RESEARCH

Improving the Exchange of Lessons Learned in Security Incident Reports: Case Studies in the Privacy of Electronic Patient Records

Ying He*, Chris Johnson and Yu Lu

*Correspondence:
yingh@dcs.gla.ac.uk
School of Computing Science, University of Glasgow, 18 Lilybank Gardens, Glasgow, UK
Full list of author information is available at the end of the article
†Equal contributor

Abstract

The increasing use of Electronic Health Records has been mirrored by a similar rise in the number of security incidents where confidential information has inadvertently been disclosed to third parties. These problems have been compounded by an apparent inability to learn from previous violations; similar security incidents have been observed across Europe, North America and Asia. This has resulted in the loss of confidence and trust of the public towards the organisations' ability to protect the patients' private information. The Generic Security Template (G.S.T.) has been proposed to communicate security lessons learned from previous security incidents. This paper conducts a series of empirical studies to evaluate the usability of the G.S.T. The first study compares the G.S.T. with the conventional text-based security incident reports. The two methods were compared in term of the users' ability to identify a number of lessons learned from investigations into previous incidents involving the disclosure of healthcare records. The study showed that the graphical approach resulted in higher accuracy in terms of number of correct answers generated by participants. However, subjective feedback raised further questions about the usability of the G.S.T. as the readers of security incident reports try to interpret the lessons that can increase the security of patient data. The second study further evaluates the usability of the G.S.T. using the Cognitive Dimensions and identifies some aspects that need to be improved.

Keywords: Lessons Learned; Security Incident; Electronic Patient Record; Generic Security Template; Empirical Study; Cognitive Dimensions

1 Introduction

According to Symantec, the healthcare accounted for 42% in the total number of attacks on electronic information systems in 2012 [1]. At 36% in 2013, healthcare continues to be the sector responsible for the largest percentage of disclosed data incidents by industry [2]. Almost identical breaches have occurred across Europe, North America and Asia [3]. Such incidents may affect the organisations' reputation and challenge the public confidence and trust toward organisations' information security management [4]. It is imperative for the organisations to learn from those incidents and take effective actions to improve the reliability and trustworthiness of their information security management systems. Learning from incidents enables the organisation to extract meaningful information from incidents, and use this information to improve security management systems [5]. Effective communication

mechanism is needed to synthesis the information from the incident into the security incident management system so as to prevent a similar incident.

Popular communication mechanisms include formal reports, less formal meetings, newsletters, emails, as well as presentations to management [5]. However, the detailed incident reports that are produced in the post-incident activity [6] have not been given enough attention. Those reports contain comprehensive information, which is typically classified into two types, business impact and remediation information [6]. Business impact information involves how the incident is affecting the organisation in terms of mission impact, financial impact, etc. For example, "The missing external hard drive is believed to contain numerous research-related files containing personally identifiable information and/or individually identifiable health information for over 250,000 veterans, and information obtained from the Centres for Medicare & Medicaid Services (CMS), Department of Health and Human Services (HHS), on over 1.3 million medical providers" [7]. Remediation information mainly refers to the suggested remediation actions, plans, procedures, and lessons learned. For example, "We recommend that the Assistant Secretary for Information and Technology revise VA Directive 6601 to require the use of encryption, or an otherwise effective tool, to properly protect personally identifiable information and other sensitive data stored on removable storage devices when used within VA." [7].

As for a purpose of sharing, it is suggested to avoid sharing business impact information with outside organisations unless there is a clear value proposition or formal reporting requirements. When sharing information with peers and partner organisations, incident response teams should focus on exchanging remediation information [6]. The remediation information reported describes (1) the security issues, e.g. "The position sensitivity level for the IT Specialist was inaccurately designated as moderate risk, which was inconsistent with his programmer privileges and resulted in a less extensive background investigation", (2) the security objectives violated during this process, e.g. "Position Sensitivity Level Assessments were Not Adequately Performed", and (3) the recommendations, e.g."We recommend that the Under Secretary for Health direct the Medical Centre Director to re-evaluate and correct position sensitivity levels and associated background investigations for positions at the Birmingham VAMC " [7]. Those granular information are inter-related, however, they are scattered documented in a pure textual based report that makes it difficult for the readers to identify the relationships among them. This issue has been compounded by the lengthy security incident report, which is usually around hundred of pages [7]. The stakeholders responsible for protecting patient data lack the time and the motivation to spend the many hours needed to read and digest existing reports. This creates significant problems within the wider scope of security management systems. It can be difficult to accurately assess the likelihood or consequences of future attacks when managers are unaware of previous incidents, which undermines the trustworthiness of the systems.

Graphical techniques can address some of these limitations. The Generic Security Template (G.S.T.) has been developed [3, 8] to help readers understand the lessons learned from previous security incidents. In particular, it extends the Goal Structuring Notations (GSN) [9] to provide an overview of previous data breaches.

The intention is to map out the security objectives, security issues and recommendations that are embedded in the many pages of text that are used in conventional reports. More information on the GSN and the G.S.T. is provided in section 3. Figure 1 provides an excerpt from one of these diagrams. It is based on a report into the disclosure of personal information about 250,000 veterans and over 1.3 million medical providers by the US Veterans Affairs Administration (VA) [7]. This incident report provides the case study that is used throughout this paper. The leaf nodes in this diagram are used to gather together the recommendations that were intended to avoid future incidents. The internal nodes are used to show how each of these findings supports higher level goals and sub-goals intended to ensure that systems meet an acceptable level of security, defined in terms of the US Government's Federal Information System Controls Audit Manual (FISCAM) [10]. Further information about the graphical technique is provided in [3, 8, 11]. The use of graphical overviews is intended to make it easier to identify recommendations that can be transferred from a previous incident to prevent similar breaches from occurring in other organisations.

Previous work has shown that the G.S.T. can be used to map common lessons from data breaches in healthcare organisations in both the United States and in China [3]. Although these incidents occurred in very different contexts, the security concerns and the consequences for patient confidentiality show remarkable similarities. Previous work provided initial case studies but did not, present empirical support to evaluate the usability of the G.S.T.

This paper, therefore, presents a series of empirical studies to evaluate the G.S.T. The remainder of the paper is structured as the following, section 2 reviews the related work, section 3 briefly introduces the Generic Security Template, section 4 presents the first empirical study, section 5 presents the second study, and section 6 summarises the paper.

2 Related Work

2.1 Trust management

There are a variety of definitions of trust [12]. Two common definitions are reliability trust [13] and decision trust [12]. In this paper, we refer to the reliability trust that is defined as the users' subjective expectation about a service provider to perform a given action on which its welfare depends [13, 14]. The public expects the healthcare organisation to securely protect their medical record while processing this information for different purposes such as medical diagnosis and researches. However, the repeats of the security incidents have undermined the public's trust in the healthcare organisations information security management systems. There is a need for the organisations to learn the lessons from the security incidents and demonstrate to the public that those incidents have been treated seriously and are under control.

