Validating a Set of Candidate Criteria for Evaluating Software Tools and Data Sources for National CSIRTs' Cyber Incident Responses

SHARIFAH ROZIAH BINTI MOHD KASSIM*, Institute of Cyber Security for Society (iCSS) & School of Computing, University of Kent, UK and CyberSecurity Malaysia, Malaysia
SHUJUN LI and BUDI ARIEF, Institute of Cyber Security for Society (iCSS) & School of Computing, University of Kent, UK

National Computer Security Incident Response Teams (CSIRTs) are established worldwide to coordinate responses to cyber security incidents at the national level. It is known that software tools (including open-source ones) and public data are routinely used to facilitate incident response in national CSIRTs. However, there is a lack of an authoritative set of criteria that can be used for a systematic evaluation to decide which software tools and data sources should be used by national CSIRTs for incident response. A prior study identified a set of potential candidate criteria for such an evaluation. The study presented in this paper aims to validate these candidate criteria empirically by asking staff members of several national CSIRTs how they perceive the candidate criteria's practical usefulness and readiness for deployment in national CSIRTs' operations. The study involved online semi-structured interviews with nine interviewees from nine national CSIRTs in Asia-Pacific, Africa, and Europe. After validating the candidate criteria using semi-structured interviews with these nine interviewees, we applied the criteria to evaluate a selection of software tools and data sources by converting each criterion into one or more relevant metrics, such as "measuring the time taken by a tool to produce results". Results from the study led to the following main findings: 1) all interviewees perceived the candidate criteria as practically useful for evaluating tools and data sources in the operations of national CSIRTs; 2) all interviewees agreed that the candidate criteria could be deployed in national CSIRTs and other types of CSIRTs; and 3) the candidate criteria can be applied relatively easily in practice. These criteria are envisaged to help national CSIRTs select the most appropriate tools and data sources to facilitate effective incident response, improve their operational practices, and improve the quality of wider security operations.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; *Usability in security and privacy*;

Additional Key Words and Phrases: CSIRT, computer security incident response team, CERT, incident response, cyber incident response, national CSIRT, semi-structured interview, tools evaluation, data evaluation, criteria

1 INTRODUCTION

National CSIRTs (computer security incident response teams) are a distinct category within the broader CSIRT landscape, and many countries and regions worldwide have formed their national CSIRTs [55]. National CSIRTs allow countries and regions to respond to and coordinate incidents at the national and regional level through a centralised contact point more quickly and systematically, empowering a wide range of stakeholders to learn from experience and build cyber security resilience [26]. They typically use tools and data from various sources to support incident responses. Such tools and data are paramount for national CSIRTs in general *to perform effective and efficient incident responses* [5, 27]. Using good-quality tools and data to facilitate incident responses in national CSIRTs is essential, which can be achieved through evaluating and implementing tools [5] and significantly increasing the effectiveness of national CSIRTs' operation. Due to their significant role at the national level, national CSIRTs are chosen for the study over sectoral or organisational CSIRTs.

*Sharifah Roziah Binti Mohd Kassim did the work during her PhD study at the University of Kent, UK.

Authors' addresses: Sharifah Roziah Binti Mohd Kassim, Institute of Cyber Security for Society (iCSS) & School of Computing, University of Kent, Canterbury, Kent, CT2 7NS, UK, roziah@cybersecurity.my and CyberSecurity Malaysia, Cyberjaya, Selangor, 63000, Malaysia; Shujun Li, s.j.li@kent.ac.uk; Budi Arief, b.arief@kent.ac.uk, Institute of Cyber Security for Society (iCSS) & School of Computing, University of Kent, Canterbury, Kent, CT2 7NS, UK.

Despite the importance of tools and data for national CSIRTs, very few studies exist concerning the adoption of free tools and public data in national CSIRTs' operations; one such area is evaluating free software tools and public data sources for quality and usability purposes to support the operations of national CSIRTs. This presents a research gap. Shedding light on this research gap can be helpful, especially in ensuring appropriate free tools and public data are selected through systematic evaluation, subsequently enhancing the operations in national CSIRTs [15, 50].

Furthermore, Mohd Kassim et al. [39] investigated the use of free software tools and public data by surveying and interviewing staff members from multiple national CSIRTs, and they found that such tools and data are less often evaluated for quality purposes than commercial ones, indicating the need to develop processes and criteria for evaluating such tools and data sources. They further investigated how free software tools and public data sources are currently evaluated in national CSIRTs in a follow-up study based on focus group discussions with staff members of national CSIRTs [40], which led to a set of candidate criteria for evaluating free software tools and public data sources that can be used by national CSIRTs. However, Mohd Kassim et al. did not conduct any validation of their proposed candidate criteria, nor have we seen other researchers' independent validation. Therefore, the practical usefulness of their candidate criteria remains unknown.

This paper aims to empirically validate the candidate criteria for evaluating software tools and public data sources proposed by Mohd Kassim et al. [40], focusing on their practical usefulness and readiness for deployment in national CSIRTs. For the sake of brevity, in the rest of the paper, we use 'tools' to indicate 'software tools'. We will also focus on free software tools and public data sources only, following what Mohd Kassim et al. did in their past studies [39, 40]. Our work has three research questions (RQs):

- **RQ1**: How do staff members of national CSIRTs perceive the practical usefulness of the candidate criteria for evaluating tools and data sources?
- **RQ2**: How do staff members of national CSIRTs perceive the readiness to deploy the candidate criteria for evaluating tools and data sources in national CSIRTs?
- **RQ3**: How easily can the candidate criteria be applied to evaluate tools and data sources?

To answer the first two RQs, we conducted online semi-structured interviews with nine staff members from nine national CSIRTs in Asia-Pacific, Africa and Europe. To answer the third RQ, we applied Mohd Kassim et al.'s candidate criteria to evaluate two sample tools and one data source widely used by national CSIRTs.

This study revealed three major findings, summarised as follows:

(1) All interviewees perceived the candidate criteria as practically useful for evaluating free tools and public data in the interviewed national CSIRTs.
(2) All interviewees agreed that the candidate criteria could be deployed in the interviewed national CSIRTs and other CSIRTs. Most (eight out of the nine) interviewees would also recommend the criteria to other national CSIRTs for deployment.
(3) The study's attempt to apply the candidate criteria to evaluate two candidate tools and a data source showed that the candidate criteria could be applied in practice to evaluate free tools and public data.

The rest of the paper is organised as follows. Section 2 provides an overview of previous work on the importance of and the current practices in tool and data evaluation. Section 3 explains the methodology used in the study, which includes the data collection and analysis methods. Section 4 presents results from the interviews and evaluations of sample tools and a data source. The results are presented separately to differentiate how the candidate criteria were validated and the results are discussed in Section 5. Section 6 presents the study's limitations and potential future work. Section 7 summarises and concludes the paper.

2 RELATED WORK

In this section, we reviewed several studies related to our research. These include studies about the importance of tools and data evaluation for national CSIRTs and ISO/IEC standards and guideline documents related to our study.

2.1 Importance of Tool and Data Evaluation for National CSIRTs

Tools and data sources are paramount for all CSIRTs to facilitate cyber incident responses [5, 27]. Several national CSIRTs, such as the National Cyber Security Centre in the Netherlands (NCSC-NL), CERT-NZ in New Zealand and CERT.at in Austria, have pointed out the importance of having appropriate tools for operations [40]. Dubois and Tatar [13] even suggested that the lack of efficient tools can cause issues in cyber incident responses, which might lead to more significant security risks. The International Telecommunications Union (ITU) highlighted the importance of CSIRTs and how the establishment and development of CSIRTs should be based on mature models, utilising international collaborations, strong procedures, *effective tools* and training [13].

Several researchers emphasised that having qualified tools and data is crucial in responding to cyber-attacks and incidents more efficiently [15, 50]. The need for quality tools and data is increasing as more and more organisations rely on tools and data in many aspects of their operations [56]. These include national CSIRTs, extensively using software tools (in particular, open-source and free tools), public data, and open-source intelligence (OSINT) to facilitate incident responses [29]. Hence, tools and data must be evaluated systematically by following specific criteria for quality purposes [19]. Such evaluation ensures compliance with security requirements [6] and effectiveness for operations in national CSIRTs [5].

