as a natural part of any form of measurement, and the absence or presence of these controls should instead be the main concern.

Expert (& Doctrinal) Knowledge

The criteria for the measurement is derived from expert opinions or authoritative doctrines, i.e. it builds upon implicit knowledge and experience. A doctrine is a document specifying procedures and rules valid for a certain organization, these are typically written by experts within the field and is therefore suited to be framed within this class as well.

A stereotypical measurement belonging to this class are the indicators from the medical domain (Gryth, Radestad, Nilsson, Nerf, Svensson, Castren, Riiter, 2010; Green, Modi, Lunney & Thomas 2003; Djalali, Carenzo, Ragazzoni, Azzaretto, Petrino, Corte & Ingrassia, 2014):

> An indicator is an observable variable that can be either absent or present, and the presence of the indicator is considered as valuable. These indicators are derived from experts and/or doctrines, however, no theories are presented to explain their relevance.

Theoretical (& Model) knowledge

For a measurement to belong to this class, an underlying theory or model needs to be presented that logically supports the relevance of the measurement. These measurements carry similarities to what Hayes (2012) calls Empirical Measurements; they require substantial planning and investments, and many organizations are unwilling to support them. However, a valid measurement based on Expert Knowledge would require the same efforts.

An example of a stereotypical measurement belonging to this class is the Quantitative Analysis of Situational Awareness (QUASA):

> QUASA is a method developed from the theory of Situation Awareness (SA). This method involves asking the participant about the ongoing situation and the given answer can be correct acceptance, incorrect acceptance, correct rejection or incorrect rejection. (McGuinnes, 2004).

The Focus Dimension

The Focus dimension concerns the choice of data; by choosing what data to collect a statement is made regarding what is important when handling a crisis. Each measurement carries specific assumption regarding its relation to actual performance and effectiveness, assumptions that can only be properly rejected or accepted through statistical testing.

Three classes are identified: Capability, Process and Macro-Cognition. Capability can be described as a more direct measurement, it evaluates the perceived performance or effectiveness. Process and Macro-Cognition can instead be understood as indirect, they measure aspects that influence the performance or effectiveness (see Figure 3).

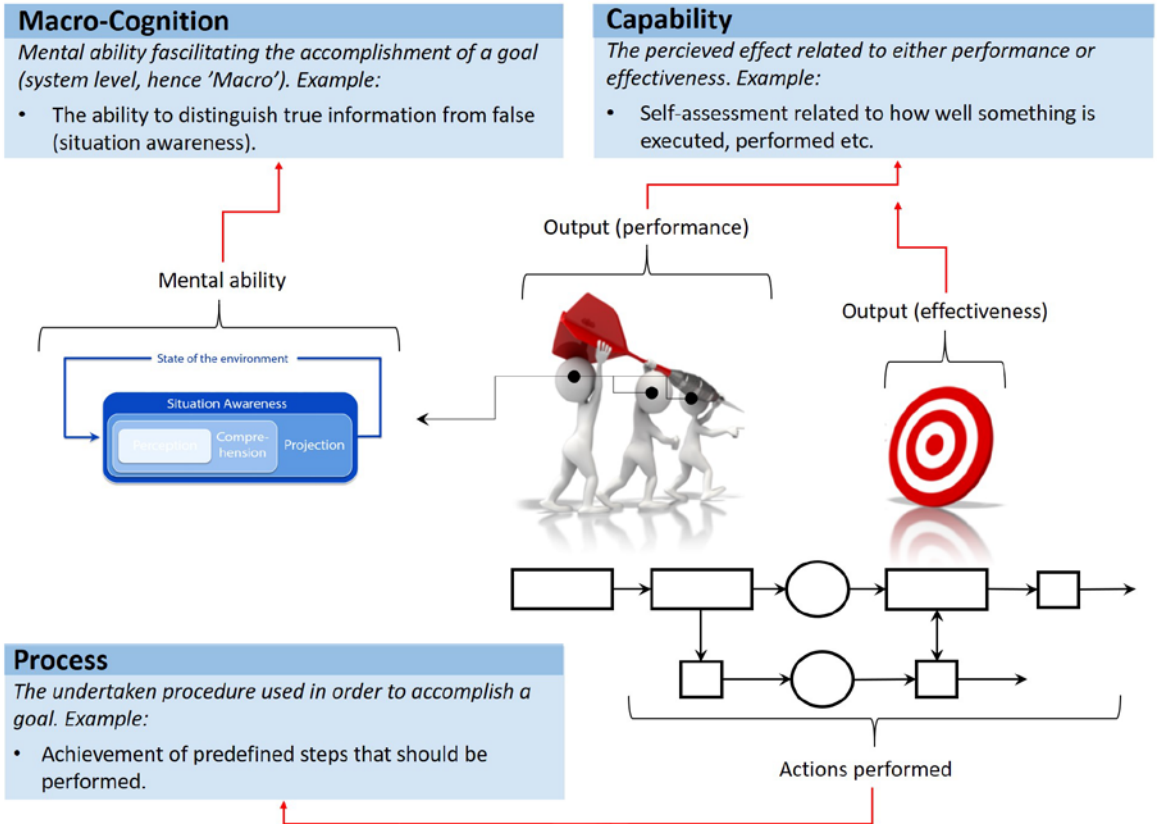

Figure 3. A visual representation of the three classes in the Focus Dimension.

Capability

Capability measurements focus on the perceived effect of either performance or effectiveness, it is important to note that *it is the perceived and not actual effect/performance* since today's measurements are based on subjective opinions. The implicit assumption is that the perceived effect is equal or similar to the actual performance or effectiveness. However, there are many influencing variables that can disrupt this correlation.

A stereotypical measurement is the Command Team Effectiveness (CTEF) questionnaire (Essens, Vogelaar, Mylle, Baranski, Goodwin, Van Buskirk, Berggren, Hof, 2010):

> CTEF is an extensive self-assessment questionnaire concerning conditions, processes, outcomes and feedback factors that influence effective teamwork. The questionnaires ask the participant to grade factors on 6 graded Likert scales.