2.2 Sharing of security lessons

NIST [15, 16] and SANS [17, 18] have stressed the importance to share security lessons. They require that the insights from previous security breaches are documented, reviewed, presented and integrated back into the incident response process

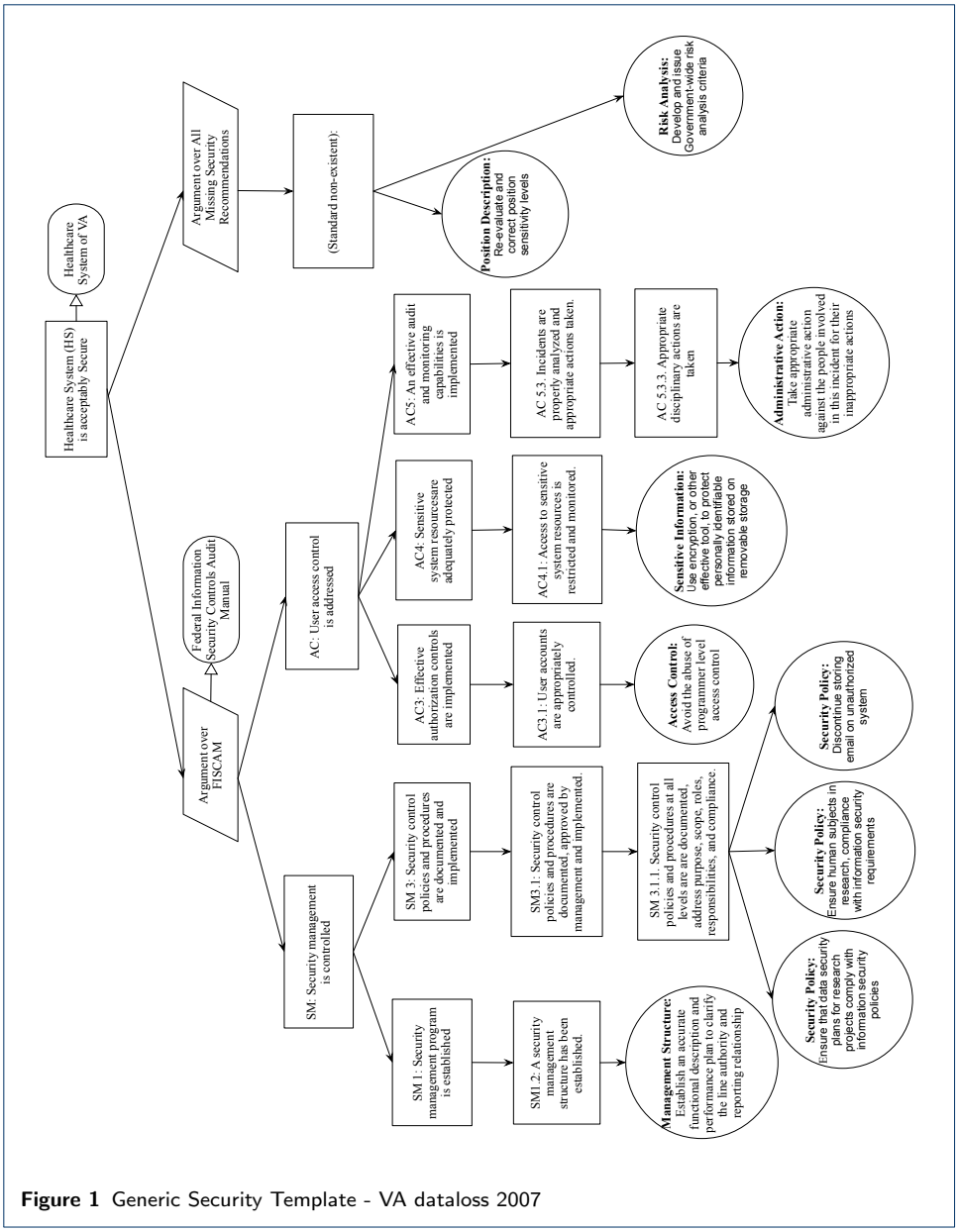

Figure 1 Generic Security Template - VA dataloss 2007

for future improvement. A growing number of regulatory agencies now provide detailed reports that are intended to help avoid any recurrence of previous failures. Security management systems have also been introduced into many healthcare organisations to ensure previous security incidents inform threat and risk assessments [19]. Improving situation awareness, in particular about security breaches, help persuade end users of the importance of existing policies and procedures. There are further benefits from the wider dissemination of incident reports. Security engineers can learn important lessons about the analysis, containment, eradication, and recovery from previous attacks [16].

2.3 Evaluation of Graphical Notations

The introduction has argued that existing, text-based reports can be supported through the use of graphical notations that provide an overview of many dozens of pages of detailed prose. It is also hoped that this use of a semi-formal notation will encourage greater consistency and correctness [20, 21]. Although graphical representations are often perceived as easier to understand, it can be difficult for readers to interpret the meaning of abstract symbols [22, 23]. The notation introduces unfamiliar syntax and semantics. There is a danger that our use of these techniques can prevent stakeholders from understanding the arguments in security incident reports [24, 25, 26]. This paper, therefore, presents a series of empirical studies to evaluate the usability of graphical representations of security incident reports.

3 The Generic Security Template

As mentioned, the G.S.T. extends the Goal Structuring Notations (GSN) [9] to provide an overview of previous security breaches. GSN is the dominant approach in the UK defence sector, increasingly being used in safety-critical industries to improve the structure, rigor, and clarity of design requirements. A particular strength is that it also links the evidence to show that particular requirements have been met. The same approach has more recently been extended to document security requirements [3, 8]. There are four principal notations used in the GSN, A *Goal* is a claim, the statements that the goal structure is designed to support. *Evidence* exists to support the truth of the claimed goal, which can be documented by providing a solution in GSN. *Strategy* is inserted between goals at two levels of abstraction, to explain how the top-level goal is addressed by the aggregation of the goals presented at the lower level. *Context* is used to declare supplementary information and provide adequate understanding of the context surrounding the claim (or strategy). Usually it presents concepts clarification introduced in the claim (or strategy) [9].