Furthermore, to ensure only qualified tools and data are selected, particularly for incident management and analysis work in national CSIRTs, the evaluation and implementation of tools is necessary [5]. CSIRTs must examine IT devices or software to identify vulnerabilities; such examination is crucial to avoid using unpatched and vulnerable tools in the operations [44]. Therefore, evaluating tools and data in national CSIRTs (and CSIRTs at large) is essential to ensure the team is equipped with quality tools and data to detect, respond to, and mitigate security incidents effectively.

A recent empirical study [40] reported that more than three-quarters of the national CSIRTs who participated did not have a systematic approach to evaluating and selecting qualified tools and data sources for their operations. Furthermore, all participants involved in the study reported a lack of systematic procedures and criteria for evaluating tools and data in their corresponding national CSIRTs. The study also identified that a systematic evaluation of tools and data guided by specific criteria is crucial for national CSIRTs' operations and proposed a set of candidate criteria for such purposes.

Nowikowska [44] made it clear that one of the key reasons for evaluating tools and software is to avoid using vulnerable IT systems that could jeopardise the confidentiality, integrity and availability (CIA) of information. This is reinforced by Bills et al. [5], who emphasised the need for tool evaluation and implementation of the evaluation to ensure that national CSIRTs use efficient and effective tools, especially for incident management and analysis. Iakovakis et al. [22] also pointed out the importance of using effective and qualified tools in CSIRTs' operations to prevent and mitigate cyber-attacks. Though several researchers have started to study the general use of tools in the operational practices of CSIRTs, such as Krstic et al. [32] and Spring and Illari [54], no attempt has been made to study tool and data evaluation in the operation of national CSIRTs. This represents a gap in the literature that we intend to address.

Some work has been done in evaluating data within the CSIRTs' community, precisely on threat data feeds. Pawlinski and Kompanek [49] studied evaluating threat intelligence data feeds in helping users choose qualified threat data feeds available in abundance. They evaluated sample threat feeds using the following factors (criteria): relevance, accuracy, completeness, timeliness and ingestibility with several metrics.

Similarly, Kührer et al. [33] evaluated blocklists data feeds based on the following criteria: vantage, volume, timeliness, accuracy, and completeness using several metrics.

All the above studies highlight the need for effective tools and data to support national CSIRTs' operations. Hence, national CSIRTs must identify qualified tools and data to facilitate cyber incident responses. Identifying such tools and data can only be achieved through systematic procedures. One such procedure involves constructing a set of criteria for evaluating tools and data, which we intend to establish in this study.

2.2 Relevant International and Industrial Standards

The following three international standards are particularly relevant for our study [37]:

(1) ISO/IEC 25010:2011 Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models [24],
(2) ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model [23], and
(3) ISO/IEC 5055:2021 Information technology – Software measurement – Software quality measurement – Automated source code quality measures [25].

The first standard describes principles for assessing software quality [4], the second one for data quality [62], and the third one for quality measures derived from static analysis of a software system's source code and architectural structure [9]. These standards have been widely adopted by researchers for software evaluation purposes [2, 31, 41, 42, 58, 59, 65] and data evaluation [62, 64]. ISO/IEC 25010:2011 is more recent and comprehensive than some older software quality evaluation models such as the McCall model [35], the Boehm model [6], FURPS [18] and the Dromey model [12]. For instance, essential factors or criteria such as *security*, *satisfaction*, *functionality* and *learnability*, which are not covered in the McCall model, the Boehm model, FURPS and the Dromey Models, are addressed in the ISO/IEC 25010:2011 standard.

3 METHODOLOGY

The study aims to validate the candidate criteria for evaluating free tools and public data proposed by Mohd Kassim et al. [40], using two methods: 1) semi-structured interviews with staff members of national CSIRTs to gain insights about their perception on the usefulness of Mohd Kassim et al.'s candidate criteria and their potential for deployment, and 2) simulated applications of Mohd Kassim et al.'s candidate criteria to evaluate two sample tools and a data source widely used by national CSIRTs to identify how the criteria could be converted to more specific quantitative metrics in practical settings. Quantitative metrics are used in the second method because they are more objective, measurable and reproducible than qualitative metrics. They can be used to compare different tools more easily and to evaluate improvements over time.

3.1 Mohd Kassim et al.'s Candidate Criteria

Mohd Kassim et al. derived eight groups of candidate criteria for evaluating free tools – *Usability*, *Maintainability*, *Security*, *Functionality*, *Compatibility*, *Reliability*, *Context Coverage* and *Other*. In our study, we refined the original eight groups of tool-oriented candidate criteria into 14 more fine-grained groups of candidate criteria guided by two broad categories defined in the ISO/IEC 25010:2011 standard [24]: "Product Quality" and "Quality in Use". The "Product Quality" category is about *software's static properties and the computer system's dynamic properties*. It covers the following nine groups of criteria: 1) Security, 2) Usability, 3) Maintainability, 4) Compatibility, 5) Functionality, 6) Performance Efficiency, 7) Reliability, 8) Compliance, and 9) Certification. The "Quality in Use" category is about *the outcome of interaction when a product is used in a particular context*. It covers the following five groups of criteria: 1) Context Coverage, 2) Usability, 3) Effectiveness, 4) Freedom from Risk, and 5) Popularity. The definitions of 32 candidate criteria belonging to the 14 groups of criteria are shown in Tables 1 and 2. Note

Table 1. Definitions of the Nine Groups of Criteria for Evaluating Tools in the "Product Quality" Category

Criteria	Definition
Security	
Confidentiality	The degree to which a product or system ensures that data is accessible only to those authorised to have access
Integrity	The degree to which a system, product or component prevents unauthorised access to, or modification of, computer programs or data
Authenticity	The degree to which the identity of a subject or resource can be proved to be the one claimed
Usability	
Learnability	The degree to which specified users can use a product or system to achieve specified goals of learning to use the product or system with effectiveness, efficiency, freedom from risk and satisfaction in a specified context of use
Operability	The degree to which a product or system has attributes that make it easy to operate and control
User interface aesthetics	This refers to properties of the product or system that increase the pleasure and satisfaction of the user, such as the use of colour and the nature of the graphical design
Accessibility	The degree to which a product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use
Maintainability	
Maintainability	The degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers
Supportability	The degree to which a product or system could provide support and assistance to users when encountering a problem
Analysability	The degree of effectiveness and efficiency with which it is possible to assess the impact on a product or system of an intended change to one or more of its parts, to diagnose a product for deficiencies or causes of failures, or to identify parts to be modified
Modifiability	Modifications can include corrections, improvements or adaptation of the software to changes in the environment and in requirements and functional specifications
Compatibility	
Interoperability	The degree to which two or more systems, products or components can exchange information and use the information that has been exchanged
Functionality	The degree to which the set of functions covers all the specified tasks, appropriateness of the tasks and user objectives
Performance Efficiency	The performance relative to the number of resources used under stated conditions
Time behaviour	The degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements
Capacity	The degree to which the maximum limits of a product or system parameter meet requirements
Money	The degree of how much money is used in relation to the results achieved
Human effort	The degree of how much human effort is used in relation to the results achieved
Material	The degree of how much material is used in relation to the results achieved
Reliability	
Reliability	The degree to which a system, product or component performs specified functions under specified conditions for a specified period of time
Availability	The degree to which a system, product or component is operational and accessible when required for use
Compliance	The degree to which tools comply with a specific policy, rules and regulations and operations
Certification	The degree to which tools are certified and accredited by reputable accreditation and certification bodies

that Usability appears under both "Product Quality" and "Quality in Use" categories since it has sub-criteria falling into these two broad categories. Among the 14 groups of criteria, three are not included in the ISO/IEC 25010:2011 standard: Compliance, Popularity, and Certification.