Process

Process measurements focus on the procedure used to accomplish set goals. The implicit assumption is that a specific work processes or patterns correlate with high performance and effectiveness, and that the desirable process can be predetermined.

A typical measurement is, again, the medical indicators (Gryth, et al., 2010; Green, et al., 2003; Djalali, et al., 2014):

> An indicator is an observable behavior that can be either absent or present, and the presence is considered as valuable. Each observable indicator is part of a standardized procedure that is accepted as the correct one. For example:

Indicator	**Standard (time frame in min)**
Declaring major incident	1
Deciding level of preparedness	3
Decision on additional resources on scene	3

Macro-Cognition

Macro-Cognitive measurements focus on the mental ability facilitating the accomplishment of a goal. Macrocognition concerns the way we think in complex situations (Klein, Ross, Moon, Klein, Hoffman & Hollnagel, 2003). 'Macro' puts the emphasis on the real world setting where multiple individuals collaborate and makes use of tools and their surroundings. For example, if a team's or organizations SA is well in tuned with reality, then they will perform better and the probability of a better outcome increases. The implicit assumption is that system performance and effectiveness is achieved through the fulfillment of higher macro-cognitive functions. Measurements belonging to this class tend to go beyond raw data, such as hits, errors and accuracy, and instead makes use of increases/decreases, changes in range or variance etc.

The Shared Priorities Instrument (SPI) is a classic Macro-Cognitive measurement (Berggren, 2016):

> Shared Priorities assumes teams to be dependent on shared understanding to achieve their joint goal. Each team member generates a list of items and rank the items in order of importance, the items on each list are then scrambled and participants rank order each other's lists. Shared understanding can then be calculated by comparing the rankings.

METHOD

The method contains two individual tracks under the headings '*The Scoping Study*' and '*The Development of The CRM Matrix*'. Articles that are identified within the Scoping Study are analyzed using the CRM Matrix. The analysis yields extensive results, making it more appropriate to divide the results based on the intended reader. In this article are the results intended for practical usage, i.e. for someone planning to evaluate C2. The results are presented under '*THE PRACTITIONERS GUIDE FOR C2 EVALUATIONS*'. Complementing results are planned for a future article that focus on future method development and implications for the scientific community.

The Scoping Study

The Scoping Study was conducted 2015-17, following the 6-step framework proposed by Arksey & O'Malley (2005). The steps are revisited iteratively, as also proposed by the authors. The written presentation follows the logical order from the point of view of the 6-step framework, while the chronological order can be found in Figure 4.

Step 1: Identifying the research question

The study used a broad research question to allow range and coverage: *How can CRM performance/effectiveness be evaluated according to the scientific literature?*

Step 2: Identifying relevant studies

To satisfy the purpose of maximizing range and comprehensiveness, multiple search strategies were used to identify relevant literature:

Scopus database

Scopus a broad database featuring peer-reviewed research in the fields of social sciences, medicine, technology, arts and humanities. Key concepts were identified within the research question and complemented with synonyms (see Table 1). All possible combinations were investigated in the database using Boolean search strings in the form of (TITLE-ABS-KEY (Variable A) AND (Variable B) AND (Variable C)). Exclusion was based on (1) the number of relevant hits, (2) if they produced duplicates of other searches, or (3) if they generated an unmanageable number of articles (i.e. the term C2 generated too many hits to be practically feasible combined with few relevant hits). If the initial hits included multiple irrelevant references, then intervening subject areas got excluded (i.e. (EXCLUDE (SUBJAREA, "MATH"))). This lead up to the final search string presented in Figure 4.

Table 1: Complete list of keywords used within the database search in Scopus.

Variable A: Command and Control
Included synonyms: Coordination, Collaboration, Response Management
Excluded synonyms: Cooperation, Management, C2
Variable A: Evaluation
Included synonyms: Analysis, Measur*, Effect*, Assessment, Performance
Excluded synonyms: Audit
Variable C: Crisis
Included synonyms: Disaster, Accident, Catastrophe
Excluded synonyms: Emergency, Accident

Proceedings of ICCRTS and the C2 journal

Preliminary results from the database search shed light on an absence of relevant references, particularly from the military domain (*see Step 6: Consultation*). ICCRTS and the C2 journal have a military perspective and were believed to cover the gap identified during the consultation. Relevant tracks in ICCRTS and the complete C2 journal were hand searched from 1999-2016. Additional searches included the Information Systems for Crisis Response and Management (ISCRAM) symposium, but these were omitted based on a shortage of relevant articles.

Existing networks

This proved to be an important complement, offering insights to research not found by neither of the other search strategies. Researchers affiliated with the following networks were advised: Swedish National Defense Research Agency (FOI), Swedish National Defense University (FHS), Lund's University (LU), Center for Teaching and Research in Disaster medicine (KMC) and the North Atlantic Treaty Organization (NATO).

Step 3: Study Selection

Because of the comprehensive approach in the searches and research question, numerous irrelevant studies were included into the results. The following inclusion criteria allowed for systematic exclusion of irrelevant studies:

1. Full text is available in English or Swedish.
2. The measurements must be quantitative, i.e. offering the result in numbers or on a scale.
3. The method should be applicable either during an exercise or real event.

It is worth noticing that articles and reports *did not have to be peer reviewed* to be included. To include only peer-reviewed articles would eliminate many relevant references.

Step 4: Charting the data

Each reference was read to identify predefined characteristics, these could either have a set of values to choose from or require an open answer (see Table 2). The references are then divided according to the categories in the CRM Matrix.

Table 2: A list of all characteristics used when charting the data. The open answers in this table are only examples.