The G.S.T. has customised the GSN. Instead of collecting evidence to support design and development requirements, it collects lessons (i.e. security causes and recommendations) from previous security incidents. These lessons are defined as the knowledge or understanding gained by experience [27]. In the G.S.T., they refer to the security issues that cause a security breach, and the security recommendations intended to avoid any recurrence. The evidence of compliance with the security objectives is presented in the form of a specific security standard or guideline applied to the organisation where the security incident happened. This has reflected the granular information described in section 1. Generic, is defined as “characteristic of

or relating to a class or group of things; not specific". In other words, the intention is to create a GSN diagram that conveys the lessons learned from specific previous security breaches at a level of abstraction that helps others to use them to improve the security of other systems.

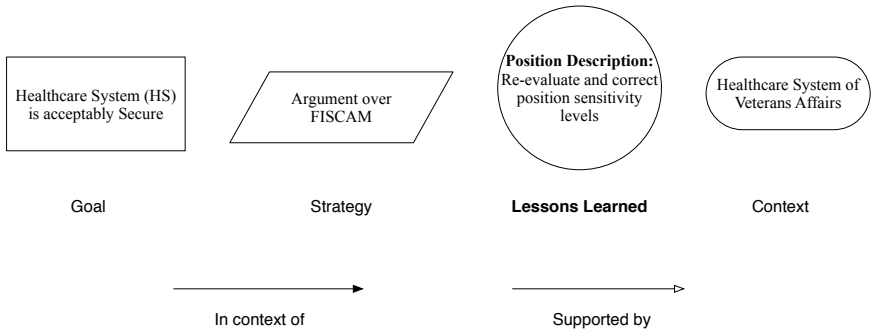

Figure 2 Customised GSN Notations

Figure 2 presents the notations used in the Generic Security Template. In particular, rather than using the evidence derived from validation and verification to support safety arguments, the G.S.T. uses the findings from previous data breach incidents (i.e. leaf nodes of Figure 1) to support security arguments in terms of the compliance with the security guideline (i.e. internal nodes of Figure 1). The other concepts remain the same between both application areas.

4 Empirical Study - Compare G.S.T. with Text-based approach

4.1 Experiment Objective and Scope

A controlled experiment was conducted to evaluate whether the use of graphical techniques helps improve the comprehension of the lessons from previous security incident reports compared to conventional text-based approaches. The aim was not to show that graphical techniques could replace conventional reports; in contrast the focus was in the use of the G.S.T. to provide a map or overview of complex text-based reports. Purely graphical representations are often less expressive than textual representation as some system properties cannot easily be specified using diagrams alone. A combined graphical representation with supporting textual representations can assist visualisation while still achieving the full expressiveness and precision of a textual representation [23]. Accuracy, efficiency and task load are compared quantitatively in this experiment and the following hypotheses are proposed for the comparison.

H1: Participants will be better able to identify the Lessons learned (security issues, security recommendations) in security incident report with the help of the G.S.T. than using Text-based Document alone;

H2: Participants will be better able to identify the compliance with the security objectives in security incident report with the help of the G.S.T. than using Text-based Document alone;

H3: The time taken to complete the designed task will be less using the G.S.T. than that using the text-based Document alone;

H4: The task load will be lower using the G.S.T. than using the text-based Document alone.

Ease of use is compared qualitatively based on the feedback obtained from participants.

4.2 Experiment Variables

4.2.1 Dependent Variables

We evaluate the usability [28] in terms of the accuracy, efficiency, ease-of-use and task load compared to the conventional, text-based approach. *Accuracy*, is measured by assessing the quality of the security causes, recommendations and the compliance with the security objectives from the security incident. *Efficiency*, is measured by the time it takes to complete the experiment task. *Ease of use*, is evaluated by the feedback obtained from the post-experiment questionnaire. *Task load*, is measured by the application of NASA's Task Load Index to assess workload [29].

4.2.2 Independent Variables

Generic Security Template (G.S.T.), we used the same G.S.T. across all participants. This presents findings from the US Veteran's Affairs administration 2007 Dataloss Incident [7]. *Text-based approach*, we developed an executive summary (reduced to four pages) and a simplified security guideline (reduced to three pages) from the FISCAM. More details on the experiment material preparation are provided in 4.3.

4.2.3 Controlled Variables

Participants, the participants were post-graduate and undergraduate students with different education background. *Tasks*, the experiment itself lasted for maximum one hour. Participants had to identify causes, recommendations and the relationships with security objectives using either a conventional text-based document or using the graphical overview plus the existing report.

4.2.4 Extraneous Variable

Experience with GSN, is defined as an extraneous variable in this experiment. People who have experience with GSN will have an obvious advantage in comprehending the security incident with the help of the G.S.T. People who have experience with GSN were excluded from this experiment.

4.3 Experiment Materials

4.3.1 Security Incident Report.

Security Incident Report. The technical context of the task focused on a data loss incident involving the Veterans Affairs' Administration [7]. The original report was around 80 pages long and hence we could not use it directly within the time available for the experiment. We also felt that our more focused approach was more appropriate for an initial study that could, in turn, inform future empirical work over a longer period of time and with a larger number of participants. We, therefore,

provided both groups with the executive summary from the VA report reduced to four pages. As is mentioned, the evidence of compliance with the security objectives is presented in the form of a specific security standard or guideline applied to the organisation where the security incident happened. Therefore, a simplified version of security guidelines (reduced to three pages) cited from the FISCAM that are relevant to this incident are also provided as a part of the security incident report.

4.3.2 The G.S.T.

The G.S.T. used in this experiment is created from the above mentioned security incident report only. It is an abstraction and extraction of the desirable information and did not bring any other information that can bias the results of the experiment.

4.3.3 The Questionnaire

We developed separate tasks description for the two groups and a post-experiment questionnaire, to provide subjective insights into perceived workload. A slightly different version of this post-experiment questionnaire was developed for the group using the graphical overview of the security incident. They were asked to provide information about the usability of the approach by completing subjective questionnaire.

4.4 The Pilot Study

Two security experts reviewed the design of the experiment pilot studies then helped to identify issues that had not been identified during the preparation of the materials. For example, they helped adjusting the complexity of the case study and suggested to reduce the incident report to four pages as mentioned in 4.3.1. In the first pilot study, participants had to identify security issues, recommendations and compliance with the security objectives; writing them down using freestyle text. This was to simulate how security incident reports are analysed in practice, where people normally have no tools assisting them throughout this process. The feedback from the participants showed that the task was very mentally demanding and they were not able to complete it within one hour. We corrected this problem by introducing a table that provided guidance on the security issues and recommendations. The participants are required to fill in the blanks cells in the table. For the measurement of compliance with the security objectives, we have used multi-choice questions as the measurement mainly focuses on the relationships between the security objectives and the recommendations for prevention. Two more participants conducted a pilot test of the new experiment design. They were able to finish the tasks and stated that the level of mental effort was acceptable.

