Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Quality of security metrics and measurements

*Reijo M. Savola**

VTT Technical Research Centre of Finland, Kaitoväylä 1, 90650 Oulu, Finland

ARTICLE INFO

Article history:
Received 13 August 2012
Received in revised form
30 April 2013
Accepted 5 May 2013

Keywords:
Security metrics
Security quantification
Quality of security metrics
Expert opinion survey
Security effectiveness

ABSTRACT

Quantification of information security can be used to obtain evidence to support decision-making about the security performance of software systems. Knowledge about the relational importance of the main quality criteria of security metrics can help build security metrology models based on practical needs. This paper presents the results of a quantitative security metrics expert survey of 141 respondents, and an associated interview study, regarding the prioritization of 19 quality criteria of security metrics identified in the literature. The interviews were used to validate the survey results and to obtain further information on the findings. The results identified three foundational quality criteria of security metrics: correctness, measurability, and meaningfulness. These criteria form the basis for credibility and sufficiency for security metrics and associated measurements. Moreover, usability was seen as an important criterion. The paper analyzes the foundational and related quality criteria and proposes a model of them.

1. Introduction

As software systems that run critical applications become more complex, connected, and dynamic in nature, the related information security management becomes more challenging. There is a need for systematic techniques with which to obtain quantitative evidence of the operational systems' security performance. Quantification is widely used in engineering and management as a means of increasing the understanding of complex real-world phenomena and enabling informed and adaptive decision-making.

Nowadays, domain-specific expertise is the standard knowledge used to manage security. However, there is a need for considerable improvement of quantitative methods and measures for security management. Not enough attention has been paid to them in different domains. A cross-cut model for *security metrology* is needed to bridge the gap between domain-specific expert judgment and holistic understanding of the security risk. Research on the quantification of security as a systematic discipline is still in its infancy, although some examples of industrial security measurement practices do exist. SM (Security Metrics) can be used to offer security evidence for security engineering, risk and security management, and internal and external evaluation.

In practice, SM developers are often asked: *Can I trust these SM?* Like traditional software engineering (Kaner and Bond, 2004), this paper asserts that secure software engineering and operational security management under-emphasize the validity and quality of measurements and metrics, which causes misunderstandings, measurement distortions, and other problems and results in the rejection of the SM. The results of this analysis will help in the development and selection of adequate and credible SM and measurement architectures. This analysis will also hopefully be a step toward the creation of the security metrology model. The following hypothesis was proposed: (*H*) *A few foundational generic qualitative properties dominate in most SM.*

The study uses the following two-step research methodology:

1. Quantitative EOS (Expert Opinion Survey) on the prioritization of the quality criteria

* Tel.: +358 40 569 6380; fax: +358 20 722 2320.
E-mail addresses: Reijo.Savola@vtt.fi, Reijo.Savola@gmail.com.
0167-4048/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cose.2013.05.002

2. EIS (Expert Interview Study) to validate and further investigate the results of Step 1

Advancements in the topic of SM and measurements are constrained by disagreement on what constitutes essential evidence. The motivation for the research reported in this paper comes from the fact that the previous work has identified a wide collection of quality criteria for SM, and it would be impossible to develop SM that meet all the demands. There is also growing interest in adopting and using SM, yet a lack of widely accepted solutions. Recently, more researchers and industrial practitioners have started to investigate SM, making it possible to use their expertise to analyze prioritization.

Section 2 discusses the background and related work of SM and their quality. Section 3 presents EOS results, while Section 4 presents the EIS results. Section 4 also evaluates the hypothesis and introduces a conceptual SM quality criteria model. Section 5 discusses the findings, before Section 6 offers conclusions and poses some future research questions.

2. Background and related work

Security metrics has become a standard term in the context of Information Technology (IT) when referring to metrics that depict the security level, security performance, security indicators, or security strength of a System under Investigation (SuI) (Savola, 2009) – a technical system, product, service, or organization. This term is misleading, however, because it implies that traditional concepts in metrology apply equally to IT (Jansen, 2009). Security cannot be measured as a universal concept due to the complexity, uncertainty, non-stationarity, limited observability of operational systems, and malice of attackers (Verendel, 2010). Therefore, terms such as *indicators* or *strength* may be more appropriate in the case of security-related objectives. This study, however, uses the most widely used term, *metrics*. ISO/IEC International Vocabulary of Metrology (ISO/IEC Guide 99:2007, 2007) defines *measurement* as a process of experimentally obtaining information about the magnitude of a quantity. Quantity is a property of a phenomenon, body, or substance to which a magnitude can be assigned. A *measurement* result indicates single-point-in-time data on a specific factor to be measured, while *metrics* are descriptions of data derived from measurements used to facilitate decision-making (Savola, 2009). Security cannot be well managed (and measured) without properly understanding the security risk: simply, there cannot be significant risk without the potential for significant loss (Jones, 2005). See our previous work (Savola, 2009, 2010; Savola et al., 2011, 2012) for further discussion on security measurement objectives and associated definitions.

2.1. Related work on quality criteria of SM

Based on the literature, the author's earlier work (Savola, 2010) identified some important quality criteria for SM. However, their prioritization has remained an open question. Table 1 summarizes these criteria, as grouped in (Savola, 2010), with references to the original discussion. Table 2 introduces a brief survey of contributions to SM quality.

Some of the listed criteria are clearly more important to SM than others, and some depend on others or may even conflict with them. As far as can be ascertained, no thorough discussions have taken place on the quality of SM or the measurements.

In *software* metrics, several authors have discussed the quality criteria, including (Kaner and Bond, 2004) and (Fenton et al., 1996). While these results are partly applicable to SM, the main measurement objectives of software metrics differ from those of SM.

2.2. Discussion of quality criteria

The term *accuracy* is often used instead of *correctness* to emphasize the fact that all-inclusive correctness is impossible. However, in this study, we prefer to use the term correctness, because it makes it possible to differentiate between accuracy and precision. Correctness is an important for the credibility of SM.

Table 3 proposes questions that are associated with the correctness quality criteria of SM and their sub-criteria. Questions Q1–Q4 are motivated by the software quality metrics *validation criteria* presented in the IEEE Standard 1061 (IEEE Standard 1061-1998, 1998), as well as risk analysis quality criteria (Rouhiainen, 1990). The reason for the use of (IEEE Standard 1061-1998, 1998) is that validity is a measure that illustrates how *correctly* a method measures the factors it is intended to measure (Rouhiainen, 1990), consisting of predictive, content, and construct validity. The latter is a central concept in industrial–organizational psychology (Westen and Rosenthal, 2003). SM with good prediction accuracy reduce the level of uncertainty in the outcome of the security risk (Rouhiainen, 1990). Tracking of risk dynamics is also important for the prediction, as Q3 emphasizes.

Questions Q5–Q12 in Table 3 are composed based on the SM quality criteria related to correctness discussed in Table 1: time dependability, representativeness, granularity, context specificity, completeness, unbiasedness, objectivity, and non-intrusiveness. In the following, we discuss their role and how they contribute to correctness.

The *time dependability* of metrics can be seen partly as a sub-criterion of prediction accuracy and partly as an independent sub-criterion of 'general' correctness. Axelrod (2003) claimed that "The biggest source of risk to a networked information system is an over-reliance on recent experience to evaluate current security." Attention should be paid to the frequency of the metric reporting, and its utilization and duty cycle when developing SM.

The *representativeness* of SM is crucial to their correctness in a security context. The *representation condition* (Fenton et al., 1996) for measurement answers Questions Q6–Q8 in Table 3. Q6 is the main question related to *construct validity* (Rouhiainen, 1990; Westen and Rosenthal, 2003).

Measurement results only provide a rough estimate of the reality (Savola, 2010). The *granularity* of a metric and associated measurement should be at least at a level where adequate decision-making based on them is possible. Q7 refers to the degree of granularity. Greater granularity can provide better *precision* up to a point, but often at a cost in terms of time and effort, which creates trade-off questions with

Table 1 – Some unclassified quality criteria of SM discussed in the literature (Savola, 2010).