Characteristics	Value	Description
Domain	Open answer: *military, medical etc.*	A domain (profession) within CRM
Data collection	Open answer: *observers, questionnaire etc.*	How the data is collected
Data type	Open answer: *indicators, Likert-scale etc.*	Format of the collected data
Analysis	Open answer: *Summarized scores, means, ratios etc.*	How the data is analyzed
Applicability	Real event, Exercise, Both	Circumstances in which the method is applicable
Baseline	Open answer: *Specified not explained, surrogate baseline, etc.*	A benchmark of a certain value used to mark levels of performance
Generalizability	Yes/ No	Whether the method is generalizable across domains
Validation	*Open answer:* Factor Analysis, correlation to performance etc.	If and how the method has been validated
Peer reviewed	Yes/ No	Whether the literature is peer reviewed

Step 5: Collating, summarizing and reporting the results

A total of 253 articles provided the basis for the analysis; 109 articles derived from the database search, 60 from hand searching journals and conference proceedings, and 84 from consulting individuals affiliated with existing networks. In the end 56 articles were included in the analysis, however, this article do not include the crossovers making it a total of 42 references.

Step 6 (optional step): Consultation

A work in progress paper was discussed at a seminar attended by multiple scientists with PhD and professor's titles, all active within the field of CRM research. The paper covered the results from the Scopus searches with merely 12 articles included in the analysis. This proved to be rewarding. Multiple relevant references could not be found among the articles, especially from the military domain. Hence, new search strategies had to be included (see *Step 2: Identifying relevant* studies).

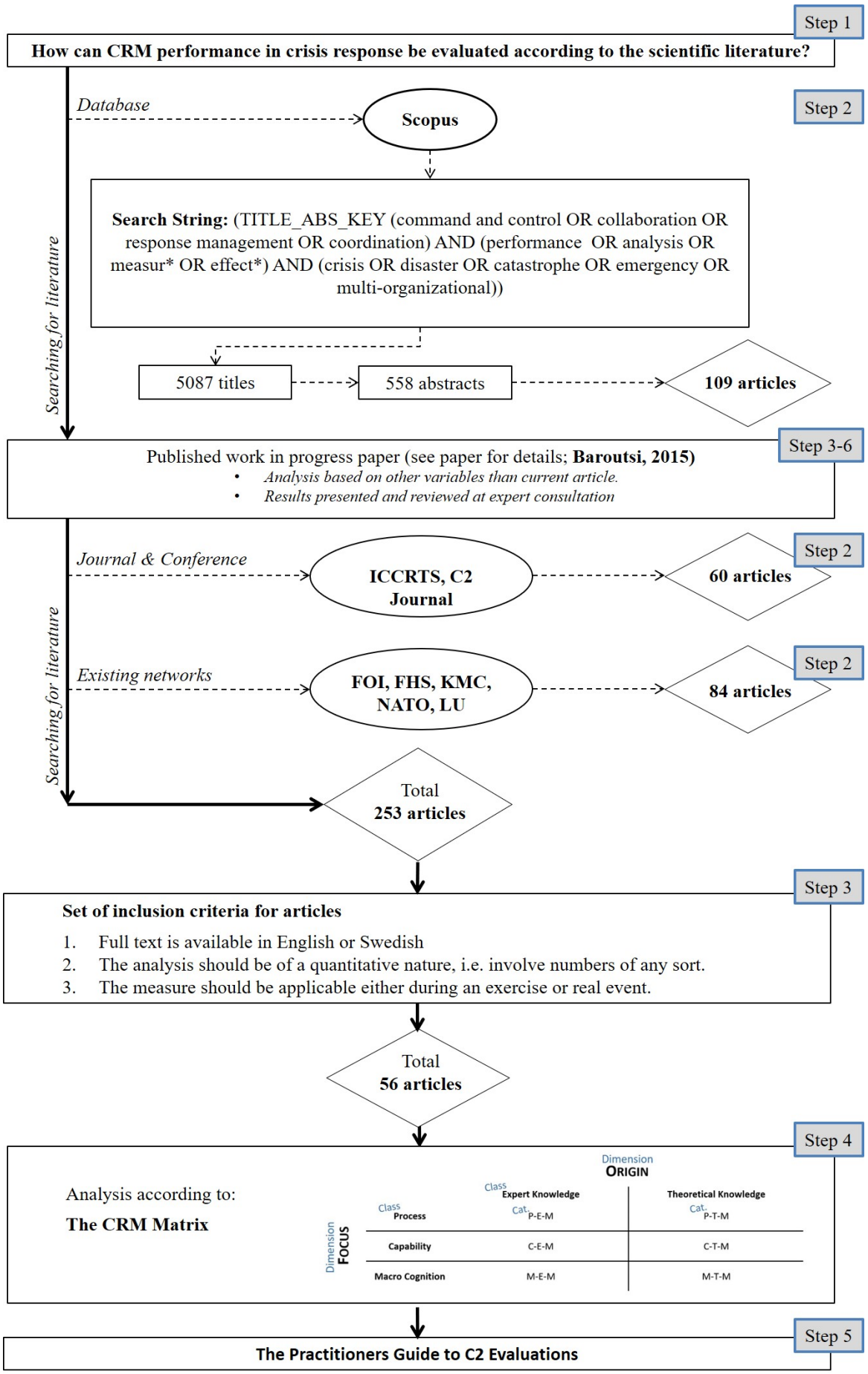

Figure 4: The method described in a chronological order.

The Development of The CRM Matrix

The initial analysis was based on Hayes (2012) classic approaches to evaluate C2, he divided measurements into Doctrinally Driven Assessments (DDA), Process Oriented Measures (POM), and Empirical Measurements (EM). However, categorizing according to these approaches proved problematic since: (1) *They are not mutually exclusive.* DDA and EM are both defined by their origin, DDA relies on expert opinions and EM rests on a theoretical underpinning. However, POM is defined by the focus of the measurement. POM should therefore not be treated as an alternative to DDA and EM, since a measurement can for example be based on expert opinions *and* focus on the process. (2) *They are not covering the range of measures available* i.e. a large portion of the reviewed literature did not fit into the defined approaches.