For evaluating data sources, the original eight groups of candidate criteria derived by Mohd Kassim et al. [40] – *Credibility*, *Confidentiality*, *Currentness*, *Understandability*, *Completeness*, *Precision*, *Accuracy* and *Efficiency* – are

Table 2. Definitions of the Five Groups of Criteria for Evaluating Tools in the "Quality in Use" Category

Criteria	Definition
Context Coverage	
Flexibility	The degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements
Usability	
Satisfaction	The degree to which user needs are satisfied when a product or system is used in a specified context of use
User experience	The degree of users' perceptions and responses that result from the use and/or anticipated use of a system, product or service
Usefulness	The degree to which user needs are satisfied with their perceived achievement of pragmatic goals, including results and consequences of use
Trust	The degree to which the user has confidence that the product will behave as intended
Comfort	The degree to which user needs are satisfied with physical comfort
Effectiveness	The degree to which accuracy and completeness with which users achieve specified goals
Freedom from Risk	
Sustainability	The degree to which a system, product or component is sustainable with freedom from risk – economic risk mitigation, health and safety risk mitigation and environmental risk mitigation
Harm from use	The degree of negative consequences regarding health, safety, finances or the environment that result from the use of the system
Popularity	The degree to which the security community (large or small) uses the tool

Table 3. Definitions of the Eight Groups of Criteria for Evaluating Data Sources

Criteria	Definition
Credibility	The degree to which data has attributes regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, commitments)
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use
Confidentiality	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorised users in a specific context of use
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use
Precision	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users and are expressed in appropriate languages, symbols and units in a specific context of use
Currentness	The degree to which data has attributes that are of the right age in a specific context of use
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of us

based on the ISO/IEC 25012:2008 standard more directly [23], so we kept them as they are in our study. Their definitions are shown in Table 3.

It should be noted that most software evaluation models in the literature use the terms "factors" or "characteristics". Nevertheless, this study uses the term "criteria", which reflects the same meaning as "factors" or "characteristics".

Table 4. National CSIRTs Our Interviewees Worked for

National CSIRTs	Website
Uganda CERT	https://www.cert.ug/
Albania CERT	https://cesk.gov.al/
CERT BUND (Germany)	https://www.bsi.bund.de/
NCSC Switzerland	https://www.ncsc.admin.ch/
CERT-MZ (Mozambique)	https://www.cert.mz/
ID-SIRTII/CC (Indonesia)	https://idsirtii.or.id/
NCSC-FI (Finland)	https://www.kyberturvallisuuskeskus.fi/en/our-activities/cert
JpCERT/CC (Japan)	https://www.jpcert.org/
INCIBE-CERT (Spain)	https://www.incibe-cert.es/

3.2 Data Collection

Semi-structured interviews [45] were used in the study to draw insights from interviewees through interactive discussions on how they perceive the candidate criteria (RQs 1 and 2). The study received a favourable opinion from the University of Kent's Central Research Ethics Advisory Group (CREAG) under the reference number (CREAG071-05-22) on 14 June 2022.

All interviewees willingly gave consent to participate in the study and to have their direct quotes included in research publications resulting from this study, with their personal information anonymised. The Consent Form and the Participant Information Sheet (PIS) used for the semi-structured interviews are available at https://cyber.kent.ac.uk/research/CSIRTs/Validation-Criteria/Consent-Form.pdf and https://cyber.kent.ac.uk/research/CSIRTs/Validation-Criteria/PIS.pdf, respectively.

The study required specific knowledge, understanding and experiences of national CSIRTs' real-world operations to supplement the necessary information to answer RQs 1 and 2. Therefore, the selection of interviewees for the study was "purposive" instead of "random" and intentionally selected staff members of national CSIRTs to participate in the study [11]. This was essential to gain accurate, meaningful, and rich insights to answer the research question [11]. Feedback from staff members of CSIRTs is vital to gain insights into the operations of CSIRTs as they have significant experience in incident responses, much needed when intending to improve CSIRT practices [1].

To recruit interviewees, six staff members from six national CSIRTs were invited to participate in the study during the 34th Annual FIRST Conference[1] and the co-located NatCSIRT 2022 Conference[2] in Dublin, Ireland, in July 2022. Three more interviewees were invited through contacts at the CERT Division of the Software Engineering Institute (SEI) of Carnegie Mellon University in the USA[3]. Formal emails were later sent to potential interviewees to formally recruit them into this study. A sample recruitment email is available at https://cyber.kent.ac.uk/research/CSIRTs/Validation-Criteria/Recruitment-Email.pdf. Finally, nine staff members from nine national CSIRTs willingly consented to participate in the study. The national CSIRTs our interviewees worked for are shown in Table 4.

The instrument used for this study was an "Interview Schedule", which guided the interviewer during the interviews. The interview schedule contains brief information about the interview process, time allocation, and interview questions. It consists of eight open-ended and semi-structured questions, arranged into four sections in sequence: 1) basic information about interviewees, 2) how staff evaluate tools and data in national CSIRTs,

[1]https://www.first.org/conference/2022/
[2]https://www.basecybersecurity.com/cyber-security-events-infosec-conferences-it-security-trainings-europe-calendar/natcsirt-2021-2/
[3]https://www.sei.cmu.edu/about/divisions/cert/

3) how staff perceive the usefulness and deployment of the candidate criteria for evaluating tools and data in national CSIRTs, and 4) any other comments about the candidate criteria.

The order of the interview questions was flexible, allowing interviewees to highlight or introduce any other points relevant to the questions. The interview schedule used for the study is available at https://cyber.kent.ac.uk/research/CSIRTs/Validation-Criteria/Interview-Schedule.pdf.

The interviews were conducted virtually via Microsoft Teams (https://teams.microsoft.com/) between 2 August and 6 October 2022. On average, each interview took approximately 30 minutes to complete. The interviews were audio recorded, as consented to by interviewees and transcribed. For some interviewees, some follow-up email exchanges took place to clarify their opinions on some criteria. In addition to audio recording the interviews, the interviewer (the first author of the paper) also took notes of important points during the interviews.

To ensure the credibility of data collection, the semi-structured interview questions were reviewed and verified by a domain expert from CyberSecurity Malaysia[4], the national cyber security agency of the Malaysian government. A pilot semi-structured interview was held with a senior staff member of the national CSIRT of New Zealand[5] on 2 June 2022, following the exact setup of the actual nine interviews to ensure the feasibility and appropriateness of the interview questions. The pilot interview also helped improve and refine the interview questions before the actual interviews. It should be noted that data from the pilot interviews were not used in the data analysis.

3.3 Data Analysis Method – Content Analysis

We found *Content Analysis* [63] to be the best method to analyse the semi-structured interview data to gain insights and put them into the context of our research questions. Content analysis also allows for categorising, quantifying and describing the data objectively. Thematic analysis was ruled out as there was no intention to explore and identify new themes or patterns across the data and interpret its underlying meaning.

Content analysis is a qualitative data analysis method that is flexible [8, 57], yet systematic and rigorous [63]. This method is suitable when an existing theory or the research literature on a particular topic of study is limited [21] – as was the case of our study. Furthermore, content analysis is the best fit for exploratory research to gain new insights, opinions and views that could answer research questions and achieve the aim of a study [14].

Our study used "codes and coding" to capture the emerging concepts, ideas and categories underpinning the interview data and helping to organise the data [17, 47]. Codes are "tags" or "labels" assigned to raw data collected in a study, e.g., from interviews and focus groups, for analysis purposes [38].

This study adopted in-vivo coding to code the semi-structured interview data [3]. It was used to capture exactly what the interviewees had said. Hence, the codes derived from in-vivo coding are concrete and specific. Erlingsson's coding model [14] was used, which we found easier to follow, to guide the coding process. The codes were developed using a data-driven approach (from the raw interview data) instead of theory-driven [38] since the study was not based on any existing theory. This required us to re-examine the raw data repeatedly to gain clearer insights about the interview data, making the study's code development an iterative process [10]. It should be noted that only one researcher (the first author) coded the whole interview data. As such, there were no issues in establishing consistency or reliability of the coding process (in comparison to a situation in which several coders were involved, which would require further checks to ensure consistency and reliability). We used one coder because the first author is an experienced staff member of the Malaysia Computer Emergency Response Team (MyCERT) with over 20 years of work experience and substantial knowledge to do the task.

In the coding process, we extracted words, phrases and sentences from the interview data as meaningful codes while considering the research questions [38]. During coding, we focused on extracting "manifest meaning"

[4]https://www.cybersecurity.my/
[5]https://www.cert.govt.nz/

(what has been said) or surface meaning of the data instead of "latent meaning" (what is intended to be said) or deeper meaning [3]. Our study needs to capture only what the interviewees said during the interview. We used the words and phrases in the text rather than interpreting the underlying meaning of the words and text [3].

3.4 Applying the Candidate Criteria

After the candidate criteria were empirically validated using semi-structured interviews, they were validated more objectively using several metrics. This was performed by applying the candidate criteria to evaluate two sample tools and one data source to derive several concrete metrics and values. Doing so gives further evidence of the practicality of the criteria in practice. This supplements the opinions from the semi-structured interviews and makes the study's findings more credible and reliable.