Dimension	Ref.	Explanation
Correctness	Williams and Jelen, 1998	The SM are correctly implemented and error-free.
Granularity	Böhme et al., 2008a	The SM allow the measurement results that differ from each other to be distinguished at an adequate level.
Objectivity and unbiasedness	Schechter, 2004; Atzeni and Lioy, 2005	*Objectivity*: The measurement results are not influenced by the measurer's will, beliefs, or actual feelings (Atzeni and Lioy, 2005). *Unbiasedness*: The results are not influenced by any bias.
Controllability	Savola, 2010	The measurement results should be kept within the defined limits or the measurement window.
Time dependability	Jelen et al., 2000; Kanoun et al., 2004; Henning et al., 2002	The time-dependent behavior of SM can be leading, coincident, or lagging (Jansen, 2009). The time dependability of the SM should be part of their specification.
Comparability	ISO/IEC 21827:2003, 2003	The SM should support comparison of the targets that they represent.
Measurability	Williams and Jelen, 1998; Jelen et al., 2000; Rathbun, 2009; Jaquith, 2007	The SM are capable of having dimensions, quantity or capacity ascertained (Williams and Jelen, 1998) in the SuI.
Attainability, availability and easiness	Atzeni and Lioy, 2005; Kanoun et al., 2004; Henning et al., 2002	*Attainability* means that measurement results can be achieved from the SuI. *Availability* implies that they are generally available. *Easiness* refers to how easy it is to achieve the measurement results.
Reproducibility, repeatability, scale reliability	Schechter, 2004; Atzeni and Lioy, 2005; Jelen et al., 2000; Kanoun et al., 2004; Henning et al., 2002; Rathbun, 2009	*Reproducibility* and *repeatability* indicate that the same results are achieved if a measurement is repeated in the same context, with exactly the same conditions (Atzeni and Lioy, 2005). *Scale reliability* refers to the reproducibility of the measurement by different persons.
Cost effectiveness	Williams and Jelen, 1998; Kanoun et al., 2004; Rathbun, 2009; Jaquith, 2007	The measurement gathering and measurement approaches should be cost effective.
Scalability, portability	Williams and Jelen, 1998; Kanoun et al., 2004	*Scalability* means that the SM should be applicable to SuIs of different sizes. *Portability* refers to the applicability of the SM to various target systems (Kanoun et al., 2004).
Non-intrusiveness	Kanoun et al., 2004	The measurements should not harm the normal operation of the SuI, require only minimum changes to the SuI, and not affect the measurement results.
Meaningfulness	Schechter, 2004; Henning et al., 2002; Rathbun, 2009	The SM should be relevant and respond to the needs.
Effectiveness	Williams and Jelen, 1998	The expectations of the adequacy of the SM sufficiency in the final use environment are satisfied.
Efficiency	Williams and Jelen, 1998	The adequate requirements of the SM are achieved with only minimal undesired use of effort and time.
Representativeness, contextual specificity	Jelen et al., 2000; Kanoun et al., 2004; Henning et al., 2002; Rathbun, 2009; Jaquith, 2007	The SM correspond to the actual system characteristics in the SuI in a contextually focused way.
Clarity, succinctness	Atzeni and Lioy, 2005	*Clarity* means that the SM are clearly formulized. *Succinctness* means that only important parameters are considered (Atzeni and Lioy, 2005).
Ability to show progression	Schechter, 2004	The SM should be able to show progression on the dimension it addresses.
Completeness	Williams and Jelen, 1998	The collection of SM should be complete from a measurement objectives perspective.

usability. *Contextual specificity* (or, inversely, *contextual independence*) is a special case of granularity. *Completeness* of SM is related to representativeness, addressing one or a collection of metrics.

It is important to aim for SM that are as *unbiased* as possible (Savola et al., 2012). In security, a remarkable degree of subjectivity is unavoidable, creating a form of correctness bias. Objectivity is a sub-dimension of unbiasedness. There is also a connection between objectivity and meaningfulness: the degree of objectivity is often driven by the clarity of the metrics.

Non-intrusiveness is an important criterion related to correctness. Security measurements and the associated metrics should not overly affect or hinder the actual functionality of software systems or the functions of an organization. Non-intrusiveness is also part of the usability of SM, as discussed below.

Measurability is a prerequisite for the meaningfulness and usability of SM. Williams and Jelen (1998) defined the measurability of a metric by the fact that it has criteria, quantity, or capacity ascertained in the SuI. The availability of measurable information is particularly important to the

Table 2 – A brief survey of SM quality investigations.

Contribution	Ref.
Literature survey on security quality dimensions and introduction of feasibility evaluation process based on the findings	Savola, 2010
Suggestions of properties relevant to security assurance	Williams and Jelen, 1998
Suggestion that certain properties (quality criteria) of SM are important	Atzeni and Lioy, 2005; Jelen et al., 2000; Jaquith, 2007
A collection of dependability benchmark validation criteria	Kanoun et al., 2004
Introduction of a validation approach for analytical dependability metrics	Böhme et al., 2008b
Analysis of validity types (face, content, criterion and construct) for SM	Bayuk, 2011

measurability of the metric. This definition is extended here as follows: *Measurability of a metric means that it has dimensions, quantity or capacity ascertained in the measurement architecture and the measurable information is attainable with sufficient precision*. Table 4 summarizes the proposal of the main questions regarding measurability and the related criteria of SM.

Attainability of measurable information is related to measurability, while *availability* is a sub-goal of attainability. A major problem in measuring security is the lack of adequate security data (MacQueen, 1967). Nowadays, data arising from the vulnerability discovery/disclosure/patching cycle offer the widest resource.

Together, *reproducibility* and *repeatability* form the *precision* of the measurement. *Scale reliability* is a sub-criterion of reproducibility. In addition to measurability, reproducibility, repeatability, and scale reliability are related to correctness.

Table 4 – Questions relevant to SM measurability.

Question	Criterion
Q13: *Can the measurement data be achieved from the system under investigation?*	Attainability
Q14: *Are the measurement data available?*	Availability
Q15: *Are the same results returned if a measurement is reproduced in the same context, with exactly the same conditions?*	Reproducibility
Q16: *Are the same results returned if a measurement is repeated in the same context, with exactly the same conditions?*	Repeatability
Q17: *Are the same results returned if a measurement is reproduced in the same context, with exactly the same conditions by different measurers?*	Scale reliability

Organizations tend to pick *convenient* SM rather than *meaningful* ones (Jaquith, 2010). To be meaningful, the metric should answer the original *essential question* that reflects the need for evidence, such as: *Is the access control solution specified in the security objectives effective enough?* or *Are the losses reduced by at least 20% due to the deployment of security controls?* or *Are we secure enough?*

Clarity is closely related to meaningfulness: the clearer the formulation of the metric, the easier it is to understand provided that the person interpreting it has enough knowledge about the underlying context. *Succinctness* increases clarity and thereby meaningfulness. Succinctness is important in security measurements because the complexity and uncertainty of security-related phenomena make it necessary to have a wide collection of metrics. Good succinctness also increases the *efficiency* of the metric, a sub-goal of usability.

In order for the SM to be meaningful, they should incorporate *applicability to decision-making*. Comparability and ability to show progression belong to this category as sub-criteria. *Comparability* of different measurement results is desirable when making selection decisions among different security

Table 3 – Questions relevant to SM correctness.