Attempts were made to alter Hayes framework, including attempts to introduce extra approaches from other authors. However, countless encounters with the reviewed material eventually led up to the realization that a new framework is a better choice. Resemblances might be found between the classes in the CRM Matrix and the approaches from the addressed literature, but the definitions are so far apart that referencing them would be inappropriate. The new framework, the CRM Matrix (see Figure 5), is grounded in the literature from the Scoping Study through the iterative process of altering the categories then revisiting the data. The iterative process can be divided into 3 main steps:

Step 1: Defining the classes

During the analysis two characteristics seemed to group measurements so that trends among the other characteristics became visible. These are the dimensions of the matrix: Origin and Focus. What was more difficult to provide was the definitions of the classes within these dimensions, i.e. to specify the exact feature that categorize a measurement within one class and not another.

Step 2: Investigating fit between definitions and literature

The measurements were sorted according to the defining criteria within each class, and each measurement belongs to two classes (at least). Measurements that did not fit within the classes definitions were investigated further to sort out what is causing the friction; either the measurement did not fit within any of the classes, or it belonged to multiple classes. If necessary, step 1 would be conducted again, and the classes were redefined based on this new knowledge.

Step 3: Conducting the analysis

The literature was divided according to the definitions of the categories, and the differences between the categories explored, i.e. investigating what makes each category unique. During the analysis new contradictions became clear, hence, step 1 and 2 had to be revisited until these contradictions were settled.

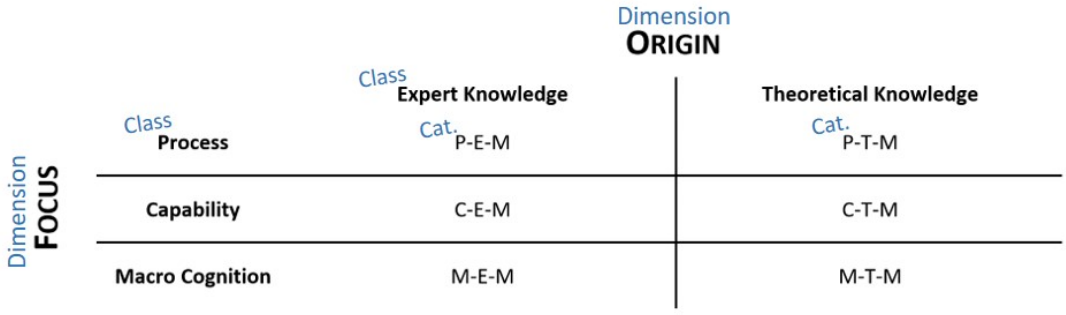

Figure 5. The CRM Matrix.

THE PRACTITIONERS GUIDE FOR C2 EVALUATIONS

This guide goes from the general to the specifics; starting out with an overview of all categories, followed by more detailed description of each one. The individual descriptions are each complemented with a table summarizing the analysis with references to the literature in the Scoping Study. Measurements that belong to multiple classes on a single dimension are called crossovers, these are excluded from the current article but will be discussed in detail in the upcoming article (see '*Conclusions*').

Field overview

Each class and category carry unique strengths and weaknesses (see Table 3), and to choose an unfit evaluation might lead to incorrect conclusions. For example, it would be recommended to choose a measurement that plays to the strengths of the evaluator; if the evaluator is highly skilled within the domain but is a novice in statistics, then it would be preferable to choose from the PEM category rather than PTM. Of course, it would be possible to let one evaluator perform the data collection and another can calculate the statistical analysis.

Measurements that are non-intrusive use data collecting methods that do not interrupt the participants in their work. For a measurement to be usable during real events the data either must be non-intrusive, or the data must be collectable after the event is completed. Post collection of data may however be less reliable, since the participants recollection of an event will become less trustworthy as time passes.

A baseline is a decided value in the analysis that marks that a satisfactory performance or effectiveness have been achieved. Baselines are scarce in all categories, which is problematic. The few that exist are not supported by any logical explanation, with one exception (Farrell, 2004). There are alternatives one can use when baselines are not available. A surrogate baseline can be used, which means that data from another evaluation is used as a standard. Another option is to only make comparisons; either follow the same organization over time to follow progress or investigate whether an organizational or technological change has had the desired effect, i.e. compare the results before and after the change. Not optimal, but these are the only options available today.

Generalizability concerns whether a measurement can be used in another domain than the one it was developed in. Some categories are more problematic to use within new domains, since they focus on procedures or concepts that are domain specific. One example is the Commanders Intent, a concept used within military organizations, but not anywhere else.

Table 3. Overview of all categories in the Practitioners Guide. For references and more detailed information, see the detailed descriptions in each category.

***1 or 2 exceptions are available.**

****This only includes statistical testing, with numbers available for the reader.**

	PEM	**PTM**	**CEM**	**CTM**	**MEM**	**MTM**
Non-intrusive	Yes	Yes*	Yes*	No*	Yes	No*
Use for exercises	Yes	Yes*	Yes	Yes	Yes	No
Use for real responses	Yes	Yes*	Yes*	Yes/No	Yes	Yes
Validity and/or reliability testing	No**	Common	No**	Common	No	Common
Baselines	A few	A few	None	A few	None	A few
Generalizability	Low	Medium/High	Low	High	Medium	High
Evaluator skills/ requirements						
Domain	Expert	Expert	Expert	Novice	Intermediate	Novice/ Intermediate
Statistics	Novice	Intermediate - Expert	Novice	Novice	Intermediate - Expert	Intermediate - Expert
Evaluation method	Intermediate	Intermediate/ Expert	Intermediate/ Expert	Novice	Intermediate - Expert	Intermediate - Expert

PEM: Process – Expert – Measurement

All reviewed papers rely on either observers or retrieved documents for the data collection, one did try to use the actors in the exercise, but the researchers deemed the data hard to trust. The methods for collecting data are non-intrusive for the participants, hence, suitable both for real events and exercises. The analysis uses raw scores, such as percentages and summed scores, making these available for evaluators with only basic experience in statistics. The evaluator must be knowledgeable in the domain and receive training on how to conduct the evaluation properly. Only two of the measurements offered a baseline, but neither gave any explanation on why this was a good standard. To generalize the method to new domains is troublesome since most of them focus on domain specific procedures, with a few exceptions being identified (Rüter, Örtenwau & Vikström, 2007; Brown & Galkovsky, 2007; Entin, Entin & Street 2001).