4.5 Experiment Task Design

In Group A, the experiment materials included the textual incident report (reduced executive summary and reduced security guidelines from FISCAM), the graphical G.S.T. and a task description. The pilot study had confirmed the arguments presented in the opening sections of this paper; that it can be difficult for readers to identify the causes, recommendation and the compliance with the security objectives of previous security incidents from existing textual reports. We, therefore, created tasks that guided the participants' analysis:

Task 1: Identify security issues and recommendations from the security incident report with the help of the G.S.T. They had to complete missing information from a table that provided partial information about the causes and recommendations. Table 1 is an exempt of the table. Issue Category and Description are provided. The participants need to fill in the blank about the recommendation description.

Table 1 An exempt of the security issue and recommendation table

Task 2: Answer multiple-choice questions on compliance of the security objectives. This removed the additional contextual support of the tabular format used in task one and provided a stepping stone towards the open ended analysis of security incident reports that proved problematic in the pilot studies.

In Group B, the experiment materials included the textual incident report without the G.S.T. but participants had the same task descriptions as the first group.

4.6 Experiment Procedures

4.6.1 Experiment Treatment

There was only one treatment in the experiment using a between groups (Group A and B) design. The empirical comparisons are between one group using a conventional text-based document and the other using the graphical overview as well as the existing report.

4.6.2 Participants

To conduct research involving human participants, this experiment adhered to the BPS ethical guidelines, and has been approved by the ethics committee of the redacted university (ref: CSE01098). Information sheets were disseminated to each school of the redacted university. This has provided suitable coverage of participants from different background across the university. The participants attended this study voluntarily. The participants completed the consent form before starting the study. Twenty-four subjects were recruted and assigned to either of the two experimental conditions – using the textual report only or using both the textual report and the graphical overview. Group A consists of one undergraduate student and eleven postgraduate students, within which three of them have information security experience; Group B has one undergraduate student and eleven postgraduate students, within which three of them have information security experience. Each of the group has three females and nine males. None of them have experience with GSN or G.S.T.

The use of the students is justified for pragmatic and also for ethical reasons. Interviews with healthcare and IT professionals in healthcare organisations revealed that many lack formal training in security incident reporting and analysis; they come from varied backgrounds [8]. The growing number of patient data breaches has also created enormous sensitivity; many employers are extremely unwilling to participate in studies of this nature even when anonymity is guaranteed [8].

4.6.3 Training of the Participants

A pre-scripted familiarisation tutorial was provided before the experiment. Participants from both Group A and B attended the same tutorial session. This was to

ensure that they received equal knowledge related to the handling of security incidents. The participants were introduced to the Goal Structuring Notations and G.S.T.

4.6.4 Experiment Execution

The experiment was conducted on a one-to-one mode to provide any support needed during the whole process including the familiarization tutorial session, the experiment session and the post-experiment questionnaire session. During the familiarisation tutorial session, the participant had unlimited time to study the material and to have any question clarified. The participants were allowed to refer to the tutorial document or notes. The participants were instructed to inform the experiment conductor if they had any trouble in understanding the questions. During the post-experiment questionnaire session, an informal interview was conducted to make sure their attitudes were consistent with the answers they have provided. They are also requested to write down their subjective feedback on the G.S.T.

4.7 Results - Prepare the data

4.7.1 Scoring Scheme for the Experiment Tasks

Sample answers for the experimental tasks were agreed on by two independent security experts.

4.7.2 Preparation for Task 1 - Open-ended questions

For Task 1, the answers expected were qualitative. The marking was based on the description of security issues and recommendations expected from the sample answers. The answers for each task were marked by two further independent experts (Rater A and B) using an agreed scoring scheme. The participants' answers were classified into four categories, which are "Correct", "Incomplete", "Wrong" and "Blank". A correct answer completely described the recommendation to support the given issue; incomplete answers show that the participant had a partial understanding of the recommendation, but lacked comprehension of an important aspect of it. Wrong answers showed that the participant did not understand a particular recommendation. Blank, no answer was provided at all. The following paragraph provides an example from task one:

The report identifies the security concern: "The IT Specialist was improperly given access to multiple data sources". An answer is marked as, *Correct*, if the participant states that the recommendation associated with this issue was to "Consider the conditions under which programmer level access may be granted for research project". A correct answer completely describes the recommendation to support the given issue; *Incomplete*, if the answer is stated as "Ensure the access control is appropriately granted". Incomplete answers show that the participant had a partial understanding of the recommendation, but lacked comprehension of an important aspect of it; *Wrong*, if the answer provided is not relevant to a particular recommendation. *Blank*, if no answer was provided at all.

Each participant was free to use his or her own words to describe the recommendations in this part of the study. The group identifiers were removed so that Rater A and B marked the answers without knowing whether or not the participants had access to the G.S.T. diagram.

4.7.3 Preparation for Task 2 - Multi-choice questions

Task 2 used multi-choice questions to examine the participant's ability in understanding the compliance with the security objectives. Less subjectivity was involved in interpreting the answers. There can be more than one correct choice for each question and participants were asked to select all of the responses they believe were relevant to the questions. Below is an example.

What are the security recommendations for addressing the security objective "User Access Control"?

a. Develop and implement policies describing the conditions under which programmer level access may be granted for research purposes.

b. Effective procedures are implemented to determine compliance with authentication policies.

c. Attempts to log on with invalid passwords are limited. Use of easily guessed passwords (such as names or words) is prohibited.

d. None of the above

Correct answer: a, b

The sample answers were prepared by the independent security expert A. Each answer was classified as, *Correct*, *Correct but broad*, *Incomplete*, *Incomplete and broad*, *Wrong*, and *Blank*. A *correct* answer contained and only contained all the acceptable choices (e.g. a, b); *Correct but broad* contained all the acceptable choices, but also incorrect choices (e.g. a, b, c); *Incomplete* answers contained only some of the acceptable choices but not all (e.g. a). *Incomplete and broad* answers contained some of the acceptable choices and also other choices. (e.g. a, c); *Wrong* answers contained none of the acceptable choices (e.g. c). There was only one blank answer out of 144 responses; therefore we ignore this in the subsequent analysis.