Questions	Criteria
Q1: *Do the results of the metric and the measured property in SuI correlate accurately enough?*	Accuracy (correctness) of correlation
Q2: *Are the results of the metric consistent enough?*	Accuracy (correctness) of consistency
Q3: Are the results of the metric able to depict the dynamics of the measured property in SuI with sufficient accuracy?	Accuracy (correctness) of dynamics tracking
Q4: Is the metric able to predict the security risk for the purposes of the use of the metric with sufficient accuracy?	Accuracy (correctness) of prediction
Q5: *Does the metric address time-related relationships of measured data in an adequate way with sufficient accuracy?*	Time dependability
Q6: *Are we measuring the attribute that we really want to measure?*	Representativeness
Q7: *Do we know enough about an attribute before it is reasonable to consider measuring it?*	Granularity
Q8: *Do we know enough about the context before it is reasonable to consider measuring it?*	Context specificity
Q9: *Does the collection of SM cover all essential security measurement objectives to a sufficient degree?*	Completeness
Q10: *Are the measurement results influenced only by a sufficiently small amount of bias?*	Unbiasedness
Q11: *Are the measurement results influenced by the measurer's will, beliefs or actual feeling to a sufficiently small degree?*	Objectivity
Q12: *Is the target of the metric affected by the measurement? Is this effect kept low enough not to make the measurement too inaccurate?*	Non-intrusiveness

controls (Information Technology, 1991). The *ability to show progression*, a special case of comparability, is a sub-criterion of meaningfulness on most security measurement uses. Trend analysis is one of the most important applications of SM.

Pironti (2007) summarized that the key to the success is the meaningfulness of SM, noting that the metrics and measurements should be focused and their value should be easily recognizable and apparent to the intended audience. Table 5 proposes the questions related to meaningfulness.

Usability of SM is important, yet not as critical as correctness, measurability and meaningfulness, because poor usability is not fatal for SM. The criteria related to usability include the following: *efficiency*, *cost effectiveness* and *controllability*, *scalability*, and *portability*. Table 6 presents the proposal of questions related to the usability of SM.

The *effectiveness* of SM is essentially the same as its usability: a metric is 'effective' because it is usable. Accordingly, the term effectiveness can be replaced by the term usability. According to (Frøkjær et al., 2000), usability comprises effectiveness, efficiency, and satisfaction. The term 'satisfaction' in this context has a very similar meaning to meaningfulness. The most desirable of all the effective measurements are the cost effective ones. *Controllability* can be seen as a usability dimension and is also related to measurability.

In addition to usability, *efficiency* is a function of attainability: if measurement results are poorly attainable, efficiency decreases because more data are needed to carry out the measurement. According to Q28, efficiency aims to consume the minimum amount of undesired resources. *Easiness* (of achieving adequate SM and measurements) and *cost effectiveness* mentioned in Table 1, are sub-goals of efficiency.

SM should be as *scalable* and as *portable* as possible, in order to become more widely used. These criteria are also related to contextual specificity: the higher the contextual independence, the higher the scalability and portability.

Table 5 – Questions relevant to SM meaningfulness.

Question	Criterion
Q18: *Is the metric meaningful in the context of its use?*	Meaningfulness
Q19: *Is the metric meaningful to the measurer in the context of its use?*	Meaningfulness to the measurer
Q20: *Is the metric meaningful to the audience in the context of its use?*	Meaningfulness to the audience
Q21: *Are the metric and associated measurements clearly formulized?*	Clarity
Q22: *Are only important parameters considered in the metric?*	Succinctness
Q23: *Is the metric applicable to the planned decision-making?*	Applicability to decision-making
Q24: *Does the metric support comparability?*	Comparability
Q25: *Is the metric able to show progression?*	Ability to show progression
Q26: *Are the metrics and related measurements useful in the decision-making?*	Usefulness

Table 6 – Questions relevant to SM usability.

Question	Criterion
Q27: *Are the measurement results kept within the defined limits or the measurement window?*	Controllability
Q28: *Are adequate SM and measurements achieved while only consuming the minimum amount of undesired resources?*	Efficiency
Q29: *Are adequate SM and measurements achieved while only consuming minimal costs?*	Cost effectiveness
Q30: *Are adequate SM and measurements achieved while only consuming the minimum amount of effort and time?*	Easiness
Q31: *Are adequate SM and measurements applicable to SuIs of different desired sizes?*	Scalability
Q32: *Are adequate SM and measurements 2applicable to different desired SuIs?*	Portability

3. Quantitative security metrics expert opinion survey

An EOS (Expert Opinion Survey) was carried out to elicit SM expert opinions about the prioritization of the SM criteria in Table 1. In the following, we discuss the survey process and its results.

3.1. *Respondents, questionnaire administration and data collection*

Security experts with experience of SM (group G_s) were chosen to respond to the questionnaire. In addition, a smaller control group consisted of IT experts with security competence. G_s was selected from the experts active on the securitymetrics.org SM e-mail discussion list (www.securitymetrics.org, 2013) and SM experts from the MASTER EU FP7 (MASTER, 2013), GEMOM EU FP7 (GEMOM, 2013), SOFIA ARTEMIS (SOFIA, 2013), and BUGYO Beyond CELTIC Eureka (BUGYO, 2013) research projects, and from the author's personal SM expert contacts. The criteria for the selection of SM experts G_s were any of the following demonstrated activities: (i) scientific or industrial SM publications, (ii) active participation in SM research projects or (iii) industrial-strength experience of the topic. The competence was measured in the case of (i) by the peer-review process outcome (accepted scientific papers) and relevance (industrial white papers), in the case of (ii), contributions to the projects, and in the case of (iii), demonstrated modeling and/or use of SM in industrial-strength cases. Only a few of the SM expert candidates originally considered to be part of the questionnaire were not chosen due to failing all the criteria, because most of them were known beforehand to meet the criteria. The criteria were not enforced for the control group G_{ns} with 29 respondents. This group had expertise in *metrics related to software engineering* (but not in SM), and the sufficient knowledge of the security field.

Table 7 presents statistical information about the respondents. There were respondents from 21 countries: 61 (43%) from Finland, 20 (14%) from the U.S., nine (6%) from

Spain, seven (5%) from each of Italy and Germany, and five (4%) from each of Austria, Ireland and Norway. The remaining respondents were from Belgium, the Czech Republic, France, Greece, Hungary, India, Luxembourg, the Netherlands, Norway, Singapore, South Africa, Sweden, and the U.K.

The questionnaire asked respondents to prioritize the 19 quality criteria in Table 1, given in the same order as in that table, and associated letters in alphabetical order from A to S. The prioritization question was: "*Write a sequence of letters associated with each criterion in priority order (for example: 'A B C D E…'). Please suggest new criteria or modifications if you think they are needed.*"

All the experts answered independently and anonymously, and their prioritization profiles had equal weight. The administration mode was computerized self-administration by the respondents. The data were gathered between April and June 2010, via e-mail.

Before the survey, the questionnaire was tested by interviewing five SM experts. They were asked to estimate the degree of importance between different ranks. The mean of the answers, clearly showing an exponential distribution, was used to scale the normalized prioritization opinion function of Eq. (1).

3.2. Quantification and weighting of ranks

Only the first six quality criteria can be seen as prominent in a prioritization carried out by humans. The first three decisions were assumed to be critical, with the first being the most critical. Taking into account the above observations and assumptions, the *normalized prioritization opinion* is defined as follows:

$$\overline{\Omega}_{i,j} = \frac{e^{\Omega_{i,j}/3}}{e^{\max(\Omega_{i,j})/3}} = \frac{e^{\Omega_{i,j}/3}}{e^{19/3}}, \quad \overline{\Omega}_{i,j} \in \left[e^{-6}, 1\right], \tag{1}$$

where i is the respondent index number $i \in G$, G is the group of respondents under investigation, $j \in \{1,2, \ldots, 19\}$ is the quality dimension under investigation, $\Omega_{i,j} \in \{1, 2, \ldots, 9\}$ is the rank associated with the quality dimension, max $(\Omega_{i,j}) = 19$, $\min(\Omega_{i,j}) = 1$, and e is Napier's constant (2.71828…). The mean of the normalized prioritization opinions for each quality dimension j is:

$$\mu_{j,G}(\overline{\Omega}) = \frac{1}{n_G} \cdot \sum_{i=1}^{n_G} \overline{\Omega}_{i,j}, \tag{2}$$

where N_G is the number of individual experts in G. Similarly, the mean of the ranks for each quality dimension j is:

$$\mu_{j,G}(\Omega) = \frac{1}{n_G} \cdot \sum_{i=1}^{n_G} \Omega_{i,j}. \tag{3}$$

In the investigation of variance based on the exponential scale $\overline{\Omega}_{i,j}$, the first priorities always have higher values; therefore, this is not applicable to reliability analysis. This is not the case in the variance calculations based on ranks $\Omega_{i,j}$. The unbiased estimate of variance $s^2_{j,R}(\Omega)$ is based on ranks:

$$s^2_{j,R}(\Omega) = \frac{1}{M-1} \sum_{i=1}^{M} \left(\Omega_{i,j} - \mu_{j,R}(\Omega)\right)^2. \tag{4}$$

The normalized prioritization opinion formula of Eq. (1) causes the prioritization decisions carried out by the experts to be *exponentially pre-distributed*. This pre-distribution is a typical practice, especially in industrial-strength risk assessment methods.