Table 4. Summary of PEM characteristics with reference to specific articles in the Scoping Study: 1. Brown & Galkovsky, 2007; **2.** Cosgrove, Jenckes, Wilson, Bass & Hsu, 2008; **3.** Djalali, et al., 2014; **4.** Entin, et al., 2001; **5.** Green, et al., 2003**; 6.** Gryth, et al., 2010; **7.** Nilsson & Rüter, 2008; **8.** Nilsson, Vikström & Jonson, 2012; **9.** Rüter, 2006; **10.** Rüter, et al., 2007.

***** Refers to the given number to each reference in the Table description.

** The actual numbers and calculations are not available to the reader.

Characteristic	**Value**	**Number of articles** (total 10)	**Articles** (ref. number*)
Domain	Medical	8	2-3, 4-10
	Team	1	4
	Military	1	1
Data Collection	Observer	8	1-7, 10
	Documents	2	8-9
	Actors	1	5
Data Type	Indicators	8	2-3, 5-10
	Matrix	1	4
	N.A.	1	1
Analysis	Raw scores (%, M, Sum, etc.)	8	3-10
	Maturity level	1	1
	N.A.	2	1, 9
Baseline	N.A.	8	1-5, 8-10
	Specified, not explained	2	6-7
Validity & Reliability	N.A.	8	1-2, 4-6, 8-10
	Discussed, not calculated	1	3, 7
	Validity claimed, not supported**	1	9
Generalizability	No	7	2-3, 5-9
	Yes	3	1, 4, 10
Peer Reviewed	Yes	9	1, 3-10
	No	1	2

PTM: Process – Theory – Measurement

The data collection displays a wider variety of methods, including questionnaires, observers and documents. All measurements are non-intrusive, except for DATMA (Berggren, Johansson, Svensson, Baroutsi & Dahlbäck, 2014; Macmillan, Paley, Entin & Entin, 2005). All measurements can be used for evaluating exercises, except two measurements that are using documents. This is because of the nature of the documents, they are not common to incorporate in training, for example patient files, governmental documents, and after-action reports. The analysis requires an evaluator knowledgeable in statistics, the exception being the Sensemaking Process Instrument that simply sums up the obtained score to form a total value (Jensen, 2006). Again, two articles offer a baseline, but no explanation is given. All but one measurement has statistically evaluated the measurements validity and/or reliability, but not all tests are significant. Many of the measurements can with changes become applicable in new domains, but this does require an experienced evaluator.

Table 5. Summary of PTM characteristics with reference to specific articles in the Scoping Study: 1. Abbasi, et al., 2010; **2.** Abbasi, et al., 2013; **3.** Berggren, et al., 2014; **4.** Hossain & Kit Guan, 2012; **5.** Jensen, 2006; **6.** Kapacu, et al., 2009; **7.** Macmillan, et al., 2005; **8.** Serfaty, Entin & Johnston, 1998.

***** Refers to the given number to each reference in the Table description.

Characteristic	**Value**	**Number of articles** (total 8)	**Articles** (ref. number*)
Domain	Civilian	2	1-2
	Medical	2	4, 6
	Team	2	3,7
	Military	1	5
	Maritime	1	8
Data Collection	Questionnaires	4	1-3, 7
	Documents	2	4, 6
	Observers	2	5, 8
Data Type	N.A.	3	1-2, 6
	Indicators	2	4, 8
	Likert-scales	1	5
	Probes	2	3, 7
Analysis	Raw scores (%, M, Sum, etc.)	3	5, 7-8
	Network Analysis Metrics	3	1, 4, 6
	N.A.	2	2-3
Baseline	N.A.	6	1-3, 5-6, 8
	Specified, not explained	2	4, 7
Validity & Reliability	Validity	5	1-5
	N.A.	3	6-8
	Reliability	1	1
Generalizability	No	4	4-6, 8
	Yes	3	1-3, 7
Peer Reviewed	Yes	6	1-6
	No	2	7-8

CEM: Capability – Expert – Measurement

Articles related to Measures of Effectiveness (MoE) are very different from the other articles within the Scoping Study. MoE differs from all other measurements in the study since they are a framework for developing variables, hence, it is not a specific measurement with specified variables (a few examples are often offered in). Many of the characteristics relate to the specific variables, hence, the label 'Not Available' becomes a recurrent theme.

The data collection for *the other articles* involves either observers or questionnaires. The MoE articles do not offer information on how to collect the data, but the results are presented in time, frequencies, ratio, etc. All measurements are applicable in both exercises and real events (MoE should be developed for real events), except for one which uses questionnaires. The analysis includes raw scores, such as means or summarized scores, making these available for evaluators with only basic understanding of statistics. No baselines are provided. One article claims that the validity has been statistically supported, but the numbers unavailable. The measurements are difficult to generalize to new domains.

Table 6. Summary of CEM characteristics with reference to specific articles in the Scoping Study: 1. Bornman, 1993; **2.** Entin, Entin & Street, 2001; **3.** Hone, Whitworth & Farmilo, 2007; **4.** Sproles, 2002.

