






ConGISATA: A Framework for Continuous Gamified Information Security Awareness Training and Assessment

Ofir Cohen^(✉), Ron Bitton, Asaf Shabtai, and Rami Puzis

Software and Information Systems Engineering and Cyber@BGU,
Ben-Gurion University of the Negev, Beer-Sheva, Israel
{cofir,ronbit}@post.bgu.ac.il, {shabtaia,puzis}@bgu.ac.il

Abstract. The incidence of cybersecurity attacks utilizing social engineering techniques has increased. Such attacks exploit the fact that in every secure system, there is at least one individual with the means to access sensitive information. Since it is easier to deceive a person than it is to bypass the defense mechanisms in place, these types of attacks have gained popularity. This situation is exacerbated by the fact that people are more likely to take risks in their passive form, i.e., risks that arise due to the failure to perform an action. Passive risk has been identified as a significant threat to cybersecurity. To address these threats, there is a need to strengthen individuals' information security awareness (ISA). Therefore, we developed ConGISATA - a continuous gamified ISA training and assessment framework based on embedded mobile sensors; a taxonomy for evaluating mobile users' security awareness served as the basis for the sensors' design. ConGISATA's continuous and gradual training process enables users to learn from their real-life mistakes and adapt their behavior accordingly. ConGISATA aims to transform passive risk situations (as perceived by an individual) into active risk situations, as people tend to underestimate the potential impact of passive risks. Our evaluation of the proposed framework demonstrates its ability to improve individuals' ISA, as assessed by the sensors and in simulations of common attack vectors.

Keywords: Information Security Awareness · Social Engineering · Human Factors · Gamification · Cybersecurity Training · Mobile Devices

1 Introduction

Defense mechanisms are deployed to prevent attackers from performing malicious activities such as hacking into networks, accessing sensitive information, and compromising computerized systems. Social engineering (SE) refers to techniques

aimed at manipulating people into performing actions that help an attacker bypass state-of-the-art defense mechanisms [1]. The ease with which the human factor can be exploited has resulted in numerous cyberattacks caused by human error [2, 20]. For mobile users, SE is one of the main attack vectors [24], and given the prevalence of smartphones today, SE poses a significant threat to society.

Approaches for mitigating the risk posed by cybersecurity attacks utilizing SE techniques consist of two essential components: assessing information security awareness (ISA) and improving it.

Various methods can be used to assess ISA, the most common being questionnaires [13, 14, 23]. However, questionnaires require users' active involvement and collaboration; moreover, they are subjective and prone to bias, as they rely on self-reported behavior [25]. Despite their widespread use, questionnaires have been shown to be an unreliable measurement tool for ISA [11].

Challenges involving simulations of common attacks are also used to measure ISA. The primary advantage of this type of assessment is that it measures users' ability to handle real-life attack scenarios. However, challenges also have a major limitation: they do not consider users' context (e.g., opening an email from home versus opening an email at work). Since human behavior often depends on the context, these methods are inherently less accurate [26].

To address the limitations of existing ISA assessment methods, Bitton *et al.* [8, 11] proposed a taxonomy for mobile users' security awareness that defines a set of measurable criteria organized by technological focus areas. These criteria are measured by a mobile agent that collects data from sensors on the users' devices. The sensors are mapped to the taxonomy's criteria, and a final passive ISA score is produced by aggregating their outputs. This ISA score can be changed dynamically based on continuous sensor readings. In this research, we use this sensor-based approach, along with challenges associated with three common attack vectors, to assess ISA.

Typically, ISA is improved by participating in security awareness programs (workshops) or performing challenges with feedback. However, the previously mentioned limitation also applies when challenges are used to improve ISA. In many cases, efforts aimed at improving ISA evoke fear, which has been shown to be counterproductive at times; such efforts also result in 'security fatigue,' in which people tire of being presented with security procedures and processes [3].

Gamification is a technique often used to overcome the limitations described above. Deterding *et al.* [4] defined gamification as "the use of game design elements in non-game contexts", and Hamari *et al.* [5] reviewed many gamification studies and concluded that this method works well in various fields, particularly for improving learning and training sessions. As a result, the use of gamification to increase ISA has grown, leading to the development of various gamified solutions for this purpose [16–18].