For each tool and data source, all candidate criteria were checked individually. First, we tried to determine if a criterion is relevant to the evaluated tool. If NO – not relevant, move on to the next criterion. If YES – relevant, identify one or more suitable metrics for the criterion and determine the value for each metric. Value can be derived from 1) factual information about the tool's features from its documentation, and 2) output and results obtained after inputting an artefact to the tools – a PDF file.

Candidate tools and data source. Two sample tools were evaluated to demonstrate the discriminatory power of the candidate criteria developed from this study. This was performed by applying the candidate criteria to evaluate two sample tools – VirusTotal[6] and Hybrid Analysis[7], using several metrics. Evaluating two sample tools allows for comparing the tools and highlighting the utility of one tool over the other. Doing so shows that the criteria can be used to compare different tools and guide national CSIRTs in selecting more appropriate tools.

VirusTotal and Hybrid Analysis were selected for the evaluation exercise due to their significance in supporting national CSIRTs' operations [60]. This is consistent with an empirical study conducted in [39], which found that VirusTotal and Hybrid Analysis tools are used in the surveyed national CSIRTs to support incident response. Additionally, in the researcher's informal conversations with several national CSIRT staff members, there is indecision in selecting between VirusTotal and Hybrid Analysis tools to best support incident response. Hence, considering the above points, the study decided to evaluate VirusTotal and Hybrid Analysis so the results from the two candidate tools could be discerned.

In contrast, only one candidate data source is evaluated due to time constraints. The candidate data source is Shadowserver, which was selected due to its significance to national CSIRTs' operations [36, 43, 61]. This is evidenced by its utilisation by some national CSIRTs [28, 30] and also reported in [39]. The first author accessed the candidate data source through contacts with MyCERT, facilitating the evaluation exercise.

Notably, the scope of the evaluation exercise for the tools was to submit a sample file to the tools online and observe the outputs. The file can be submitted by clicking the "Choose file" in VirusTotal and "Drag & Drop for Instant Analysis" in Hybrid Analysis and observing the outputs. The Shadowserver data source was evaluated by reviewing the sample data in its original CSV file format obtained from a national CSIRT and observing if it fulfilled the candidate criteria requirement.

The candidate tools and a data source are described briefly below:

(1) VirusTotal is a free SAAS (software-as-a-service) tool owned by Chronicle (https://chronicle.security/), a subsidiary of Google. VirusTotal can be accessed online at https://www.virustotal.com/ to analyse suspicious files, hashes or URLs, which uses several back-end Antivirus engines to facilitate the detection of malware [34]. It is the largest online anti-malware scanning service, and security researchers widely use it for malware analysis.

[6]https://www.virustotal.com/
[7]https://https://www.hybrid-analysis.com/

Table 5. Opinions about Usefulness of Mohd Kassim et al.'s Candidate Criteria for National CSIRTs

How interviewees perceived usefulness of criteria (inductive codes)	Number of interviewees
The criteria provided are good, nice, great	8
Can help national CSIRTs to select tools and data	7
Useful for operations	4
Comprehensive and complete criteria	3
Approach of the criteria is good, helpful and interesting	3
The criteria are important	2
Criteria are valuable	2
A valid research area	1
Criteria have valid points	1
The basic idea around the criteria is interesting	1
The research tackled both sides, tools and data	1
Methodology used and the evaluation is nice	1
Needed by the National CSIRTs	1
The criteria fit	1
Increase quality of incident response reports	1
Positive with the criteria	1
There is no problem with the criteria	1
Would not take out any points from the criteria	1
Easy-to-understand criteria	1
Good point	1
Big help	1

VirusTotal inspects files or URLs submitted by users using more than 70 state-of-the-art anti-malware engines and returns engines' detection results if the file or URL is malicious or not [66]. VirusTotal only provides detection results from all the detections employed by back-end engines, whether the file or URL is malicious or benign.

(2) Hybrid Analysis is a free SAAS tool owned by CrowdStrike (https://www.crowdstrike.com/). It is a web-based service to detect and analyse malware using a unique Hybrid Analysis technology [51]. The tool can be accessed at https://www.hybrid-analysis.com. Hybrid Analysis is an open-source malware analysis platform that can sandbox malicious software and executables. It provides file/URL sandboxing, file collections, reports search, and sandbox results with IOCs and screenshots.

(3) Shadowserver data source [53] contains data about malicious Internet activities worldwide (e.g., malware, botnets, spam and computer fraud). The data is structured to include the following fields: date, timestamp, IP address, hostname, geolocation, URL, ASN, port numbers, protocol, name of malware, and file hash. The data is essential for national CSIRTs to notify respective service providers concerning malicious activities originating from their IP addresses or domains. Doing so helps to address emerging threats worldwide and for cyber crime investigations [28].

4 RESULTS

4.1 RQ1 Results: How interviewees perceived the practical usefulness of the candidate criteria

All nine interviewees generally perceived the candidate criteria as useful for evaluating tools and data in national CSIRTs. The majority (8) of interviewees perceived the candidate criteria very positively ("good", "nice", and "great"). Seven interviewees expressed that the candidate criteria could help national CSIRTs select the right

tools and data sources, and three commented that the candidate criteria were comprehensive and complete. A complete list of opinions captured from interviewees concerning how they perceived the usefulness of the candidate criteria is shown in Table 5.

Besides evaluating tools and data, the candidate criteria were also considered useful for software tool development in national CSIRTs. One participant said that national CSIRTs could refer to the criteria when planning to develop software, as commented below:

> "*This was also a very important for us when we decided to develop a new tool.*" (NCSC-Switzerland)

One participant mentioned that their national CSIRT had technical criteria to evaluate tools and data sources. After looking at the candidate criteria from our study, the participant realised that their criteria were incomplete. This is because user perspective criteria are essential but missing in their own criteria. Therefore, this participant perceived comprehensive criteria like ours as good and useful for national CSIRTs. This participant gave the following statement:

> "*We have lots of technical criteria, but not in the end user perspective. So I think this is a good, good starting point.*" (NCSC-FI)

The same participant pointed out that, in general, the criterion "Usability" is often missing when assessing open-source tools. Hence, the candidate criteria are perceived better as it includes "Usability", as mentioned below:

> "*I think this is better in the sense that you have thought about the usability. Also I think that's one criteria that is never, not never but it's fairly, not at least in open source, it's kind of missing.*" (NCSC-FI)

Besides being useful, three interviewees perceived the approach and method used by the study to establish the candidate criteria as interesting and useful. This is reflected in the below comment:

> "*I guess the general approach is like interesting and I think that that's quite useful.*" (JpCERT/CC)

4.2 RQ2 Results: How interviewees perceived deployment readiness of the candidate criteria

All (9) interviewees perceived that the candidate criteria could be deployed in national CSIRTs to evaluate tools and data, and two interviewees perceived that they could be deployed in other types of CSIRTs.

Furthermore, all interviewees also expressed willingness to adopt the criteria in their operation once they are available as a public resource, while two of them considered the criteria as a good option for deployment. One participant said they would focus on deploying the criterion "Usability" since it is largely missing from their current criteria. Three interviewees expressed their willingness to deploy the criteria from this study at any time in their national CSIRTs.

Notably, one participant suggested that the criteria should be translated into guidelines and best practices for deployment in national CSIRTs for evaluating tools and data. The guidelines and best practices are not necessarily mandatory but recommended for national CSIRTs. This was commented as below:

> "*Saying that the criteria can be a guideline or best practice.*" (CERT-BUND)

One participant expressed their positivity about deploying the candidate criteria in their national CSIRT operation, knowing that it leverages ISO/IEC standards:

> "*Especially for the criteria, which is already referred to the applicable international standard reports.*" (IDSIRTII)

Three interviewees perceived that the candidate criteria would greatly help, especially the new national CSIRTs when deployed, to select appropriate tools and data. This was commented on below by an interviewee:

Table 6. Opinions about Deployment Readiness of Mohd Kassim et al.'s Criteria in National CSIRTs

How interviewees perceived deployment readiness of criteria (inductive codes)	Number of interviewees
Can be used in national CSIRTs	7
Don't have criteria like this	2
Can be used in CSIRTs	1
Can be used in all CSIRTs	1
Worth a try	1
Can be a best practice	1
Good if implemented in our organisation	1
Can be a guideline	1
To evaluate new tools, of course	1
They will help us – national CSIRT	1
We want to borrow the criteria	1
It is good for new CSIRTs	1
Beneficial for us	1
Not a must-have	1
Certainly	1

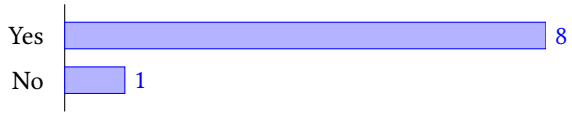

Fig. 1. Recommending Mohd Kassim et al.'s Criteria to Other National CSIRTs

> "*Yes, I think I'm sure the tool [the criteria] is very important, especially because we are in our early stages, and that can help us use the criteria to actually select the tools that we're going to use.*" (CERT-Mozambique)

This is further supported by another participant who perceived that the candidate criteria would greatly help new national CSIRTs who may not have knowledge of tools and data evaluation practices. This is mentioned below:

> "*Yes, I think this is a very good help, probably for new CERTs who do not know how to evaluate products.*" (NCSC-Switzerland)

A complete list of opinions from interviewees concerning how they perceived the deployment of the candidate criteria in national CSIRTs is shown in Table 6.