4.8 Results - Analysis

4.8.1 Results for Accuracy (Task 1)

Since the results are categorical data, we use cross-tabulation analysis to analyse the results. A data set with 168 rows was imported into SPSS. Within the cross-tabulation analysis, groups were set as rows and task results were set as columns. Chi-square statistics was selected to test the hypothesis. Recall that these open ended questions were assessed by two independent raters. For Rater A, as is shown in Table 2, the results from the cross-tabulation analysis show that 62.7% of the responses from Group A were correct, which is 18.8% higher than Group B. This might seem a relatively low level of accuracy. However, it is important to recall that our marking scheme was careful to distinguish between complete, perfect responses and partially correct or incomplete answers. The total percentage of incomplete and correct answer is 81.9% in Group A, which is 16% higher than Group B. As is shown in Table 3, the Chi-Square Test ($P = 0.031 < 0.05$) shows that these results are statistically significant. Therefore, hypothesis H1 "Participants will be better able to identify the recommendations and causes in security reports with the help of a graphical method than using text alone" is supported based on Rater A's

judgement.

Table 2 The performance of Task 1 using Cross-tabulation by Rater A

Table 3 Chi-Square Tests performance of Task 1 using Cross-tabulation by Rater A

For Rater B, as is shown in Table 4, the results from the cross-tabulation analysis show that 65.1% of the responses from Group A were correct, which is 20% higher than Group B. The total percentage of Incomplete and Correct answer is 83.1% in Group A, which is 9.6% higher than Group B. As is shown in Table 5, the Chi-Square Test (P = 0.019 < 0.05) shows that these results are statistically significant. Therefore, hypothesis H1 "Participants will be better able to identify the recommendations and causes in security reports with the help of a graphical method than using text alone" is again supported based on Rater B's judgement.

Table 4 The performance of Task 1 using Cross-tabulation by Rater B

Table 5 Chi-Square Tests performance of Task 1 using Cross-tabulation by Rater B

Since these open ended questions were assessed by two independent raters, inter-rater reliability was checked for each question in Task 1. The results in Table 6 - 12 shows that the two raters have achieved agreements on judging the accuracy of the lessons learned identified by the participants and the results are statistically significant ($Approx.Sig.$ < 0.001). Landis and Koch proposed the benchmark scale on how the extent of agreement among raters should be interpreted and how the extent of agreement among raters should be interpreted [30], as is shown in Table 13. They have recommended this as useful guideline and Everitt also supported this benchmark scale [31]. Questions 1 ($KappaValue$ = 0.706), 3 ($KappaValue$ = 0.715), 5 ($KappaValue$ = 0.723), and 6 ($KappaValue$ = 0.782) have achieved "substantial agreement". Questions 2 ($KappaValue$ = 0.801) have achieved "almost perfect agreement"; Questions 4 ($KappaValue$ = 0.574) and 7 ($KappaValue$ = 0.497) has achieved "Moderate agreement".

Table 6 Inter-rater reliability for Task1 Question 1 (Rater A and B)

Table 7 Inter-rater reliability for Task1 Question 2 (Rater A and B)

Table 8 Inter-rater reliability for Task1 Question 3 (Rater A and B)

Table 9 Inter-rater reliability for Task1 Question 4 (Rater A and B)

Table 10 Inter-rater reliability for Task1 Question 5 (Rater A and B)

Table 11 Inter-rater reliability for Task1 Question 6 (Rater A and B)

Table 12 Inter-rater reliability for Task1 Question 7 (Rater A and B)

Table 13 Landis and Koch-Kappa's benchmark scale

4.8.2 The Results for Accuracy (Task 2)

As is shown in Table 14, the results from the cross-tabulation analysis show that the participants from Group A achieved a 33.3% accuracy rate, which is 9.7% higher than Group B. The total percentage of Correct, Broad, Incomplete, and Incomplete but broad answer is 87.5%, which is 18.1% higher than Group B. As is shown in Table 15, the Chi-Square Test ($P = 0.038 < 0.05$) shows that these results are statistically significant. Therefore, hypothesis H2 "Participants will be better able to identify the security arguments on the supportive relationships between the lessons and the security requirements with the help of the G.S.T. than using text-based document alone;" is supported in Task 2.

Table 14 The performance of Task 2 using Cross-tabulation

Table 15 Chi-Square Tests performance of Task 2 using Cross-tabulation

4.8.3 The Results for Efficiency (Time)

The mean total time used by Group A was almost equal that in Group B; 47.3 versus 47.8 minutes. The total time taken across all tasks is not statistically significant ($P = 0.932 > 0.05$). Therefore, we can accept the null hypothesis that "the mean time taken to complete our experimental tasks using a textual security incident report and a textual report with a graphical overview are not significantly different." Hypothesis H3 is not supported. One interpretation of these results is that significant time is required to understand security incidents, irrespective of whether they are presented in graphical or textual format. However, this would require further empirical support to determine whether or not other graphical notations might lead to significant differences in the time taken to understand security incident reports. It is also important for further work to consider the learning effects that might be expected through repeated use of the graphical maps.

4.8.4 The Results for Task Load Index (TLX)

We used NASA's Task Load Index [29] to assess workload using a post-evaluation questionnaire. The t-test results show a significant difference ($P = 0.047 < 0.05$) in the first dimension of the task load index regarding "how mentally demanding was the whole task". With a mean value of task load, 12.75 versus 15.50, participants expressed a lower subjective level of workload in terms of "mentally demand" when using the G.S.T. The results for the other four dimensions of the Task Load Index are not significantly different. However, a more sustained analysis is required to replicate these findings across a wider range of workload measures and with a larger sample of potential users.

4.9 Subjective Feedback

In Group A, approximately half of the participants expressed some difficulty in understanding the text based Security Incident Report. Half of the participants reported that they have no difficulty in completing task 1 of Group A: identifying security elements from the security incident report with the help of the G.S.T. Group B reported a slightly higher level of understanding of the Security Incident Report. However, less than half of the participants suggested that they have no difficulty in completing task 1 of Group B: identifying lessons learned from the security incident report, and the rate is much lower than that of Group A. These subjective findings are consistent with the quantitative results in section 6.3.

The participants' answers to the open questions regarding the overall experience of using the graphical overviews suggested that a longer training session might have helped them to better prepare for the tasks. Several participants mentioned that they had experienced learning effects; their confidence in answering the questions increased as they worked their way through the questions. This finding from Group A reveals generally positive feedback for the G.S.T. Group B did not use the G.S.T. during the experiment. They were asked to review the G.S.T. after the experiment and provide the feedback by completing Questionnaire Section 6 designed for Group B. Almost all of them suggested that they would have no difficulty in understanding the G.S.T. and agreed that the G.S.T. can help them better comprehend existing security incident reports. Two thirds of the participants reported their willingness to use the G.S.T. if they are requested to do a similar task in future. "It will help to understand terminologies security elements easily, less confusing, very structured and don't have to waste time, most importantly very easy to understand with less information". In summary, the participants' overall experience with the G.S.T. is positive, however, questions remain about the ability of participants to apply the lessons from the report within their own organisation rather than answering directed questions about the contents of a security report.