3.3. Survey results

3.3.1. General ranking

Table 8 shows the prioritization order compiled from the respondent data (R_{all}, $n = 141$), indicating that correctness is the most important quality dimension for SM, followed by measurability and meaningfulness. Therefore, CMM (Correctness, Measurability, and Meaningfulness) quality criteria can be seen as *foundational quality criteria* for SM. In the table, the overall prioritization order is based on the mean of prioritization opinions per quality, and $\mu_{all} = \mu_{j,R_{all}}(\overline{\Omega})$, p_x = rank based on $\mu_{j,C_x}(\overline{\Omega})$, (1 = first priority, 19 = last priority), and $s^2_x = s^2_{j,C_x}(\Omega)$ is the unbiased estimate of expert opinion cluster C_x variance. This choice is based on the independence, anonymity and equal weighting of the responses.

Table 7 – Statistical background information on survey respondents.

Property			n	% of n_{all}	n in clusters by k-means clustering					
					C_1	C_2	C_3	C_4	C_5	C_6
All respondents	G_{all}		141	100%	54	28	26	15	11	7
Employment	Private sector		58	41%	19	15	14	5	3	2
	Research		70	50%	32	12	9	6	8	3
	Government		13	9%	3	1	3	4	0	2
Expertise	SM expertise	Security engineering G_{seng}	43	30%	16	10	3	5	6	3
		Security management G_{smgt}	46	33%	12	14	8	9	2	0
		Risk management G_{rmgt}	31	22%	7	10	5	6	2	0
		Other security G_{os}	55	39%	17	12	9	7	6	4
	SM expertise (any of the above) G_s		112	79%	37	25	20	15	9	6
	Non-SM expertise G_{ns}		29	21%	20	3	6	0	2	1
SuI thought in response	*Tech*: technical system, product or service		41	29%	15	6	10	4	5	1
	Org: organization		5	4%	2	1	0	2	0	0
	Hol: holistic picture consisting of elements of *Tech* and *Org*		89	63%	36	19	14	9	6	5
	Oth: other		6	4%	1	2	2	0	0	1

The prioritization order is approximately the same for the respondent group with SM expertise G_s ($n_s = 112$) as for all G_{all} ($n_{all} = 141$). This is expected, as G_{ns} is comprised of persons with engineering background, concerned mainly with technical metrics that typically incorporate less need for interpretation.

Fig. 1 shows the rankings in the G_s sub-groups G_{seng}, G_{smgt}, G_{rmgt}, and G_{os}. Unlike others, risk management group G_{rmgt} scores measurability as the highest priority, together with correctness.

3.3.2. Clustering

The similarities among the experts' rankings were tested by investigating the presence of expert opinion *prioritization profile clusters*. The quantified results were clustered using the non-hierarchical *k*-means clustering method (MacQueen, 1967), which aims to partition *n* answers into *k* clusters by minimizing the Within-Cluster Sum of Squares ($WCSS_{\overline{\Omega}}$) given by the following equation:

$$WCSS_{\overline{\Omega}} = \arg\min_{R} \sum_{m=1}^{k} \sum_{\mathbf{o}_p \in C_i} \left\| \mathbf{o}_p - \mu_m \right\|^2, \quad (5)$$

where $\mathbf{o}_p$ is a 19-dimensional vector of normalized prioritization opinions $\overline{\Omega}_{i,j}$ for all 19 quality criteria j ('a prioritization profile for a respondent with index i'), and μ_m is the mean of normalized ranks in cluster C_p, $1 \le p \le k$, $G_{all} = \{C_1, C_2, \ldots, C_k\}$.

The optimal number of clusters was tried between two and 32 clusters, using silhouette, cluster distance, and centroid distance algorithms. The optimal number in all cases was 32 clusters. To simplify the interpretation, a fixed number of clusters, $k = 6$, was chosen, because most of the rankings were observed to have similar patterns and the rest were more or less individual prioritizations. Table 8 shows the ranks of quality criteria calculated based on the means of the ranks of each cluster $\mu_{j,C_k}(\overline{\Omega})$, $k \in \{1,2,3,4,5,6\}$.

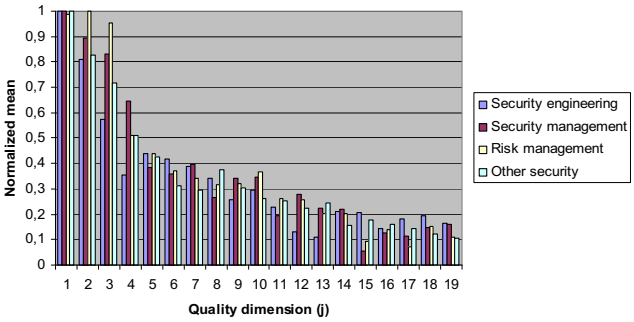

Fig. 1 – Quality dimension priorities based on the means of normalized ranking in different sub-groups of security expertise.

The clustering results show different layers of experts' opinions. The biggest clusters, C_1, C_2, and C_3, which represent the majority (108) of the respondents, rated all the foundational criteria and objectivity as the highest priorities, with a slightly different order in each cluster. These are denoted as 'CMM Clusters.' The 'CMM pattern' or at least the 'CM pattern' is visible in all clusters. In the following discussion, usability refers to effectiveness, efficiency, and cost effectiveness.

The identified clusters based on raw data can be characterized as follows:

1. C_1 – '*Researcher CMM Cluster with correctness emphasis*' ($n = 54$): Accentuates correctness, objectiveness, and unbiasedness – closely related criteria to correctness. Usability is not rated high. This cluster has more respondents from research and fewer from private companies and governments.
2. C_2 – '*CMM Cluster with measurability and usability emphasis*' ($n = 28$): Rates measurability as the top priority, followed by

Table 8 – Prioritization results of all respondents.