On recommending the candidate criteria to other national CSIRTs, the majority (8) of interviewees said they would recommend the candidate criteria to other national CSIRTs for deployment, as shown in Figure 1. Only one participant mentioned needing to deploy the candidate criteria in their national CSIRT before recommending it to others. Nevertheless, the participant perceived that the candidate criteria should work fine when deployed in their national CSIRT environment.

4.3 RQ3: Evaluation Results of Two Candidate Tools and a Data Source

4.3.1 Results of Evaluating Two Software Tools. The software tools evaluated in our study are *VirusTotal* and *Hybrid Analysis*. The first author of this paper manually evaluated the tools as the sole tester between 30 and 31 October 2023. The first author was considered a representative tester because of her rich experience working at a

national CSIRT. All results and observations were recorded in a table format using a Word document on the same day for subsequent analysis. The results were discussed with other co-authors to get their feedback and to reach a consensus among all co-authors.

The candidate criteria for evaluating tools (shown in Tables 1 and 2) were first reviewed to understand each criterion's requirement(s). Then, the online documentation of each tool was read to understand its functionalities and features. Each criterion was examined to determine if it was relevant to each tool. When a criterion was considered relevant, it was translated into one or more concrete metrics that could cover the requirement(s) of the criterion. The values of the metrics could be 1) binary (YES/NO), 2) categorical, 3) numeric (e.g., the time taken to complete a task), or 4) descriptive (e.g., the tool supports a graphical user interface (GUI)).

For quantitative metrics, a sample artefact was fed into each tool to estimate each metric's value. For qualitative metrics, the value was derived from the first author's personal judgment based on her work experience as a staff member of a national CSIRT. Then, based on the value(s) of the corresponding metric (s), it was determined if the tools fulfilled each criterion's requirement (s).

For example, let us use the criterion "Supportability" to illustrate how the above evaluation process worked. This criterion is relevant for the tools since staff members of national CSIRTs are concerned that they can get support and help if any issues arise. One metric defined for this criterion is categorical: "what type of support is provided by the tool", with three possible values – "24x7 Live Support", "Online Support" and "Chat Bot". After evaluating the two sample tools, the following values were determined: "Chat Bot" for VirusTotal and "Online Support" for Hybrid Analysis.

A second example is the criterion "Time behaviour". This criterion is relevant because users are concerned about how fast they can get the results. One metric is defined for this criterion: "The amount of time taken to obtain the results when uploading a file or inputting a URL (from the start to the return of the results)" and the value is numeric (in seconds). After evaluating the sample tools, the following numeric values were determined: 2 for VirusTotal and 9 for Hybrid Analysis.

A third example is the criterion "Interoperability". This criterion is relevant because users are concerned about integrating the tools with other third-party applications. One categorical metric is defined for this criterion:"How the tool supports integration and information exchange with third-party applications", with the following four identified values – "API", "an export feature for data exchange", "an import feature for data exchange", and "No support". After evaluating the two sample tools, the following values were determined: "API" for both VirusTotal and for Hybrid Analysis.

A fourth example is the criterion "Functionality". This criterion is considered relevant as users are concerned that a tool's functionalities cover all the specified tasks and if they meet the user's objectives. For the two sample tools, one Boolean metric identified is "if the tool has a file scanning feature and performs the specified functionality (feature) accordingly". After evaluating the two sample tools, the value was determined to be "Yes" for both VirusTotal and Hybrid Analysis since both tools have such a file scanning feature.

The detailed evaluation results of VirusTotal and Hybrid Analysis are available at https://anonymous.4open.science/r/Evaluation-Result-of-Tools-8898/Evaluation-Sample-Tools.pdf. The metrics used for each criterion and the values generated are appended in the results.

4.3.2 Results of Evaluating a Data Source. The data source evaluated in our study is *Shadowserver*. The data source was manually evaluated by the first author on 12 February 2023, and the results and observations were recorded in the same way as the evaluation exercise of the two sample tools.

The candidate criteria for evaluating data sources (as shown in Table 3) were first reviewed to understand each criterion's requirement(s). Then, the data source's online documentation was read to understand its features better. Each criterion was examined to determine if it was relevant to the data source, and one or more metrics

were identified for each relevant criterion. If the data source fulfils each criterion's requirement(s), it is judged based on the value(s) of the metric(s) identified for the criterion.

To illustrate how the evaluation was done, take the criterion "Efficiency" as an example. This criterion is relevant to the data source as staff members of national CSIRTs are concerned with the time spent identifying an indicator of compromise (IOC) in the data. One metric specified for the criterion is numeric: "the average time taken to analyse and identify the IOC" (in seconds). The value identified after evaluating the data source is 5.

A second example is the criterion "Understandability". This criterion was considered relevant for the data source as users are concerned that the data is understandable by staff members of CSIRTs, hence it is presented in a way that can be easily understood. One metric identified is multi-valued and categorical: "the format of human-understandable output", with two identified values – "as a CSV file" and "displayed visually in a table". After the evaluation, it was confirmed that the data source has both values.

A third example is the criterion "Precision". This criterion is relevant as users care about whether the data is precise enough to take further action, such as takedowns of phishing or malware-hosting websites. Less precise data would deter further analysis and action from a CSIRT end. One metric identified is multi-valued and categorical: "what specific details the data has about an incident". The values identified include: 'complete URL', 'source IP address', 'destination IP address', 'timestamp', 'hash value(s) of related software/malware', 'network protocol type', 'port number(s)', 'geo-location(s) obtained from IP addresses'. After evaluating the sample data source, the following values were determined: 'complete URL', 'source IP address', 'destination IP address', 'timestamp', 'hash value(s) of related software/malware', 'network protocol type', 'port number(s)', and 'geo-location(s) obtained from IP addresses'.

The Shadowserver data source evaluation results are available at https://anonymous.4open.science/r/Evaluation-Result-of-Tools-8898/Evaluation-Data-Shadowserver.pdf. The metrics used for each criterion and the values generated are appended in the results.

4.3.3 Implications of the Evaluation Results.

Operationalising the Criteria for Evaluating Tools and Data. The evaluation results show that the criteria could be potentially operationalised for evaluating tools and data in national CSIRTs. The evaluation results show how the criteria can be contextualised and translated into concrete metrics when applied in real-world operations to evaluate tools and data.

Demonstrating the Criteria's Usefulness. The evaluation of two different tools, VirusTotal and Hybrid Analysis, presented the differences between the two tools, though some similarities were also present. During the evaluation exercise, the differences between VirusTotal and Hybrid Analysis were observed in terms of "User interface aesthetics", "Accessibility", "Performance efficiency", "Interoperability", "Learnability" and "Supportability". Such differences show the potential ability of the criteria to distinguish different tools from each other. This could highlight the utility of one tool over the other and potentially help in decision-making to identify suitable tools to support incident responses. Moreover, it helps to provide a more systematic way of identifying suitable tools to support incident response by evaluating them with a set of criteria, as opposed to current practices reported in [39, 40] – a key gap identified and aimed to address in this research. This implies that the criteria validated in this research could help national CSIRTs select suitable tools systematically for their incident response operational needs.

5 FURTHER DISCUSSIONS

Based on the results reported in Section 4, this section discusses high-level insights into the RQs.