4.10 Validity Analysis

Threats to validity [32] are factors other than the independent variables that can affect the dependent variables.

4.10.1 Internal validity

Internal validity is concerned about the cause-effect relationships induced from the study. *Maturity effects*, there is a threat that the participants would tend to be bored and performed worse towards the end of the experiment session. However, we do not think that maturity effects will have undermined the validity of our results. As mentioned previously, several participants reported that their confidence in using the different reporting formats increased as they progressed through the tasks. *Learning effect*, there was not a learning effect in this experiment as there was only one treatment. *Testing effects*, all participants have studied the same material in the familiarisation tutorial session. Very few students have experience with security incident analysis. There was not any cheating because the experiment was on a one-to-one session. *Instrument effects*, the participants were given the same type of tasks and the answers were evaluated by the same marking scheme.

Evaluator bias was addressed through the use of two independent security experts during the assessment phase. The group identifiers were removed so that Rater A and B marked the answers without knowing whether or not the participants had access to the G.S.T. diagram.

4.10.2 External validity

External validity is the possibility to generalise the results beyond the current experiment. We addressed these concerns by selecting a broad cross section of participants including individuals with diverse background to reflect the those of managers and technical staff who must cooperate to implement the recommendations in security incident reports. The participants were undergraduate and postgraduate students. Using students in such experiment is common for practical reasons when the professionals are less available and expensive. However, the generalisation of the results to different target groups needs to be carefully considered.

5 Extended Study - Evaluation using Cognitive Dimensions

The main empirical study has identified the benefits and difficulties using the G.S.T. However, it has not systematically evaluated the usability of G.S.T. as graphical notations. The extended study further explores the strength and weaknesses of the G.S.T. using Cognitive Dimensions framework [33].

5.1 Cognitive Dimensions framework

Cognitive Dimensions framework provides a generic approach to measure various usability characteristics of notations and their environments [33]. Previous research has argued the importance of using Cognitive Dimensions to evaluate graphical notations [34] and this approach has been assessed for validity and reliability by a number of other researchers [35, 36, 37]. There are fourteen dimensions in the full framework and below provides an example of the Visibility dimension,

Example:

(Visibility) It is easy to see or find the various parts of the Generic Security Template while it is being used?

A. Strongly disagree B. Disagree C. Neutral D. Agree E. Strongly agree

Explain what kind of things is difficult to see or find?

For our study, we did not ask about the creation or modification of the notation. Therefore we have selected five dimensions, which include Visibility, Diffuseness, Hard Mental Operation, Closeness of Mapping, and Role Expressiveness. The evaluation questionnaire was based on the Cognitive Dimensions of Notations Usability Framework.

5.2 Participants

The extended study is a follow-up study of the main empirical study. We interviewed the twelve participants in Group A. They already had experience using the G.S.T. Each individual contact lasted for approximately 20 minutes. The use of the Cognitive Dimensions aimed to identify the strengths and weakness of the G.S.T., for example, any strength that the user is in favour of, any weakness that affects usability, any opportunity for further improvements.

5.3 Results

CD-Visibility Dimension. Ten out of twelve of the participants agreed or strongly agreed that "It is easy to see or find the various parts of the G.S.T. while it is being used". One participant disagreed and argued about the visibility of the goal structure. The comment was "might be difficult to differentiate between goals and sub goals". The suggestions were "use of colour may help visual interpretation" and "introduction of colours to identify the different levels/layers".

CD-Diffuseness Dimension. Eleven out of twelve of the participants agreed or strongly agreed that "the G.S.T. lets you say what you want reasonably brief". There was one participant against it and the reason was "too many words". Too much information will undermine the effectiveness of the graphical presentation, while too little information will make it difficult to understand. Future work should focus on identifying an appropriate level of abstraction of the G.S.T.

CD-Hard Mental Operation Dimension. Six out of twelve of the participants disagreed that "There seem some things especially complex or difficult to understand in your head while using the G.S.T.". Two out of twelve of the participants agreed or strongly agreed and stated this was caused by "too many words within one notation". They suggested to "separate recommendation into different or individual circles".

CD-Closeness of Mapping Dimension. Nine out of twelve of the participants agreed or strongly agreed that "the G.S.T. describes the problem accurately and completely for the security incident stated in the textual document". There was one participant against it and the feedback was "the case is not generic enough", this is consistent with the comments on CD-Hard Mental Operation Dimension "with many words". The participants also argue about separation of the recommendation notations, "it's in some cases hard to separate the individual solutions in one bottom node into separate issues", which had some overlap with the finding in the CD-Hard Mental Operation Dimension.

CD-Role Expressiveness Dimension. Seven out of twelve of the participants agreed or strongly agreed that "while reading the G.S.T., it is easy to tell what each part is for in the overall scheme". One participant disagreed with it. The feedback was "might be hard to see whether the user wants to work on the high or low level of the hierarchy". They suggested that "could use multiple cases" for different target groups with interest towards different level of information. This is consistent with the finding in an industiral evaluation of the G.S.T. They require a multi-view feature to meet the needs of different target groups including security managers, engineers and healthcare professionals [38].

5.4 Discussion

Table 16 summarises the suggested improvements identified during the evaluation of the G.S.T. using Cognitive Dimensions. One of the key questions for future work is to determine an appropriate level of abstraction for G.S.T. Too much information might affect the readers' motivation, and ability to analyse the causes of a previous incident. There may also be problems in navigating and interpreting the resulting graphical structures. While too little information will make it difficult to understand why an incident occurred and may provide insufficient contextual information to focus future interventions. There are some existing works on model abstraction. For example, Polyvyanyy proposed an abstraction slider to allows user control of the model abstraction level [39]. Smirnov presented abstraction approach, addressing specific features of BPMN [40]. Future work can focus on model abstraction as well as business intelligence [41, 42, 43] to generate lessons learned with a desirable level of details. However, questions remain on how to customise the G.S.T. to fit into the needs of particular healthcare organisations.

Table 16 Improvement suggestions for the G.S.T.

6 Conclusions and Future Work

The graphical G.S.T. captures the lessons learned from the security incidents and reasons about the information system security by attaching them to the security objectives of the organisation. It provides a way to communicate the information to different parties on how the remedial actions are taken in compliance with the organisational objectives as ways of avoiding future violations, hence increases their confidence and trust of organisations' information security management systems [3]. In this paper, we have presented the results derived from an initial study into the use of Generic Security Template to represent and reason about the recommendations made in a report of a data confidentiality breach involving the US Veterans' Affairs Administration. We were able to show significant benefits from the use of a graphical technique in answering a number of comprehension questions when compared to the more conventional use of text-based incident reports. However, we could not demonstrate any significant benefits in terms of the time taken to complete our experimental tasks.