j	Quality dimension	μ_{all}	C_1		C_2		C_3		C_4		C_5		C_6	
			p_1	s_1^2	p_2	s_2^2	p_3	s_3^2	p_4	s_4^2	p_5	s_5^2	p_6	s_6^2
1	Correctness	0.528	**1**	0.2	4	7.5	**2**	12.4	8	0.7	7	5.5	8	32.3
2	Measurability	0.378	**3**	19.6	**1**	1.2	**3**	22.5	15	18.2	10	12.6	**3**	8.3
3	Meaningfulness	0.373	4	14.1	**3**	21.9	**1**	0.1	12	19.5	5	33.5	10	30.3
4	Objectivity, unbiasedness	0.260	**2**	12.8	**2**	33.5	4	18.1	**1**	31.1	15	9.0	18	19.0
5	Reproducibility, repeatability, scale reliability	0.204	5	25.4	6	19.3	5	17.5	**3**	5.3	11	22.6	5	13.6
6	Effectiveness	0.168	10	23.2	10	18.4	12	12.1	7	8.6	**1**	0.5	17	4.0
7	Comparability	0.164	7	21.4	5	28.1	11	17.7	9	22.4	19	4.7	4	12.7
8	Attainability, availability, easiness	0.160	6	19.8	7	21.1	6	9.8	13	17.3	6	24.7	6	12.2
9	Clarity, succinctness	0.143	11	23.3	12	12.7	7	24.9	**2**	24.0	9	35.2	**1**	25.8
10	Cost effectiveness	0.140	15	19.7	9	26.4	8	29.2	5	34.9	**2**	18.2	14	31.0
11	Representativeness, contextual specificity	0.113	14	21.7	8	27.0	9	31.1	11	16.3	8	28.2	13	19.0
12	Controllability	0.107	8	19.4	15	19.8	13	24.8	6	6.2	12	26.0	12	25.2
13	Ability to show progression	0.089	17	29.3	16	20.4	10	32.9	4	25.7	17	26.2	11	44.3
14	Efficiency	0.084	16	18.0	11	18.9	14	16.0	17	8.9	4	24.6	9	12.7
15	Non-intrusiveness	0.081	12	34.3	18	23.2	17	25.8	19	0.9	**3**	41.8	16	52.0
16	Completeness	0.076	9	39.6	17	18.8	16	23.7	16	17.4	18	30.9	7	40.3
17	Scalability, portability	0.071	19	17.2	19	8.7	15	15.8	14	47.3	13	31.5	**2**	1.2
18	Granularity	0.062	18	21.6	14	25.7	18	27.9	10	49.8	14	36.8	15	30.0
19	Time dependability	0.055	13	28.7	13	23.5	19	17.3	18	8.9	16	25.5	19	0.3

The bold entries are used to emphasize the three most important quality criteria in each cluster (i.e. all rank numbers 1, 2 and 3).

objectivity and meaningfulness. The ranking in this cluster is very similar to C_1, but representativeness and contextual specificity, as well as usability, are seen as more important. The emphasis on usability is somewhat expected because this cluster has most respondents from private companies.

3. C_3 – '*CMM Cluster with meaningfulness emphasis*' ($n = 26$): Ranks meaningfulness as the most important. Clarity and succinctness, both related to meaningfulness, are also relatively highly ranked. Similarly to C_2, cost effectiveness is emphasized. The cluster has fewer respondents from research.
4. C_4 – '*Unbiasedness Cluster with underlying CMM pattern*' ($n = 15$): Does not explicitly rate foundational criteria highly. Instead, it emphasizes closely related criteria: objectivity and unbiasedness (related to correctness), followed by clarity and succinctness (related to meaningfulness), and then in third place by reproducibility, repeatability, and scale reliability (related to measurability). Evidently, the respondents of C_4 concentrated more on detailed criteria than the more abstract CMM criteria. This cluster comprises a higher number of respondents from governments than from private companies and research.
5. C_5 – '*Usability Cluster with underlying CMM pattern*' ($n = 11$): Emphasizes effectiveness, cost effectiveness, non-intrusiveness, and efficiency, all of which are related to usability. In this cluster, the 'CMM pattern' is prioritized directly below usability: meaningfulness, attainability, availability and easiness, closely related to measurability, and, finally, correctness. C_5 has a higher number of respondents from research, than from private companies, and none from governmental institutions. Interestingly, experts from private companies generally do not seem to emphasize the usability criteria as much as C_5 does.
6. C_6 – '*Remainder Cluster*' ($n = 7$): Does not explicitly prioritize correctness and meaningfulness. However, clarity and succinctness, which are closely related to meaningfulness, rank at the top, followed by measurability. It should be noted that the variance of opinions within this cluster is high, making it present the 'remainder' of the opinions. The number of respondents in C_6 is quite even between research, private companies, and governments.

3.4. *Validity and reliability*

There are some factors that affect to the validity and reliability of the EOS results, either confirming or not confirming the results. In the following, we discuss these factors and related assumptions.

3.4.1. *Selection of respondents*

SM and security metrology are still in their infancy, due to various reasons discussed in detail in (Savola, 2010). There are therefore not that many *experts* in this area. The persons who work with or are interested in SM tend to be SM researchers, other security researchers, industrial security engineering and management practitioners, and risk management experts. The EOS contacted and received responses from all of these groups. The main available forums mentioned above were used to reach the respondents. It is difficult to estimate the total number of SM experts globally. Only a few workshops are arranged under the topic annually, with less than 100 participants in total. Only tens of researchers worked on the topic in the research projects mentioned in Section 3.1. The most widely known resource for SM discussion was the securitymetrics.org e-mail discussion list with about 800 subscribers (Securitymetrics.org, 2011) during the EOS execution. Many of the subscribers are not SM experts, and it cannot be claimed that all SM experts subscribe to the list. However, the number of subscribers gives a rough estimate of how narrow the current activity is.

Based on these observations, the total number of SM experts globally can be assumed to be over 1000 persons but not over 10,000 persons. This is far fewer than the number of experts in the *information security field as a whole*. Taking into account this estimate, a considerable number of SM experts responded to the survey. It would have been very difficult to obtain more responses than those that are included. We assume that the results widely represent the current state-of-the-art research knowledge and state-of-practice knowledge of SM.

3.4.2. *Expertise in security metrics for carrying out prioritization*

Little published empirical evidence exists about the deployment of SM. There are therefore many degrees of freedom in the current knowledge on SM and their quality criteria. In addition, it is difficult to define widely accepted criteria for an SM expert. The selection criteria presented earlier were not too strict, allowing the inclusion of enough experts. However, the criteria ruled out persons without SM knowledge. Very strict criteria were not suitable, in consideration of the low maturity of the field. In the EOS, it is assumed that the current average expertise level of SM researchers and practitioners is sufficient to be able to carry out prioritization of SM quality criteria. Moreover, it is assumed that the experts did not overestimate or underestimate the accuracy of their beliefs, used the most recent available knowledge they had, and did not misunderstand the questionnaire (linguistic uncertainty).

3.4.3. *Cultural factors*

Although the majority of the respondents were from Finland and the U.S., most of them work in global businesses. The responses are clearly based on global views, not local ones. The only cultural phenomenon discovered from the data is that the respondents from the U.S. tend to prioritize only the most important criteria, whereas most Finnish respondents prioritize them all. This does not have any significant effect on the results because of the exponential weighting of the ranks.

3.4.4. *Consensus on foundational quality criteria*

With regard to the first three ranks, the foundational quality criteria of SM, there is a sufficient degree of consensus among SM experts. Clusters C_1, C_2, and C_3 represent the majority (77%) of the respondents. Table 9 elaborates on the reasoning behind the high priority of the foundational quality criteria. In the table, Variance 's^2' is denoted as 'very low' if $s^2_{j,C_x}(\Omega) < 1.5$, 'low' if $1.5 \leq s^2_{j,C_x}(\Omega) < 10.0$, 'medium', if $10.0 \leq s^2_{j,C_x}(\Omega) < 20.0$, and 'high' if $s^2_{j,C_x}(\Omega) \geq 20.0$.

3.4.5. *Partial consensus on usability*

Usability criteria – effectiveness, efficiency, and cost effectiveness – are ranked top by C_5, effectiveness with very low

variance. Moreover, C_4 ranked effectiveness relatively high with low variance. Effectiveness was rated better than mediocre by all the remaining clusters except C_6, which ranked it low. Cost effectiveness was ranked second by C_5 and fifth by C_4. C_2 and C_3 rated it better than mediocre. Efficiency was ranked fourth by C_5 and fifth by C_4. C_2 and C_6 rated it better than mediocre. These results indicate *partial consensus* on the importance of usability criteria.