RQ1. As shown in the results in Section 4, enlightening feedback was received from the interviewees regarding whether the candidate criteria are useful for national CSIRTs' operations. Interviewees also confirmed the comprehensiveness of the candidate criteria, and some appreciated the research method underlying the identification and validation of the candidate criteria in the study. It is worth noting that some interviewees reported using some criteria for tool and data evaluation, primarily criteria representing "Product Quality", in their national CSIRTs, such as in NCSC-FI and CERT-INCIBE. However, these criteria are just a smaller subset of the candidate criteria we validated in our work, often with "Quality in Use" criteria missing. This shows that current approaches used to evaluate tools and data in some National CSIRTs are less systematic and could be improved by deploying the study's criteria.

The interviewees' positive feedback on our work also provided further evidence for more research on how the operations of national CSIRTs can be improved. There seems to be a general lack of systematic treatment and rigour in how decisions are currently made within national CSIRTs. For instance, this is reflected in one participant's comment below:

> "*Research should probably be proactive about this kind of issues and ... I think that your approach will be helpful.*" (JpCERT/CC)

RQ2. The positive feedback from interviewees made the impression that the candidate criteria are ready to be deployed in operational practices at their national CSIRTs. The positivity could be better explained by the lack of systematic procedures and comprehensive criteria for evaluating tools and data in the current operational practices in national CSIRTs. This is consistent with previous studies, which found a lack of systematic procedures for evaluating tools and data in national CSIRTs [39, 40]. One representative comment from a participant is given below:

> "*There has to be a formal way of doing something. It should, even if it is an SOC, even if it is a CERT that it has a lot of it is deals with a lot of many constituencies, many stakeholders. There has to be a baseline at least.*" (Uganda CERT)

It is particularly encouraging to see that most interviewees were willing to recommend the candidate criteria to other national CSIRTs. This implies that interviewees could see potential benefits that all national CSIRTs will gain from such candidate criteria and a more systematic tool and data evaluation approach. Another aspect commented on by the interviewees is that the candidate criteria can be a valuable reference for national CSIRTs to consider how to tailor them to meet their local needs. One comment on this aspect is given below:

> "*... if there's any, like any sort of like criteria or recommendations for national CSIRTs, I can definitely show them, you know, hey, this is something that they're using for the national CSIRT community. So I mean, I don't really ask them to comply completely, but they can refer to it.*" (JpCERT/CC)

Overall, findings from the interview results validated the candidate criteria for usefulness and deployment-readiness in national CSIRTs. The fact that no interviewees pointed out any changes or refinements to the candidate criteria indicates that they are at least good enough as a first set of empirically validated criteria for evaluating tools and data. Hence, we consider RQ2 answered positively.

RQ3. Although interviewees gave positive feedback on RQs 1 and 2, applying the candidate criteria to concrete tools and data sources to gain more direct evidence on the practical usefulness and deployment readiness is essential. Our exercises of applying the candidate criteria to two sample tools and one sample data source indicate that special care is needed when considering the relevancy of each criterion and what metric(s) could be defined to capture more concrete requirement(s) of each criterion. This process is not trivial and requires some guidelines and good case studies as examples to inform staff of national CSIRTs on how the general criteria can be contextualised for different tools and data sources.

Our evaluation exercises showed that it is easier to translate "Product Quality" criteria into concrete metrics since they are primarily about static and factual properties of the tool or data source evaluated. It is generally more complicated to consider how to handle "Quality in Use" criteria since they are often about end users' opinions and other more subjective judgments on the tool and data source evaluated. In our evaluation exercises, the first author played the role of a staff member of a national CSIRT, and the values of identified metrics were determined according to her personal opinions.

Overall, the results of both empirical studies confirmed the completeness, comprehensiveness, practical usefulness and deployment readiness of the candidate criteria. The criteria and the sample evaluation results will be released as public resources to help national CSIRTs and other types of CSIRTs consider adopting them in their operational practices.

6 LIMITATIONS AND FUTURE WORK

6.1 Limitations

One limitation is that the small sample size (9) used for the semi-structured interviews might make the findings less generalisable [22]. Although we attempted to recruit more interviewees, it proved challenging. This was not unexpected given that the target pool of interviewees is very niche and that staff within national CSIRTs are generally very busy with their work. Similar difficulties have also been reported in [7, 39, 46, 48].

Another limitation is that the candidate criteria are largely very high-level, so translating them into more concrete metrics and values is not trivial, which was one of the key observations from the first author's evaluation exercises reported in Section 4.3.

Potential limitations might exist for ensuring the smooth application of the criteria in national CSIRTs. Nevertheless, this could be mitigated by allocating ample time to study and consider the relevancy of each criterion and what metric(s) could be defined to capture more concrete requirements of each criterion. It is also essential that, for "Product Quality" criteria, more objective and consistent metrics can be defined, e.g., time taken to complete or learn a task.

Despite these limitations, we consider the main findings of our study still valid and reliable, especially when we take into account the following facts:

- The nine interviewees' opinions are highly consistent, and there is a consensus on both RQs 1 and 2, so the saturation effect [16] is already observed even with such a small sample. This indicates that further data collection is likely unnecessary [52]. This is further exemplified by Hennink et al. [20] in their study, which found code saturation (91%) was reached at nine interviews, with concrete codes and a stabilised codebook to capture the themes.
- The main findings are coherent with the results presented in [40] regarding the candidate criteria for evaluating and selecting tools and data sources for CSIRT operations.
- The main findings also match the first author's experience as an employee of a national CSIRT for over 20 years.

Although we are very confident about our work's main findings, it would still be helpful if other independent researchers and stakeholders did some re-validation work.

6.2 Future Work

As mentioned above, future work has been suggested on the criterion-to-metric process and creating more detailed guidelines and case studies. Such future work would be helpful as the candidate criteria are largely high-level; translating them into more concrete metrics and values can be challenging. Nonetheless, these guidelines and case studies would be helpful to various CSIRTs when evaluating tools and data in their operations.

Another future research direction is to expand the criteria and metrics from this research to construct an even more comprehensive taxonomy or ontology that will connect the criteria, different types of tools and data sources used by (national and non-national) CSIRTs. Doing so will inform the development of more practical operational guidelines and potentially enable partial automation of the tool and data evaluation procedure. In addition, more metrics and scoring systems can be identified and tested for even more quantitative and reproducible evaluation exercises. If the taxonomy/ontology can be made machine-readable, we can also construct an online system that can help automatically evaluate and recommend tools and data sources to the staff of national CSIRTs and other end users.

We also suggest conducting future work on more in-depth studies of the criteria not currently available in the ISO/IEC 25000 SQuaRE Model, e.g., Compliance, Popularity, and Certification. We treated these as separate criteria in our study, but they may be merged into other criteria as sub-criteria. They could also be relevant for evaluating data sources, although we did not examine this possibility in this research. If it is confirmed that these criteria should remain separate without merging, it will be helpful to work with the international standardisation community to add them to future editions of relevant international standards, especially ISO/IEC 25010 and ISO/IEC 25012.

7 CONCLUSION

This work presents the results of our work on validating a set of candidate criteria proposed by Mohd Kassim et al. [40] for evaluating tools and data sources in the context of cyber incident response operations of national CSIRTs. The validation was done via nine semi-structured interviews to understand national CSIRT staff's perception of the candidate criteria and by applying the candidate criteria to evaluate two sample tools and one sample data source widely used by national CSIRTs. The results of both empirical studies confirmed the completeness, comprehensiveness, practical usefulness and deployment readiness of the candidate criteria. Our evaluation results will be released as public resources to help national CSIRTs and other CSIRTs consider adopting the study's criteria in their incident response operational practices.

REFERENCES

[1] Atif Ahmad, Sean B. Maynard, and Graeme Shanks. 2015. A case analysis of information systems and security incident responses. *International Journal of Information Management* 35, 6 (2015), 717–723. https://doi.org/10.1016/j.ijinfomgt.2015.08.001

[2] Jose Antonio Mulet Alberola and Irene Fassi. 2022. Towards the assessment of performance-based interactions in collaborative CPPS. *Procedia Computer Science* 200 (2022), 1636–1645.