***** Refers to the given number to each reference in the Table description.

** The actual numbers and calculations are not available to the reader.

Characteristic	**Value**	**Number of articles** (total 4)	**Articles** (ref. number*)
Domain	Military	3	1, 3-4
	Team	1	2
Data Collection	N.A.	2	1, 4
	Observers	1	2
	Questionnaires	1	3
Data Type	Frequencies, time, intervals, etc.	2	1, 4
	Continuous scales	1	3
	Likert-scales	1	2
Analysis	Raw scores (%, M, Sum, etc.)	3	1-2, 4
	N.A.	1	3
Baseline	N.A.	4	1-4
Validity & Reliability	N.A.	3	1, 3-4
	Validity claimed, not supported**	1	2
Generalizability	No	4	1-4
	Yes	1	2
Peer Reviewed	Yes	3	1
	No	1	2-4

CTM: Capability – Theory – Measurement

The data collection almost exclusively uses questionnaires, with one exception that used observers. All measurements do however use Likert-scales, sometimes complemented with other forms of questions. The advantage of this simple approach is that anyone can perform the data collection, questionnaires are simply handed out and then collected again. About half of the measurements are intrusive and should be applied during pauses in the work, hence not suitable for real events. The rest are possible to conduct post event, and one measurement use observer data. Many do not mention how to conduct the analysis or present the results because the focus of the articles is on testing the measurement and not to conduct an evaluation. The references that do mention the analysis use simple techniques, hence, making them suitable for evaluators with basic knowledge in statistics. Only one article specifies a baseline, but the logic behind it is not explained. Validity and reliability calculations are common and are even the main purpose for several of the articles. The high number of studies focusing on validations makes it surprising that this category has the highest number of non-peer-reviewed articles. All measurements are generalizable across domains.

Table 7. Summary of CTM characteristics with reference to specific articles in the Scoping Study: 1. Berggren, et al., 2014; **2.** Essens, et al., 2010; **3.** Essens, et al., 2005; **4.** Hof, de Konig & Essens, 2010; **5.** Macmillan, et al., 2005; **6.** Matthews & Beal, 2002; **7.** McGuinnes & Foy, 2000; **8.** Salmon, Stanton, Walker, Jenkins, Ladva, Rafferty & Young, 2009; **9.** Schraagen, de Koning, Hof & van Dongen, 2010; **10.** Yanakiev & Horton, 2012.

***** Refers to the given number to each reference in the Table description.

Characteristic	**Value**	**Number of articles** (total 10)	**Articles** (ref. number*)
Domain	Military	5	2-4, 6, 10
	Team	4	1, 5, 7-8
	Civilian	1	9
Data Collection	Questionnaires	10	1-10
	Observers	1	6
Data Type	Likert-scales	10	1-10
	Probes	2	1, 5
Analysis	N.A.	6	1, 3, 6, 7, 9-10
	Raw scores (%, M, Sum, etc.)	5	3, 4-6, 8
Baseline	N.A.	9	1-4, 6-10
	Specified, not explained	1	5
Validity & Reliability	Validity	6	1-2, 6, 8-10
	N.A.	3	4-5, 7
	Reliability	1	10
	Reliability claimed, not supported*	1	9
	Discussed, not calculated	1	3
Generalizability	Yes	10	1-10
Peer Reviewed	No	6	2-3, 5-7, 10
	Yes	4	1, 4, 8-9

MEM: Macro cognitive – Expert – Measurement

Seemingly small, but this category was not believed to exist, since Macro-cognition is a theoretical concept and should have a theory behind it. However, decision making is also a concept that exists in everyday language, and no theories or models are presented in the article. A software is used to collect the data, a software that must be used in the working procedures of the actors being evaluated. This measurement is applicable both for training and real events, but it does impose the working procedure that the software supports. A variety of analyses become available through the software, including cognitive workload over time, operational tempo, shared events analyzer, etc. The evaluator needs experience and understanding of both the software and statistics to use it properly. No validations are presented in the article, and no baseline is offered. It is unclear for what domains the software and measurement are suitable.

Table 8. Summary of MEM characteristics with reference to specific articles in the Scoping Study: 1. Buchler, Neill, Sokoloff & Bakdash, 2013.

***** Refers to the given number to each reference in the Table description.

Characteristic	Value	Number of articles (total 1)	Articles (ref. number*)
Domain	Military	1	1
Data Collection	Software	1	1
Data Type	N.A.	1	1
Analysis	Time series, operational tempo, resource management flow, shared event analyzer, etc.	1	1
Baseline	N.A.	1	1
Validity & Reliability	N.A.	1	1
Generalizability	Yes	1	1
Peer Reviewed	Yes	1	1

MTM: Macro cognitive – Theory – Measurement

Almost every measurement relies on questionnaires to collect that data, hence, making them less applicable for real events. A few exceptions use observers and documents, making them applicable also for real events. The analysis shows clever ways of reflecting higher-level functions through many varieties of similarity calculations, cohesion, sensitivity, and so on. For example, a measurement does not compare a participant's result against a checklist with correct and incorrect answers, but instead compares it to the other team members to calculate cohesiveness amongst the actors (which relates to the concepts of Shared Mental Models). These analyses lead to interesting results and insights not available in most of the measurements belonging to other categories (exceptions are available), but they do require an evaluator knowledgeable both in conducting evaluations and statistics. A few baselines are available, and one even explains the logic behind the decided value, making it the only reference in this study (Farrel, 2004). All but two measurements can be generalized across domains.

Table 9. Summary of MTM characteristics with reference to specific articles in the Scoping Study: 1. Berggren, 2016; **2.** Berggren, Johansson & Nicoletta, 2016; **3.** Entin, et al., 2001; **4.** Farrel, 2004; **5.** Gorman, Cooke, Pederson, Connor & DeJoode, 2005; **6.** Hansberger, Schreiber & Spain, 2008; **7.** Höglund & Berggren, 2012; **8.** Leggatt, 2004; **9.** McGuinnes, 2004; **10.** Rencrantz, Lindoff & Andersson, 2005; **11.** Salmon, et al., 2009; **12.** Seet, The, Soo & Teo, 2004; **13.** Weil, Freeman, Carley, Cooke, Diesner & Weil, 2006.