3.4.6. *Ranking disagreement of other criteria*

Kendall's coefficient of concordance W (Kendall and Babington Smith, 1939) can be used for assessing agreement among raters, ranging from 0 (no agreement) to 1 (complete agreement). W is defined as (Kendall and Babington Smith, 1939):

$$W = \frac{12S}{m^2(n^3 - n)}, \tag{6}$$

where $S = \sum_{j=1}^{n}(R_j - \overline{R})$ is the sum of the squared deviations, $\overline{R} = 0.5 \cdot m(n+1)$ is the mean value of the total ranks, $R_j = \sum_{i=1}^{m} \Omega(i,j)$ is the total rank given by experts with index number i to the quality dimension of index j, m is the number of experts, and n is the total number of quality criteria ($n = 19$). The calculation of W is affected by all 19 prioritization decisions. Many respondents did not prioritize all of the quality criteria, focusing instead on the most important ones. As the calculation of W assumes that all the experts made decisions for all the quality criteria, the rankings with missing information had to be omitted from this calculation.

Table 10 summarizes the results of the W calculations, typical findings of which should be between 0.65 and 0.90. The results clearly show that the degree of concordance is generally low. W calculated for all applicable responses was 0.28, indicating ranking disagreement. The W calculation is also confirmed by the fact that the optimal cluster count was found to be the largest one, 32. However, the consensus on foundational quality criteria and partial consensus on usability still holds. This leads to the conclusion that there is a lack of strong agreement among the experts for prioritization of quality criteria *other than* the foundational criteria and usability.

3.4.7. *Routine listing and weighting*

It was apparent from the raw results that many respondents started to use *routine listing* after the sixth or seventh prioritization decision. In practice, some experts arranged letters representing rank in alphabetical order from that point on. Consequently, only the first six quality criteria can been seen as prominent. This challenge was compensated for by the choice of exponential pre-distribution of prioritization decisions in Eq. (1).

3.4.8. *Interdependencies between quality criteria*

There are interdependencies between the quality criteria. In fact, it would be impossible to find a complete set of quality criteria with no interdependencies. Obviously, a thorough investigation of the interdependencies would result in a complicated analysis and would require investigation of special cases of SM deployment. For this reason, a generic formal analysis of interdependencies would not be valuable. However, we are interested in whether the main outcomes of the EOS are affected by the interdependencies.

An investigation of the interdependencies between correctness and other criteria connected to it results in the following reasoning. If the other criteria incorporating interdependencies with correctness from Table 8 were neglected, correctness alone would be ranked as the most dominating criterion. Moreover, taking into account the interdependencies *emphasizes even more* the role of correctness, because none of the other criteria affects correctness in a

Table 9 – Reasoning behind the consensus of the foundational quality criteria.

Correctness	Measurability	Meaningfulness
• In G_{all}, correctness is clearly ranked as the first priority • In C_1, it is the first priority, with 'very low' s^2 • In C_3, it is the second priority, with 'medium' s^2. Ignoring two rankings makes the s^2 'low' • In C_2, rank number 4 with 'low' s^2 is assigned to it • The second priority in C_1 and C_2, and the first priority in C_4, objectivity and unbiasedness ($j = 4$), respectively are closely related to correctness • In C_4, C_5, and C_6, correctness is ranked better than mediocre priority, and in C_4 with 'very small', in C_4 with 'small' s^2.	• In G_{all}, measurability is ranked as the second priority • In C_2, it is the first priority, with 'very low' s^2 • In C_1 and C_3, it is the third priority, with 'medium' s^2. Ignoring a few responses in C_1 and C_2 changes their s^2 to 'low' • In C_4, reproducibility, repeatability, and scale reliability ($j = 5$), measurability-related criteria are the third priority, with 'low' s^2 • In C_1, C_3, and C_6, $j = 5$ is ranked relatively high • In C_1, C_3, C_5, and C_6, attainability, availability, and easiness ($j = 8$), all except the last measurability-related criteria, are ranked relatively high	• In G_{all}, meaningfulness is ranked as the third priority. • In C_3, it is the first priority, with 'very low' s^2. • Ignoring a few rankings in C_1, C_2, and C_5 changes s^2 to 'low' or 'very low' for meaningfulness (C_2: 3rd priority, C_1: 4^{rh} priority, and C_5: 5th priority) • In C_6, clarity and succinctness ($j = 13$) is ranked as the first priority, with 'high' s^2. Neglecting one ranking makes s^2 'low' • In C_4, $j = 13$ is ranked as the second priority, with 'high' s^2. Neglecting one ranking makes s^2 'low' • In C_6, $j = 7$ is ranked fourth, and in C_2 fifth, and in C_4, $j = 13$ is ranked fourth.

Table 10 – W values for different clusters.

Result data	α					
	C_1	C_2	C_3	C_4	C_5	C_6
W	0.378	0.332	0.382	0.605	0.483	0.858
Number of total respondents in cluster	54	28	26	15	11	7
Number of respondents used in W calculation	37	11	16	3	4	2

degrading way. In the survey results, unbiasedness and objectivity are ranked quite high but they cannot be chosen as fundamental criteria because they are strongly related to correctness and the ranks of other foundational criteria are higher. The conclusion is that degrees of interdependencies between correctness and other criteria related to it do not affect the result that correctness is ranked as the highest quality criterion. An investigation of interdependencies with measurability and meaningfulness results in similar conclusions as in the case of correctness: if other criteria incorporating interdependencies were neglected, measurability alone would be ranked as the second dominating criterion after correctness, with meaningfulness third.

The original list of criteria did not include usability. However, its sub-criteria were included. These criteria *together* are given a high priority in some prioritization clusters, but low in others. Obviously, the needs for usability of SM are more subdued than for CMM criteria, but still important. Sub-criteria of usability incorporate interdependencies with each other. This does not alter the main conclusions from the survey concerning usability being an important quality criterion, with partial disagreement among the experts.

4. Expert interview study and conceptual quality criteria model

To validate the results from the EOS and to learn more about the reasons for the disagreement among the experts in the EOS, a semi-structured EIS of 21 interviewees was conducted between October 2011 and August 2012. In this section, we also draw conclusions about the original hypothesis and introduce a conceptual quality criteria model for SM.

4.1. *Interviewees and structured questions*

15 Interviewees were chosen based on the clustering results, with three experts from each cluster C_1–C_5. There was at least one year between the original responses and the interview, in some cases one and a half years, to increase the objectivity. Respondents of the 'remainder cluster' C_6 were not interviewed. Moreover, six interviewees who were not part of the EOS were selected. The same selection criteria were enforced as for the EOS. The interviewees were divided into *IGs* (Interviewee Groups) in the following way: IG_x was formed from respondents of Cluster C_x, (x = 1,2,3,4,5), and IG_6 of the six interviewees not part of the EOS. Before the actual interview, IG_6 was asked to carry out rigorously the same prioritization task as was done by the respondents of the EOS.

Table 11 summarizes the structured interview questions. Interview questions IQ1–IQ3 were asked before the overall and clustered results were revealed. IQ4–IQ6 were asked after this.

4.2. *Findings*

There are two main findings from the EIS: (i) the real existence of C_1–C_5 was supported by strong opinions in favor of 'own' cluster prioritizations, and varying degrees of surprise of other clusters' opinions, and (ii) the exact priority of correctness, measurability, meaningfulness, and usability ('CMM + U') depends on the phase of the SM lifecycle (SM definition, preliminary use, and established use).

The main difference in opinions was between respondents from IG_1 and IG_5, the former emphasizing CMM criteria and the latter usability, as illustrated by the following quotes from some interviewees:

- "*The reason why I did not rank correctness as the most important criterion is that I think that SM are subjective anyway.*" (an interviewee from IG_5).
- "*The importance of CMM is intuitively understandable. In the overall results, it is odd that effectiveness is not ranked high.*" (IG_5)
- "*The existence of C_5 is strange. I suspect that the respondents belonging to this cluster have not thought the criteria in a profound way.*" (IG_1)
- "*In my opinion, clusters C_1–C_4 are easy to understand. However, C_5 is not. You cannot do much with SM that is not correct.*" (IG_1)

The real-world existence of clusters C_1–C_5 was supported by the interviews, for the first few ranks (up to 6th–7th). Moreover, it is evident that experts in favor of CMM (most clearly IG_1–IG_4) consider mainly the development phase of SM, and experts emphasizing usability (IG_5) consider the actual deployment of SM or using SM as a tool. It becomes

Table 11 – Interview questions.