A list of suggestion to improve the Generic Security Template had been identified using the Cognitive Dimensions and from the subjective feedback. There are recommendations regarding the visibility of the Generic Security Template, to add colour to the Generic Security Template to improve the visualisation, and decomposition of the lessons learned notation to decompose the complex lessons learned notation that contains more than one learning points and the multi-view design for different target users. We will consider those recommendations in the future design of the Generic Security Template. In particular the use of students is a limitation, healthcare security professionals need to be involved in future validation.

Our work yields important insights into the difficulties that engineers face when trying to understand the implications that previous security incident reports have for their own organisations. There have been numerous empirical studies to evaluate the usability of graphical notations, including Entity-Relationship diagrams

[44], UML [45, 46] etc. However, as far as we are aware, there have been no previous studies to assess the strengths and weaknesses of graphical notations to help transfer the lessons learned from previous security incidents. These studies are urgently needed as both the Obama administration and the European Commission have recently published proposals to support the mandatory reporting of security incidents across national critical infrastructures, including healthcare. This provides the foundation for future work to further evaluate this approach with people working in healthcare industry.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
The first and second authors have equal contributions to all the sections of this article. The third author contributes to section 4.7 of this article.

Acknowledgements
We would like to thank Scottish Informatics and Computer Science Alliance (SICSA) for funding this research work.

References

1. Symantec: Internal security threat report 2011 trends, vol. 17 (2012)
2. Symantec: Internet security threat report 2013, vol. 18 (2013)
3. He, Y., Johnson, C.: Generic security cases for information system security in healthcare systems. In: Proceedings of the 7th IET International Conference on System Safety, Incorporating the Cyber Security Conference (2012)
4. Ahmad, A., Hadgkiss, J., Ruighaver, A.B.: Incident response teams-challenges in supporting the organisational security function. Computers & Security **31**(5), 643–652 (2012)
5. Hadgkiss, J.: Computer security incident response teams: Exploring the incident learning capability. (2004)
6. Hadgkiss, J.: Computer security incident handling, step-by-step. (1997)
7. Veterans' Affairs Administration, U.: Administrative investigation loss of VA information VA medical center birmingham, al, vol. Report No. 07-01083-157 (2007)
8. He, Y., Johnson, C., Renaud, K., Lu, Y., Jebriel, S.: An empirical study on the use of the generic security template for structuring the lessons from information security incidents. In: Proceedings of the 6th International Conference on Computer Science and Information Technology, pp. 178–188 (2014)
9. Kelly, T.P.: Arguing safety - A systematic approach to safety case management. (1998)
10. Dacey, R.F.: Federal Information System Controls Audit Manual (FISCAM), (2010)
11. He, Y., Johnson, C., Lu, Y., Lin, Y.: Improving the information security management: An industrial study in the privacy of electronic patient records. In: The 27th International Symposium on Computer-Based Medical Systems (2014)
12. McKnight, D.H., Chervany, N.L.: The meanings of trust (1996)
13. Gambetta, D.: Can we trust trust. Trust: Making and breaking cooperative relations **2000**, 213–237 (2000)
14. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision support systems **43**(2), 618–644 (2007)
15. West-Brown, M.J., Stikvoort, D., Kossakowski, K.-P., Killcrece, G., Ruefle, R.: Handbook for computer security incident response teams (csirts). Technical report, DTIC Document (2003)
16. Grance, T., Kent, K., Kim, B.: Computer security incident handling guide. NIST Special Publication, 800–61 (2004)
17. Northcutt, S.: Computer security incident handling: Step by step, a survival guide for computer security incident handling. Sans Institute (2001)
18. Murray, J.: Analysis of the incident handling six-step process. In: SANS Reading Room (2007)
19. Commissioner, E.: DIRECTIVE 2009/140/EC of the European Parliament And Of The Council of 25 November 2009. (2009)
20. Craigen, D.: Formal methods technology transfer: Impediments and innovation. In: CONCUR'95: Concurrency Theory, pp. 328–332 (1995)
21. Hinchey, M.G.: Confessions of a formal methodist. In: SCS, pp. 17–20 (2002)
22. Bauer, M.I., Johnson-Laird, P.N.: How diagrams can improve reasoning. Psychological Science **4**(6), 372–378 (1993)
23. Stenning, K., Oberlander, J.: A cognitive theory of graphical and linguistic reasoning: Logic and implementation. Cognitive science **19**(1), 97–140 (1995)
24. Finney, K., Fedorec, A.: An empirical study of specification readability. Teaching and Learning Formal Methods, Academic Press, New York (1996)
25. Finney, K.: Mathematical notation in formal specification: Too difficult for the masses? Software Engineering, IEEE Transactions on **22**(2), 158–159 (1996)
26. Carew, D., Exton, C., Buckley, J.: An empirical investigation of the comprehensibility of requirements specifications. In: Empirical Software Engineering, 2005. 2005 International Symposium On, p. 10 (2005). IEEE
27. Weber, R., Aha, D.W., Becerra-Fernandez, I.: Intelligent lessons learned systems. Expert Systems with Applications **20**(1), 17–34 (2001)