***** Refers to the given number to each reference in the Table description.

** The actual numbers and calculations are not available to the reader.

Characteristic	Value	Number of articles (total 13)	Articles (ref. number*)
Domain	Team	7	1-5, 7, 11
	Military	6	6, 8-10, 12-13
Data Collection	Questionnaires	11	1-4, 6-12
	Documents	2	6, 13
	Observations	1	5
Data Type	Likert scales	6	3-4, 6, 8-10
	Probes	5	4, 8-9, 11-12
	Ranking scales	3	1-2, 7
	Networks	2	6, 13
	Scoring sheet	1	5
	Questions (open, multiple choice)	1	12
	Frequencies, time, intervals, etc.	1	4
Analysis	Coherence	7	1-3, 6-7, 10, 12
	Raw scores (%, M, Sum, etc.)	4	4-5, 9, 11
	Similarity	2	6, 12
	Sensitivity	2	8-9
	Bias	2	8-9
	Consistency	1	4
	Difference	1	4
	Complementary	1	12
	Overlap	1	13
	Prediction	1	13
Baseline	N.A.	12	1-5, 7-13
	Specified, logic explained	1	4
	Specified, not explained	1	4
	Surrogate baseline	1	6
Validity & Reliability	Validity	7	1-5, 10-11
	N.A.	3	8-9, 13
	Validity claimed, not supported**	2	6, 12
	Not significant	1	7
Generalizability	Yes	11	1-3, 5-7, 9-13
	No	2	4, 8
Peer Reviewed	Yes	12	1-9, 11-13
	No	1	10

CONCLUSIONS

Each evaluation has specific requirements and purpose to be met, for example and to accomplish the best result, these requirements ought to be reflected in the chosen measurement. It may be tempting to simply choose a questionnaire, hand it out and then collect it again, but the evaluation will not be more rewarding than the knowledge of the evaluator.

The purpose of the CRM Matrix is to categorize measurements available today and offer guidance. This framework should not be seen as restrictive, and future measurements might not fall within any of the classes or current definitions.

FUTURE WORK

Not included in this article are the crossovers; measurements that belong to multiple classes on a single dimension. These crossovers behave differently depending on which dimension they appear on. For crossovers in the Origin Dimension, the measurement is initially developed based on knowledge from one class, and then knowledge from the other class is incorporated in a modified version (e.g. Jensen, 2016). This seems to strengthen the measurement and have a positive effect. Crossovers in Focus dimension are only present for measurements containing multiple variables. One variable cannot belong to two classes, but the individual variables can belong to different classes. Measurements that show a clear distinction between the two types of variables are included in this article but divided according to each variable (e.g. DATMA, Macmillan, et al. 2005). Some measurements do not show a clear distinction and therefore become difficult to analyse (e.g. HEAT, Hayes & Wheatley, 2001). To not clearly distinguish between these types of variables within a measurement could have negative effects. This will all be covered in the future article that is meant to complement this guide, with a focus on future method development and implications for the scientific community.

REFERENCES

Abbasi, A., Hossain, L., Hamra, J., & Owen, C. (2010) Social networks perspective of firefighters' adaptive behaviour and coordination among them. *IEEE/ACM International Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*, 819–824.

Abbasi, A., Owen, C., Hossain, L., & Hamra, J. (2013) Social connectedness and adaptive team coordination during fire events. *Fire Safety Journal*, 59, 30–36.

Arksey, H. & O'Malley, L. (2005) Scoping studies: Towards a methodological framework. *International Journal Social Research Methodology*, 8 (1), 19-32.

Berggren, P. (2016) *Assessing Shared Strategic Understanding.* Linköping Studies in Arts and Science, Dissertation No. 677. Linköping, Sweden: Linköping University Electronic Press.

Berggren, P., Johansson, B. J. E., & Baroutsi, N. (2017) Assessing the quality of Shared Priorities in teams using content analysis in a microworld experiment. *Theoretical Issues in Ergonomics Science*, 18(2), 128-146.

Berggren, P., Johansson, B. J. E., Svensson, E., Baroutsi, N., & Dahlbäck, N. (2014) Statistical modelling of team training in a microworld study. In *Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting* (pp. 894–898). Chicago.

Bornman Jr., L. G. (1993) *Command and Control Measures of Effectiveness Handbook.* [Technical Document TRAC-TD-0393]. Fort Leavenworth, Kansas.

Brehmer, B. (2007) Understanding the functions of C2 is the key to progress. *International C2 Journal*, 1, 211-232.

Brown, K. M., & Galkovsky, M. (2007) Evaluating and Enhancing C2 Systems with the Decision-Making Assessment Process (D-MAP). In *International Command and Control Research and Technology Symposium.*

Buchler, N., Neill, D. O., Sokoloff, S., & Bakdash, J. Z. (2013) The Warfighter Associate: Decision-Support and Metrics for Mission Command. In *the International Command and Control Research and Technology Symposium.*

Cosgrove, S. E., Jenckes, M. W., Wilson, L. M., Bass, E. B., & Hsu, E. B. (2008) *Tool for Evaluating Core Elements of Hospital Disaster Drills.* Prepared by John Hopkins Evidence-based Practice Center under Contract No. 290-02-0018

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

Endsley, M. R., (1988) Design and evaluation for situation awareness enhancement. *In Proceedings of the Human Factors and Ergonomics Society 32nd Annual Meeting*, pp. 97–102 (Santa Monica, CA: Human Factors and Ergonomics Society).

Entin, E. E., Entin, E. B., & Street, G. (2001) Measures for Evaluation of Team Processes and Performance in Experiments and Exercises. In *International Command and Control Research and Technology Symposium.*

Essens, P., Vogelaar, A., Mylle, J., Baranski, J., Goodwin, G., Van Buskirk, W., Berggren, P. & Hof, T. (2010) *CTEF 2.0- Assessment and Improvement of Command Team Effectiveness: Verification of Model and Instrument* (Vol. 323). Retrieved from http://handle.dtic.mil/100.2/ADA534290.

Essens, P., Vogelaar, A., Mylle, J., Blendell, C., Paris, C., Stanley, H., & Baranski, J. (2005) *Military command team effectiveness: Model and instrument for assessment and improvement.* [ISBN 92-837-1135-1]. Neuilly-sur-Seine Cedex, France.

Farrell, P. S. E. (2004) Measuring Common Intent during Effects Based Planning. In *International Command and Control Research and Technology Symposium.*

Gorman, J. C., Cooke, N. J., Pederson, H. K., Connor, O. O., & Dejoode, J. A. (2005) Coordinated awareness of situation by teams (CAST): Measuring team situation awareness of a communication glitch. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 2–5).