Question	IG
IQ1: *Would you change the original prioritization you made? How? Why?*	IG_1–IG_6
IQ2: *How would you justify the prioritization you made?*	IG_1–IG_6
IQ3: *How confident are you of your prioritization?*	IG_1–IG_6
IQ4: *How would you explain the overall results from the EOS?*	All
IQ5: *How would you explain the clustering results?*	All
IQ6: *Do you think that the results can be used for better SM? How?*	All

clear in the interviews that 'CMM + U' are needed in different phases of the SM lifecycle. IG_2 identifies itself as quite close to C_1, with the exception of the first priority. In some interviews, the need for both CMM and usability was recognized, but usability was generally seen as more flexible than the CMM criteria. The IG_3 interviewees think that their opinions are between the most distant opinions of C_1 and C_5. Interestingly, all the interviewees of IG_3 emphasized *iteration* in the SM definition and use, whereas no other interviewees mentioned it. Obviously, iteration helps to achieve meaningfulness, ranked high by C_3. C_4 is an interesting cluster, focusing on 'trees' rather than on the 'forest' in its prioritization. IG_4 criticized the fact that objectivity and unbiasedness, more detailed criteria related to correctness, were not emphasized enough by other clusters than C_4. The background of IG_4 is specific: all the interviewees explained that they were engineers or government employees with some research mindset: "*I found it difficult to choose between 'researcher' and 'private company' in the background part.*"

In general, the interviewees did not change their main rankings as a reaction to Q1, except two who originally ranked effectiveness as the main priority and changed correctness to a higher priority during the EIS. This reflects even stronger consensus on the importance of CMM. Many respondents, however, suggested a new ordering for rankings 7 and below. This can be interpreted to reflect the difficulty of prioritizing more than 6–7 main rankings, confirming the ranking disagreement outcome of other criteria than the foundational ones from the EOS. All the respondents were either confident or fairly confident of their most important prioritizations (Q2).

The interviewees of IG_6 carried out the prioritization with rigor, as shown in one of their comments: "*I spent a considerable amount of time carrying out the prioritization, with several iterations.*" No actual differences were detected in their prioritizations compared with those of others: all new prioritizations represented typical profiles of the clusters discussed earlier. Two new prioritizations had profiles of C_3 and C_5, and one each of C_1 and C_4.

According to several interviewees (Q6), the results can be used for better SM, because it is possible to develop dedicated SM according to preferences of the EOS clustering results. The overall prioritization results can be used as the cookbook of SM development and deployment.

In general, time dependability was seen as the least important quality criterion. Four interviewees, however, criticized this, raising it to 'mid-category'. Many interviewees reiterated the current lack of widely accepted SM concepts, stating that if more relevant information became available, they might change their prioritization.

H asserts that a few foundational generic qualitative properties dominate in most SM. Based on the combined results of the EOS and the EIS, it can be concluded that correctness, measurability, meaningfulness and usability clearly dominate in SM. Therefore, the results provide support for *H*.

4.3. *Conceptual security metrics quality criteria model*

Based on the results from the literature survey, the EOS, and the EIS, we propose a conceptual SM quality criteria model in Fig. 2, visualizing the 'CMM + U' as the main quality criteria for SM (Criteria 1–4), but still incorporating relevant interdependencies with other related criteria. The main interdependencies are shown with solid lines and other interdependencies with dashed lines. The exact priority order of the foundational criteria depends on the SuI and the SM lifecycle. This model can be used as a basis for the feasibility analysis work aimed at increasing the quality of SM. For instance, the model offers input for weight factor determination when analyzing the relational importance of different SM quality factors.

5. Discussion

The results from the EOS indicated that correctness, measurability and meaningfulness are the core quality criteria for SM. Their importance was supported by good variance characteristics of the opinions. Moreover, SM experts have a *partial* consensus that usability criteria are important to SM. The EIS confirmed these results. Due to the factors of deteriorating validity and reliability of the EOS, the ordering of the rest of the quality dimensions is not statistically significant because there is disagreement on their prioritization.

According to the discussions during the EIS, the importance of the 'CMM + U' criteria is justified by the following reasoning: Firstly, if an SM does not correctly represent a measurable attribute, it cannot be relied upon; that is, it is not credible. Secondly, if measurement data cannot be achieved from the SuI, there are not sufficient data to use the SM. Thirdly, if an SM and related measurements are not meaningful, the metric is not credible and does not offer sufficient data for the decision-making. The usability of SM was 'raised' to the same level as the foundational criteria, because, in practice, the usability of SM is important in order for them to be used in the first place. Within the hectic schedules of a business, not all relevant security practices and controls are deployed, let alone SM with poor usability.

The EOS revealed certain expert opinion prioritization clusters, all of which at least partly incorporate the foundational criteria. Due to the validity and reliability constraints, the real-world existence of these prioritization clusters was not fully justified based on the results of the EOS. However, the EIS results supported the real existence of C_1–C_5, if the first few ranks were considered. Meaningful conclusions cannot be drawn based on the small 'remainder' cluster, C_6. Its existence was expected as in any opinion survey.

With the current state of SM research, a survey with a limited number of SM expert opinions was seen more valuable than a survey carried out among security experts in general with a statistically more significant population. Therefore, the number of respondents to the EOS is sufficient, given the limited number of SM experts. However, interestingly, there were no remarkable differences in the opinions of the SM experts and the control group consisting of security experts. This refers to the fact that no remarkable metrics-specific knowledge was imparted by the SM experts compared with the non-SM security experts. This can also be interpreted as all the respondents having sufficient security expertise; their understanding of security evidence quality is quite uniform. This is a good basis for SM development: the SM should not control security, but rather offer systematic support for this

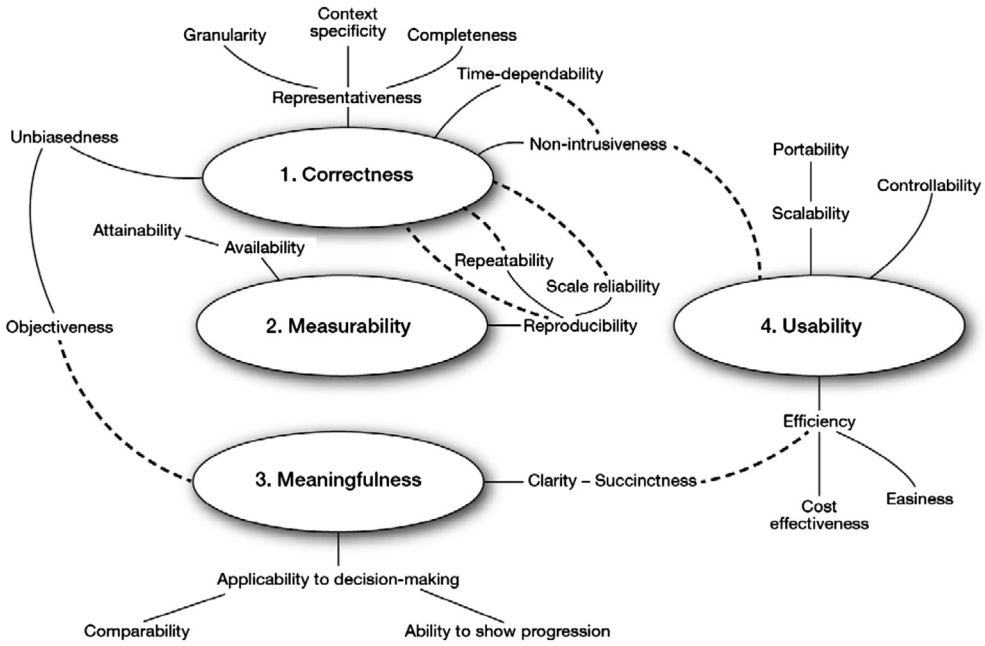

Fig. 2 – A simplified model of SM quality criteria with some of their interdependencies.

task. The overall similarity of opinions between SM experts and security experts indicates that the SM quality dimension analysis does not necessarily require SM development experience. The security practice background did not have a relevant effect on the results either. However, some interesting findings were discovered with respect to opinions, for instance, the very strong need for iteration required by a specific opinion cluster. As no one other than security experts was involved in the EOS and EIS, the results are only applicable to SM, not to other types of metrics.