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

[4] Nigel Bevan. 2001. International standards for HCI and usability. *International Journal of Human-Computer Studies* 55, 4 (2001), 533–552. https://doi.org/10.1006/ijhc.2001.0483

[5] Tracy Bills, Brittany Manley, and James Lord. 2022. *Enabling the Sustainability and Success of a National Computer Security Incident Response Team.* Handbook. Carnegie Mellon University. https://resources.sei.cmu.edu/asset_files/Handbook/2022_002_001_885865.pdf

[6] Barry W. Boehm, John R. Brown, and Mlity Lipow. 1976. Quantitative Evaluation of Software Quality. In *Proceedings of the 2nd International Conference on Software Engineering*. ACM, 592–605. https://doi.org/10.5555/800253.807736

[7] David Botta, Rodrigo Werlinger, André Gagné, Konstantin Beznosov, Lee Iverson, Sidney Fels, and Brian Fisher. 2007. Towards Understanding IT Security Professionals and Their Tools. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*. ACM, 100–111. https://doi.org/10.1145/1280680.1280693

[8] Stephen Cavanagh. 1997. Content analysis: concepts, methods and applications. *Nurse Researcher* 4, 3 (1997), 5–16. https://doi.org/10.7748/nr.4.3.5.s2

[9] Bill Curtis, Robert A. Martin, and Philippe-Emmanuel Douziech. 2022. Measuring the Structural Quality of Software Systems. *Computer* 55, 3 (2022), 87–90. https://doi.org/10.1109/MC.2022.3145265

[10] Jessica T. DeCuir-Gunby, Patricia L. Marshall, and Allison W. McCulloch. 2011. Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project. *Field Methods* 23, 2 (2011), 136–155. https://doi.org/10.1177/1525822X103884

[11] Tom Deliens, Peter Clarys, Ilse De Bourdeaudhuij, and Benedicte Deforche. 2014. Determinants of eating behaviour in university students: a qualitative study using focus group discussions. *BMC Public Health* 14, 1, Article 53 (2014), 12 pages. https://doi.org/10.1186/1471-2458-14-53

[12] R. Geoff Dromey. 1995. A Model for Software Product Quality. *IEEE Transactions on Software Engineering* 21, 2 (1995), 146–162. https://doi.org/10.1109/32.345830

[13] Elisabeth Dubois and Unal Tatar. 2022. Mitigating Global Cyber Risk Through Bridging the National Incident Response Capacity Gap. In *Proceedings of the 17th International Conference on Information Warfare and Security*, Vol. 17. Academic Conferences International Limited, 527–531. https://doi.org/10.34190/iccws.17.1.66

[14] Christen Erlingsson and Petra Brysiewicz. 2017. A hands-on guide to doing content analysis. *African Journal of Emergency Medicine* 7, 3 (2017), 93–99. https://doi.org/10.1016/j.afjem.2017.08.001

[15] Steven Furnell, Pete Fischer, and Amanda Finch. 2017. Can't get the staff? The growing need for cyber-security skills. *Computer Fraud & Security* 2017, 2 (2017), 5–10. https://doi.org/10.1016/S1361-3723(17)30013-1

[16] Patricia I. Fusch and Lawrence R. Ness. 2015. Are We There Yet? Data Saturation in Qualitative Research. *The Qualitative Report* 20, 9 (2015), 1408–1416. https://doi.org/10.46743/2160-3715/2015.2281

[17] Lisa M. Given. 2008. *The SAGE Encyclopedia of Qualitative Research Methods*. SAGE. https://uk.sagepub.com/en-gb/eur/the-sage-encyclopedia-of-qualitative-research-methods/book229805

[18] Robert B. Grady. 1992. *Practical Software Metrics for Project Management and Process Improvement*. Prentice-Hall, Inc.

[19] Glenn Gumba, Deborah G. Brosas, and Jessie R. Paragas. 2021. Assessment of SIAS Application Using Software Quality Model. In *Proceedings of the 2021 3rd International Conference on Research and Academic Community Services*. IEEE, 197–202. https://doi.org/10.1109/ICRACOS53680.2021.9701982

[20] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. 2017. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative Health Research* 27, 4 (2017), 591–608. https://doi.org/10.1177/1049732316665344

[21] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15, 9 (2005), 1277–1288. https://doi.org/10.1177/1049732305276687

[22] George Iakovakis, Constantinos Giovanni Xarhoulacos, Konstantinos Giovas, and Dimitris Gritzalis. 2021. Analysis and Classification of Mitigation Tools against Cyberattacks in COVID-19 Era. *Security and Communication Networks* 2021, 1, Article 3187205 (2021), 21 pages. https://doi.org/10.1155/2021/3187205

[23] International Organization for Standardization (ISO). 2008. Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model. web page. https://www.iso.org/standard/35736.html

[24] International Organization for Standardization (ISO). 2011. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models. web page. https://www.iso.org/standard/35733.html

[25] International Organization for Standardization (ISO). 2021. Information technology — Software measurement — Software quality measurement — Automated source code quality measures. web page. https://www.iso.org/en/contents/data/standard/08/06/80623.html

[26] International Telecommunication Unit. 2023. Global Cybersecurity Index 2020. web page. https://www.itu.int/epublications/publication/D-STR-GCI.01-2021-HTM-E

[27] Internet Governance Forum (IGF). 2014. Internet Governance Forum (IGF) 2014: Best Practice Forum on Establishing and Supporting Computer Security Incident Response Teams (CSIRT) for Internet Security. Online document. https://www.intgovforum.org/cms/documents/best-practice-forums/establishing-and-supporting-computer-emergency-response-teams-certs-for-internet-security/409-bpf-2014-outcome-document-computer-security-incident-response-teams/file

[28] Kamol Kaemarungsi, Nawattapon Yoskamtorn, Kitisak Jirawannakool, Nuttapong Sanglerdsinlapachai, and Chanin Luangingkasut. 2009. Botnet Statistical Analysis Tool for Limited Resource Computer Emergency Response Team. In *Proceedings of the 2009 Fifth International Conference on IT Security Incident Management and IT Forensics*. IEEE, 27–40. https://doi.org/10.1109/IMF.2009.13

[29] Sharifah Roziah Binti Mohd Kassim, Solahuddin Bin Shamsuddin, Shujun Li, and Budi Arief. 2022. How National CSIRTs Operate: Personal Observations and Opinions from MyCERT. In *Proceedings of the 2022 IEEE Conference on Dependable and Secure Computing*. IEEE, 2 pages. https://doi.org/10.1109/DSC54232.2022.9888803

[30] Erka Koivunen. 2010. "Why Wasn't I Notified?": Information Security Incident Reporting Demystified. In *Information Security Technology for Applications: 15th Nordic Conference on Secure IT Systems, NordSec 2010, Espoo, Finland, October 27-29, 2010, Revised Selected Papers*. Springer, 55–70. https://doi.org/10.1007/978-3-642-27937-9_5

[31] Toshihiro Komiyama, Shin'ichi Fukuzumi, Motoei Azuma, Hironori Washizaki, and Naohiko Tsuda. 2020. Usability of Software–Intensive Systems from Developers' Point of View. In *Human-Computer Interaction. Design and User Experience: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I*. Springer, 450–463. https://doi.org/10.1007/978-3-030-49059-1_33

[32] Marko Krstic, Milan Cabarkapa, and Aleksandar Jevremovic. 2019. Machine Learning Applications in Computer Emergency Response Team Operations. In *Proceedings of the 27th Telecommunications Forum*. IEEE, 4 pages. https://doi.org/10.1109/TELFOR48224.2019.8971040

[33] Marc Kührer, Christian Rossow, and Thorsten Holz. 2014. Paint It Black: Evaluating the Effectiveness of Malware Blacklists. In *Research in Attacks, Intrusions and Defenses: 17th International Symposium, RAID 2014, Gothenburg, Sweden, September 17-19, 2014. Proceedings*. Springer, 1–21. https://doi.org/10.1007/978-3-319-11379-1_1

[34] Rima Masri and Monther Aldwairi. 2017. Automated Malicious Advertisement Detection Using VirusTotal, URLVoid, and TrendMicro. In *Proceedings of the 2017 8th International Conference on Information and Communication Systems*. 336–341. https://doi.org/10.1109/IACS.2017.7921994

[35] Jim A. McCall, Paul K. Richards, and Gene F. Walters. 1977. *Factors in Software Quality. Volume I. Concepts and Definitions of Software Quality*. Technical Report ADA049014. General Electric Company. https://apps.dtic.mil/sti/citations/ADA049014

[36] Clinton J. Mielke and Hsinchun Chen. 2008. Botnets, and the Cybercriminal Underground. In *Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 206–211. https://doi.org/10.1109/ISI.2008.4565058

[37] José P Miguel, David Mauricio, and Glen Rodríguez. 2014. A Review of Software Quality Models for the Evaluation of Software Products. *International Journal of Software Engineering & Applications* 5, 6 (2014), 31–54. https://doi.org/10.5121/ijsea.2014.5603

[38] Matthew B. Miles and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. SAGE.