Green, G. B., Modi, S., Lunney, K., & Thomas, T. L. (2003) Generic evaluation methods for disaster drills in developing countries. *Annals of Emergency Medicine*, 41(5), 689–99.

Gryth, D., Radestad, M., Nilsson, H., Nerf, O., Svensson, L., Castren, M., & Riiter, A. (2010) Evaluation of Medical Command and Control Using Performance Indicators in a Full-Scale, Major Aircraft Accident Exercise. *Prehospital and Disaster Medicine*, 25(2), 118–125.

Hansberger, J. T., Schreiber, C., & Spain, R. D. (2008) C2 Network Analysis: Insights into Coordination & Understanding. In *International Command and Control Research and Technology Symposium* (pp. 81–90).

Hayes, R. E. (2012) Measuring Command and Control Effectiveness [PowerPoint slides]. *Presented at MORS Workshop.* Retrieved at https://pdfs.semanticscholar.org/0dd7/a67af43425122df4bd1c6ddc130e21fc24f3.pdf

Hayes, R. E., & Wheatley, G. (2001) The Evolution of the Headquarters Effectiveness Assessment Tool (HEAT) and Its Applications to Joint Experimentation. In *International Command and Control Research and Technology Symposium.*

Hof, T., de Koning, L., & Essens, P. (2010) Measuring Effectiveness of Teams and Multi-team Systems in Operation. In *International Command and Control Research and Technology Symposium.*

Hone, G., Whitworth, I. R., & Farmilo, A. (2007) Assessing the transmission of Command Intent. In *International Command and Control Research and Technology Symposium.*

Hossain, L., & Kit Guan, D. C. (2012) Modelling coordination in hospital emergency departments through social network analysis. *Disasters,* 36(2), 338–365.

Höglund, F., & Berggren, P. (2010) Using Shared Priorities to Measure Shared Situation Awareness. In *Information Systems for Crisis Response and Management* (pp. 1–5).

Jensen, E. (2006) Good Sensemaking is More Important than Information for the Quality of Plans Good Sensemaking is More Important than Information for the Quality of Plans. In *International Command and Control Research and Technology Symposium.*

Jensen, E. (2016) Sensemaking in military planning: A methodological study of command teams Sensemaking in military planning. *Cognition Technology and Work,* 11, 103-118.

Kapucu, N., Augustin, M. E., & Garayev, V. (2009) Interstate Partnerships in Emergency Management: Emergency Management Assistance Compact in Response to Catastrophic Disasters. *Public Administration Review*, 69(2), 297–314.

Klein, G.A., Ross, K.G., Moon, B.M., Klein, D.E., Hoffman, R.R., & Hollnagel, E. (2003) Macrocognition. *IEEE Intelligent Systems*, 18(3), 81-85.

Leggatt, A. (2004) Objectively measuring the promulgation of commander's intent in a coalition effects based planning experiment (MNE3). In *International Command and Control Research and Technology Symposium.*

Macmillan, J., Paley, M. J., Entin, E. B., & Entin, E. E. (2005) Questionnaires for Distributed Assessment of Team Mutual Awareness. In N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. Hendricks. *Handbook of Human Factors Methods.* London: Taylor and Francis.

Matthews, M. D., & Beal, S. A. (2002) *Assessing Situation Awareness in Field Training Exercises.* [Research Report 1795]. West Point, New York: U.S. Army Research Institute for the Behavioral and Social Sciences Research.

Mcguinness, B. (2004) Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings. In *International Command and Control Research and Technology Symposium.*

McGuinness, B., & Foy, L. (2000) A subjective measure of SA: the Crew Awareness Rating Scale (CARS). In *Human Performance, Situational Awareness an Automation Conference.*

Nilsson, H., & Rüter, A. (2008) Management of resources at major incidents and disasters in relation to patient outcome: a pilot study of an educational model. *European Journal of Emergency Medicine*, 162–165.

Nilsson, H., Vikström, T., & Jonson, C.-O. (2012) Performance indicators for initial regional medical response to major incidents: a possible quality control tool. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 20, 81.

Rencrantz, C., Lindoff, J., & Andersson, J. (2005) *Är Det Viktigt Att Förstå Varandra För Att Prestera Bra?* Stockholm.

Rüter, A. (2006) *Disaster medicine- performance indicators, information support and documentation A study of an evaluation tool.* Linköping University Medical Dissertations No. 975.

Rüter, A., Örtenwall, P., & Vikstrom, T. (2007) Staff Procedure Skills in Management Groups during Exercises in Disaster Medicine. *Prehospital and Disaster Medicine*, 22(4).

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009) Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, 39(3), 490–500.

Schanteau, J. (1992) Competence in Experts: The Role of Task Characteristics. *Organizational Behavior and Human Decision Processes*, 53(October), 252–266. https://doi.org/10.1016/0749-5978(92)90064-E.

Schraagen, J. M., de Konig, L., Hof, T. & van Dongen, K. (2010) Development of a self-rating instrument to measure team situation awareness. In *International Command and Control Research and Technology Symposium.*

Seet, A. W. K., Teh, C. A., Soo, J. K. T., & Teo, L. (2004) Constructible Assessment for Situation Awareness in a Distributed C2 Environment. In *International Command and Control Research and Technology Symposium.*

Serfaty, D., Entin, E. E., & Johnston, J. H. (1998) Team coordination training. In E. Salas & J. A. Cannon-Bowers (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 221–245). Washington, DC.

Sproles, N. (2002) Formulating Measures of Effectiveness. *Systems Engineering*, 5(4), 253–263.

Weil, S. A., Freeman, J., Carley, K. M., Cooke, N. J., Diesner, J., & Weil, S. (2006) Measuring Situational Awareness through Analysis of Communications: A Preliminary Exercise. In *International Command and Control Research and Technology Symposium.*

Yanakiev, Y., & Horton, J. (2012) *Improving the Organisational Effectiveness of Coalition Operations* (Vol. 323).