The results of this study can be used for the development of more feasible SM. The differences in experts' opinion can also guide the development of specific tailored metrics. It is evident that analyzing the SM quality criteria is a challenging task, because they cannot be understood in an absolute way and there is always room for interpretation. The study should therefore be seen as an initial step toward understanding what kind of SM should be developed to make them useful for practical purposes.

6. Conclusions and future work

We presented the results from a quantitative security metrics expert survey and an expert interview study on the prioritization of 19 main quality criteria of security metrics identified in the literature. The interviews were used to validate the survey results and to obtain further information on the findings. The combined results indicate that security metrics experts consider correctness, measurability, and meaningfulness to be foundational quality criteria for security metrics in general. Moreover, usability is ranked high and, in practice, should be emphasized as a foundational criterion. Furthermore, the study revealed certain clusters of quality dimension prioritization, all of which at least partly incorporate the above-mentioned criteria.

Based on the results, we proposed a conceptual security metrics quality criteria model. The model emphasizes correctness, measurability, meaningfulness and usability, yet incorporates relationships between them and relevant subcriteria.

This study offers a high-level view of the quality of security metrics. However, depending on the security objectives of specific use cases, there can be differences, especially in the prioritization order of other quality criteria than the foundational ones. There are plans to obtain more detailed case-specific security expert opinions based on the findings of this study. Not much empirical evidence on the deployment of security metrics has been published. The availability of this kind of evidence would help to increase the experts' knowledge considerably.

Acknowledgments

The work presented here has been carried out in three research projects: GEMOM FP7 (2008–2010), jointly funded by the European Commission and VTT; SASER-Siegfried Celtic-Plus project (2012–2015), jointly funded by Tekes and VTT; and ASSET (2012–2015), funded by the Research Council of Norway in the VERDIKT program.

REFERENCES

Atzeni A, Lioy A. Why to adopt a security metric? A little survey Sep. 2005. p. 1–12.

Axelrod R. Risk in networked information systems. Gerald R. Ford School of Public Policy, University of Michigan; 2003.

Bayuk JL. Measuring systems security: an initial security theoretical construct framework. Doctoral thesis. Hoboken, N.J.: Stevens Institute of Technology; 2011.
Böhme R, Freiling FC. On metrics and measurements. In: Eusgeld I, Freiling FC, Reussner R, editors. Dependability metrics, LNCS 4909. Springer-Verlag; 2008. p. 7–13.
Böhme R, Reussner R. Validation of predictions with measurements. In: Eusgeld I, Freiling FC, Reussner R, editors. Dependability metrics, LNCS 4909. Springer-Verlag; 2008. p. 14–8.
Building security assurance in open infrastructure, beyond (BUGYO) beyond project. www.celticplus.eu/Projects/Celtic-projects/Call5/BUGYO-BEYOND/bugyo-beyond-default.asp; May 11, 2013.
Dependability Benchmarking (DBench) Project (EU IST-2000-25425). Kanoun K, Madeira H, Crouzet Y, dal Cin M, Moreira F, Ruiz García J-C, editors. DBench dependability benchmarks 2004. 235 p.
E-mail message from securitymetrics.org list administrator stating that there were 835 subscribers on the list as of March 8, 2011.
Fenton NE, Melton A. Measurement theory and software measurement. In: Melton A, editor. Software measurement. International Thomson Computer Press; 1996. p. 27–38.
Frøkjær E, Hertzum M, Hornbæk K. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? April 1–6, 2000. p. 345–52.
Genetic message oriented secure middleware (GEMOM) EU FP7 project. cordis.europa.eu; May 11, 2013.
Henning R, editor. Proceedings of workshop on information security system, scoring and ranking – information system security attribute quantification or ordering. Williamsburg, Virginia: ACSA and MITRE, May 2001; 2002.
IEEE Standard 1061-1998. Standard for a software quality metrics methodology. Piscataway, N.J.: Institute of Electrical and Electronics Engineers Standards Department; 1998.
Information technology security evaluation criteria (ITSEC), Version 1.2. Commission for the European Communities; 1991.
ISO/IEC 21827:2003. Information technology – systems security engineering – capability maturity model (SSE-CMM). International Organization for Standardization and the International Electrotechnical Commission; 2003.
ISO/IEC Guide 99:2007. International vocabulary of metrology – basic and general concepts and associated terms (VIM). International Organization for Standardization and the International Electrotechnical Commission; 2007.
Jansen W. Directions in security metrics research. NISTIR 7564. U.S. National Institute of Standards and Technology; Apr. 2009. 21 p.
Jaquith A. Security metrics – replacing fear, uncertainty and doubt. Addison-Wesley; 2007. p. 306.
Jaquith A. Proving your worth – follow these steps to create a successful security metrics program. Information Security March 2010;12(2):29–33.
Jelen G. SSE-CMM security metrics Jun. 13–14, 2000.
Jones JA. An introduction to factor analysis of information risk (FAIR), Risk management insight 2005.
Kaner C, Bond WP. Software engineering metrics: what do they measure and how do we know?. In: 10th International software metrics symposium, Chicago Sept. 2004.
Kendall MG, Babington Smith B. The problem of *m* rankings. The Annals of Mathematical Statistics 1939;10(3):275–87.
MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Berkeley symposium on mathematical statistics and probabilityvol. 5. University of California Press; 1967. p. 281–97.
Managing assurance, security and trust for sERvices (MASTER) EU FP7 project. cordis.europa.eu; May 11, 2013.
Pironti JP. Developing metrics for effective information security governance. Information Systems Control Journal 2007;2.
Rathbun D. Gathering security metrics and reaping the rewards. SANS Institute Information Security Reading Room; Oct. 7, 2009. 21 p.
Rouhiainen V. The quality assessment of safety analysis. Doctoral thesis. Espoo, Finland: VTT Technical Research Centre of Finland; 1990.
Savola R. A security metrics taxonomization model for software-intensive systems. Journal of Information Processing Systems Dec. 2009;5(4):197–206.
Savola R, Heinonen P. A visualization and modeling tool for security metrics and measurements management Aug. 15–17, 2011. 8 p.
Savola R, Frühwirth C, Pietikäinen A. Risk-driven security metrics in agile software development – an industrial pilot study. Journal of Universal Computer Science Sept. 2012;18(12):1679–702.
Savola R. On the feasibility of utilizing security metrics in software-intensive systems. International Journal of Computer Science and Network Security Jan. 2010;10(1):230–9.
Schechter SE. Computer security strength & risk: a quantitative approach. PhD thesis. Cambridge, Massachusetts: Harvard University; May 2004.
Smart objects for intelligent applications (SOFIA) ARTEMIS project. cordis.europa.eu; May 11, 2013.
Verendel V. Some problems in quantified security. Licentiate thesis. Gothenburg, Sweden: Chalmers University of Technology; 2010.
Westen D, Rosenthal R. Quantifying construct validity: two simple measures. Journal of Personality and Social Psychology 2003;84(3):608–18.
Williams JR, Jelen GF. A framework for reasoning about assurance. Arca Systems, Inc; 1998. p. 43.
www.securitymetrics.org; May 11, 2013.

Mr. Reijo Savola received the degree of M.Sc. from the University of Oulu, 1992, and the degree of Licentiate of Technology from the Tampere University of Technology, 1995. He is an author of 108 journal, conference, and workshop papers and has been chair, program chair, and a member of technical committees several conferences. His current research interests include security metrics and security requirements engineering. He is currently working as a Principal Scientist at VTT Technical Research Centre of Finland. Previously he has worked as a software engineer for Elektrobit Group Plc. in Oulu, Finland and in Redmond, WA, United States.