[39] Sharifah Roziah Binti Mohd Kassim, Shujun Li, and Budi Arief. 2022. How National CSIRTs Leverage Public Data, OSINT and Free Tools in Operational Practices: An Empirical Study. *Cyber Security: A Peer-Reviewed Journal* 5, 3 (2022), 251–276. https://www.ingentaconnect.com/contentone/hsp/jcs/2022/00000005/00000003/art00007

[40] Sharifah Roziah Binti Mohd Kassim, Shujun Li, and Budi Arief. 2023. Understanding How National CSIRTs Evaluate Cyber Incident Response Tools and Data: Findings from Focus Group Discussions. *ACM Digital Threats: Research and Practice* 4, 3, Article 45 (2023), 24 pages. https://doi.org/10.1145/3609230

[41] Radka Nacheva and Anita Jansone. 2020. Evaluation of Business Process Modelling Tools through Software Quality Metrics. *Baltic Journal of Modern Computing* 8, 4 (2020), 534–542. https://doi.org/10.22364/bjmc.2020.8.4.04

[42] Hidenori Nakai, Naohiko Tsuda, Kiyoshi Honda, Hironori Washizaki, and Yoshiaki Fukazawa. 2016. A SQuaRE-based Software Quality Evaluation Framework and its Case Study. In *Proceedings of the 2016 IEEE Region 10 Conference*. IEEE, 3704–3707. https://doi.org/10.1109/TENCON.2016.7848750

[43] Marcin Nawrocki, Maynard Koch, Thomas C Schmidt, and Matthias Wählisch. 2021. Transparent Forwarders: An Unnoticed Component of the Open DNS Infrastructure. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*. ACM, 454–462. https://doi.org/10.1145/3485983.3494872

[44] Monika Nowikowska. 2022. The Main Tasks of the Network of Computer Security Incident Response Teams in the Light of the Act on the National Cybersecurity System in Poland. In *Cybersecurity in Poland*. Springer, 223–241. https://doi.org/10.1007/978-3-030-78551-2_15

[45] Briony J. Oates. 2005. *Researching Information Systems and Computing*. SAGE. https://us.sagepub.com/en-us/nam/researching-information-systems-and-computing/book226898

[46] Sean Oesch, Robert Bridges, Jared Smith, Justin Beaver, John Goodall, Kelly Huffer, Craig Miles, and Dan Scofield. 2020. An Assessment of the Usability of Machine Learning Based Tools for the Security Operations Center. In *Proceedings of the 2020 IEEE Congress on Cybermatics: 2020 International Conferences on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data and IEEE Congress on Cybermatics*. IEEE, 634–641. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00111

[47] Anthony J. Onwuegbuzie, Wendy B. Dickinson, Nancy L. Leech, and Annmarie G. Zoran. 2009. A Qualitative Framework for Collecting and Analyzing Data in Focus Group Research. *International Journal of Qualitative Methods* 8, 3 (2009), 1–21. https://doi.org/10.1177/160940690900800301

[48] Celeste Lyn Paul. 2014. Human-Centered Study of a Network Operations Center: Experience Report and Lessons Learned. In *Proceedings of the 2014 ACM Workshop on Security Information Workers*. ACM, 39–42. https://doi.org/10.1145/2663887.2663899

[49] Paweł Pawlinski and Andrew Kompanek. 2016. Evaluating Threat Intelligence Feeds. Presentation slides. https://www.first.org/resources/papers/munich2016/kompanek-pawlinski-evaluating-threat-ntelligence-feeds.pdf

[50] Zane Pokorny (Ed.). 2019. *The Threat Intelligence Handbook: Moving Toward a Security Intelligence Program* (2nd ed.). CyberEdge Group, LLC. https://go.recordedfuture.com/book-2

[51] Reischaga, Charles Lim, and Yohanes Syailendra Kotualubun. 2020. Uncovering Malware Traits Using Hybrid Analysis. In *Proceedings of the 2021 International Conference on Engineering and Information Technology for Sustainable Industry*. ACM, Article 30, 6 pages. https://doi.org/10.1145/3429789.3429867

[52] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity* 52 (2018), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8

[53] Shadowserver Foundation. [n. d.]. Shadowserver. website. https://www.shadowserver.org/

[54] Jonathan M. Spring and Phyllis Illari. 2021. Review of Human Decision-making during Computer Security Incident Analysis. *Digital Threats: Research and Practice* 2, 2, Article 11 (2021), 47 pages. https://doi.org/10.1145/3427787

[55] Julie Steinke, Balca Bolunmez, Laura Fletcher, Vicki Wang, Alan J. Tomassetti, Kristin M. Repchick, Stephen J. Zaccaro, Reeshad S. Dalal, and Lois E. Tetrick. 2015. Improving Cybersecurity Incident Response Team Effectiveness using Teams-based Research. *IEEE Security & Privacy* 13, 4 (2015), 20–29. https://doi.org/10.1109/MSP.2015.71

[56] Manoj Wadhwa Suman and MDU Rohtak. 2014. A comparative study of software quality models. *International Journal of Computer Science and Information Technologies* 5, 4 (2014), 5634–5638. https://ijcsit.com/docs/Volume%205/vol5issue04/ijcsit20140504177.pdf

[57] Renata Tesch. 2013. *Qualitative Research: Analysis Types and Software*. Routledge. https://doi.org/10.4324/9781315067339

[58] Shohei Toyama and Masayuki Hirayama. 2018. User Interface Design Method Considering UI Device in Internet of Things System. In *Proceedings of the 2018 6th International Conference on Future Internet of Things and Cloud Workshops*. IEEE, 1–6. https://doi.org/10.1109/W-FiCloud.2018.00007

[59] Naohiko Tsuda, Hironori Washizaki, Kiyoshi Honda, Hidenori Nakai, Yoshiaki Fukazawa, Motoei Azuma, Toshihiro Komiyama, Tadashi Nakano, Hirotsugu Suzuki, Sumie Morita, Katsue Kojima, and Akiyoshi Hando. 2019. WSQF: Comprehensive Software Quality Evaluation Framework and Benchmark Based on SQuaRE. In *Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice*. IEEE, 312–321. https://doi.org/10.1109/ICSE-SEIP.2019.00045

[60] Martijn van der Heide. 2017. *Establishing a CSIRT, Version 1.2*. Technical report. ThaiCERT and ETDA. https://www.first.org/resources/guides/Establishing-CSIRT-v1.2.pdf

[61] Michel van Eeten, Qasim Lone, Hadi Asghari TUD, and Hadi Asghari. 2015. WP4 Evaluating and Incentivizing Botnet Mitigation. (2015), 67 pages. https://www.acdc-project.eu/wp-content/uploads/2015/05/ACDC_D4.2_Statistical_Evaluation_final.pdf

[62] Václav Vostrovskỳ, Jan Tyrychtr, and Roman Kvasnička. 2020. Open Data Quality Management Based on ISO/IEC SQuaRE Series Standards in Intelligent Systems. In *Applied Informatics and Cybernetics in Intelligent Systems: Proceedings of the 9th Computer Science On-line Conference 2020, Volume 3*. Springer, 625–631. https://doi.org/10.1007/978-3-030-51974-2_58

[63] Marilyn Domas White and Emily E. Marsh. 2006. Content Analysis: A Flexible Methodology. *Library Trends* 55, 1 (2006), 22–45. https://doi.org/10.1353/lib.2006.0053

[64] Tangxiao Yuan, Kondo Hloindo Adjallah, Alexandre Sava, Huifen Wang, and Linyan Liu. 2021. Issues of Intelligent Data Acquisition and Quality for Manufacturing Decision-Support in an Industry 4.0 Context. In *Proceedings of the 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Vol. 2. IEEE, 1200–1205. https://doi.org/10.1109/IDAACS53288.2021.9660957

[65] Mohammad Zarour. 2020. A rigorous user needs experience evaluation method based on software quality standards. *Telkomnika Journal* 18, 5 (2020). https://doi.org/10.12928/TELKOMNIKA.v18i5.16061

[66] Shuofei Zhu, Ziyi Zhang, Limin Yang, Linhai Song, and Gang Wang. 2020. Benchmarking Label Dynamics of VirusTotal Engines. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2081–2083. https://doi.org/10.1145/3372297.3420013