28. Folmer, E., Bosch, J.: Architecting for usability: a survey. Journal of systems and software **70**(1), 61–78 (2004)
29. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): Results of empirical and theoretical research. Human mental workload **1**(3), 139–183 (1988)
30. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics, 159–174 (1977)
31. Everitt, B.S.: The analysis of contingency tables. vol. 45. CRC Press (1992)
32. Campbell, D.T., Stanley, J.C., Gage, N.L.: Experimental and Quasi-experimental Designs for Research, (1963)
33. Green, T.R.G., Petre, M.: Usability analysis of visual programming environments: a'cognitive dimensions' framework. Journal of Visual Languages & Computing **7**(2), 131–174 (1996)
34. He, Y., Johnson, C., Evangelopoulou, M., Lin, Z.-S.: Diagraming approach to structure the security lessons: Evaluation using cognitive dimensions. In: Trust and Trustworthy Computing, pp. 216–217 (2014)
35. Kutar, M., Britton, C., Barker, T.: A comparison of empirical study and cognitive dimensions analysis in the evaluation of UML diagrams. In: Procs of the 14th Workshop of the Psychology of Programming Interest Group (PPIG 14) (2002)
36. Triffitt, E., Khazaei, B.: A study of usability of z formalism based on cognitive dimensions. In: Proceedings of the 14th Annual Meeting of the Psychology of Programming Interest Group (PPIG, (2002). Citeseer
37. Blackwell, A.F., Green, T.R.: A cognitive dimensions questionnaire optimised for users. In: Proceedings of the Twelfth Annual Meeting of the Psychology of Programming Interest Group, pp. 137–152 (2000)
38. He, Y., Johnson, C.: Improving incident learning in healthcare: An industrial study in the protection of electronic patient records. In: International Journal of Medical Informatics (2014). under review
39. Polyvyanyy, A., Smirnov, S., Weske, M.: Process model abstraction: A slider approach. In: 12th International IEEE Enterprise Distributed Object Computing Conference, pp. 325–331 (2008)
40. Smirnov, S.: Structural aspects of business process diagram abstraction. In: IEEE Conference on Commerce and Enterprise Computing, pp. 375–382 (2009)
41. Bench-Capon, T.J., Dunne, P.E.: Argumentation in artificial intelligence. Artificial intelligence **171**(10), 619–641 (2007)
42. Mackinlay, J.: Automating the design of graphical presentations of relational information. ACM Transactions on Graphics (TOG) **5**(2), 110–141 (1986)
43. Negnevitsky, M.: Artificial Intelligence: a Guide to Intelligent Systems, (2005)
44. Shoval, P., Shiran, S.: Entity-relationship and object-oriented data modeling-an experimental comparison of design quality, pp. 297–315 (1997)
45. Glezer, C., Last, M., Nachmany, E., Shoval, P.: Quality and comprehension of UML interaction diagrams-an experimental comparison, pp. 675–692 (2005)
46. Razali, R., Snook, C., Poppleton, M., Garratt, P., Walters, R.: Usability assessment of a UML-based formal modelling method. In: 19th Annual Psychology of Programming Workshop (PPIG'07), pp. 56–71 (2007). Citeseer

Table 1 An exempt of the security issue and recommendation table

Issue Category	Issue description	Recommendations description
Access Control Related	The IT Specialist was improperly given access to multiple data sources.	

Table 2 The performance of Task 1 using Cross-tabulation by Rater A

		Wrong	Incomplete	Correct	Total
Group A	Count	15	16	52	83
	within Group	18.1%	19.3%16	62.7%	100.0%
Group B	Count	28	18	36	82
	within Group	34.1%	22.0%16	43.9%	100.0%
Total	Count	43	34	88	165
	within Group	26.1%	20.6%16	53.3%	100.0%

Table 3 Chi-Square Tests performance of Task 1 using Cross-tabulation by Rater A

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.951a	2	.031
Likelihood Ratio	7.029	2	.030
Linear-by-Linear Association	6.909	2	.009
N of Valid Cases	165		

Table 4 The performance of Task 1 using Cross-tabulation by Rater B

		Wrong	Incomplete	Correct	Total
Group A	Count	14	15	54	83
	within Group	16.9%	18.1%16	65.1%	100.0%
Group B	Count	28	17	37	82
	within Group	34.1%	20.7%16	45.1%	100.0%
Total	Count	42	32	91	165
	within Group	25.5%	19.4%16	55.2%	100.0%

Table 5 Chi-Square Tests performance of Task 1 using Cross-tabulation by Rater B

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.962a	2	.019
Likelihood Ratio	8.071	2	.018
Linear-by-Linear Association	7.911	1	.005
N of Valid Cases	165		

Table 6 Inter-rater reliability for Task1 Question 1 (Rater A and B)

	Value	Asymp.Std. Error a	Approx. Tb	Approx. Sig.
Measure of Agreement Kappa	.706	.117	4.254	.000
N of Valid Cases	24			

Table 7 Inter-rater reliability for Task1 Question 2 (Rater A and B)

	Value	Asymp.Std. Error a	Approx. Tb	Approx. Sig.
Measure of Agreement Kappa	.801	.105	5.415	.000
N of Valid Cases	23			

Table 8 Inter-rater reliability for Task1 Question 3 (Rater A and B)

	Value	Asymp.Std. Error a	Approx. Tb	Approx. Sig.
Measure of Agreement Kappa	.715	.127	4.786	.000
N of Valid Cases	24			

Table 9 Inter-rater reliability for Task1 Question 4 (Rater A and B)

	Value	Asymp.Std. Error a	Approx. Tb	Approx. Sig.
Measure of Agreement Kappa	.574	.128	3.962	.000
N of Valid Cases	22			

Table 10 Inter-rater reliability for Task1 Question 5 (Rater A and B)

	Value	Asymp.Std. Error a	Approx. Tb	Approx. Sig.
Measure of Agreement Kappa	.723	.120	4.796	.000
N of Valid Cases	24			

Table 11 Inter-rater reliability for Task1 Question 6 (Rater A and B)

	Value	Asymp.Std. Error a	Approx. Tb	Approx. Sig.
Measure of Agreement Kappa	.782	.104	5.325	.000
N of Valid Cases	24			

Table 12 Inter-rater reliability for Task1 Question 7 (Rater A and B)

	Value	Asymp.Std. Error a	Approx. Tb	Approx. Sig.
Measure of Agreement Kappa	.497	.154	3.251	.001
N of Valid Cases	24			

Table 13 Landis and Koch-Kappa's benchmark scale [30]

Kappa Statistic	Strength of Agreement
Less than 0.0	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

Table 14 The performance of Task 2 using Cross-tabulation

		Wrong	Incomplete and broad	Incomplete	Broad	Correct	Total
Group A	Count	9	11	19	9	24	83
	within Group	12.5%	15.3%	26.4%	12.5%	33/3%	100.0%
Group B	Count	22	4	18	11	17	82
	within Group	30.6%	5.6%16	25.0%	15.3%	23.6%	100.0%
Total	Count	31	15	37	20	41	165
	within Group	21.5%	10.4%16	25.7%	13.9%	28.5%	100.0%

Table 15 Chi-Square Tests performance of Task 2 using Cross-tabulation

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.140a	4	.038
Likelihood Ratio	10.449	4	.034
Linear-by-Linear Association	2.995	1	.084
N of Valid Cases	144		

Table 16 Improvement suggestions for the G.S.T.

Improvements	Dimentions	Comments
Use of Color	CD-Visibility	... use of color may help visual interpretation; ... introduction of colors to identify the different levels/layers....
Level of abstraction	CD-Diffuseness	...too many words...; ...too many words within one notation...; ...the template is not generic enough...
Separation of notations	CD-Closeness of Mapping; CD-Hard Mental Operation	... separate recommendation into different or individual circles; ... the recommendation part needs to be simplified or separate individually [under suitable category]; ... it is in some cases hard to separate the individual solutions in one bottom node into separate issues...
Multi-view	CD-Role Expressiveness	... could use multiple templates; ... might be hard to see whether the user wants to work on the high or low level of the hierarchy...