Nevertheless, standard gamification alone is insufficient. A literature review performed by Böckle *et al.* [27] highlighted the problem of the "one size fits all" approach, which may result in declining engagement and loss of interest in overly simple challenges. To overcome this, the authors suggested using adaptive

gamification that dynamically re-engages users. Our approach for improving ISA utilizes adaptive gamification through a personalized feedback loop tailored to the outputs of the sensors.

A behavioral aspect that was not considered in previously proposed gamified solutions is passive/active risk-taking. Studies have classified risks as either active or passive risks [9, 10]. Active risk describes actions people take that put them at risk, while passive risk is “*risk brought on or magnified by inaction or avoidance*”. One example from the cybersecurity domain is the risk of having malware on your mobile device. In its active form, this risk derives from the possibility of unknowingly downloading a malicious file, whereas in its passive form, it stems from failing to install anti-malware software on the device in advance. These studies showed that passive risks are perceived as being less risky than equivalent active risks. Therefore, our framework aims to reduce passive risk-taking (PRT), by transforming passive risks into active risks. By deducting game points from users who fail in a passive-risk-related scenario, we impose an immediate punishment on passive behaviors. By doing so, we can help users gradually overcome the human tendency to overlook passive risks.

In this research, we propose ConGISATA, a continuous gamified ISA training and assessment framework, which addresses the problems of existing gamification-based methods described above. Our approach is implemented by a mobile agent (an app) that collects data from the set of sensors used in the taxonomy and assessment method of Bitton *et al.* [8, 11]. The app has a graphical user interface with the key components of a gamified system: a detailed home screen, a leaderboard, and a learning screen. The learning screen is composed of sections, one for each criterion in the taxonomy. For each criterion, there is a score and a link to an article or blog post that should help users improve their behavior with regard to the criterion. The scores on this screen are updated daily according to the sensors’ readings and highlight the areas in which the user needs to improve. Challenges are also presented throughout the learning process to help assess users’ ISA as they progress.

To evaluate the proposed framework, we performed an extensive experiment involving 70 subjects, each of whom installed our mobile app on their smartphone and used ConGISATA for a period of five weeks. We compared our method with a baseline method inspired by methods commonly used in academia and industry today. In the baseline method, users were provided personalized articles/blog posts based on their performance in the challenges, without taking the sensor data into account. Our results show that users who were trained using the ConGISATA framework had greater improvement than those trained using the baseline method for almost all criteria of the ISA taxonomy. In addition, a significant correlation between the use of our app and users’ ISA improvement was observed. Importantly, by using simulations of three common attack vectors, we found that ConGISATA helps users deal with real-life SE scenarios.

Our contributions can be summarized as follows: (1) We propose a novel framework for improving and assessing mobile users’ ISA. (2) To the best of our knowledge, we are the first to take continuous sensor readings and show their

impact on improving ISA in an adaptive gamification setting. (3) We empirically demonstrate the importance of considering passive risk-taking in the ISA training domain.

2 Background

Active Versus Passive Risk-Taking: Keinan et al. [9] established passive risk as a unique and separate construct. The authors provide the following explanation: “People are often held less responsible for their omissions than for their commissions. This lack of perceived responsibility may lower the motivation to act. People are usually less likely to do something if they believe they will not be held accountable for failing to do it. However, risk aversion often increases with personal accountability, since accountability stimulates self-critical forms of thought and increases awareness of one’s own judgment processes. It seems plausible that once people feel accountable they process information better, realize that they are in a risky situation, and be motivated to act to avoid risk.”

A follow-up paper [10] showed that a passive risk is judged as less risky than a completely equivalent active risk. For example, the following scenario was presented in both active and passive forms: actively parking your car in a restricted zone or not moving your car once you realize it is parked in a restricted zone. When asked to rate scenarios by risk level, in its active form this scenario was rated as riskier than in its passive form.

The authors suggest that “this inferior ability to devote attention to the absence of events leads to passive risks being less available to our consciousness, to be underestimated, and thus to be perceived as less risky. We need to be motivated to devote attention to passive risks.” The authors add the following recommendation: “The findings of the current research suggest stressing people’s personal responsibility for complying with recommended preventive measures may raise risk perception and increase preventive action.”

Finally, Arend et al. [19] examined how self-reported passive risk behavior predicts cybersecurity behavioral intentions and their relation to actual cybersecurity behavior. This series of three studies showed that passive risk had a notable impact on cybersecurity intentions, meaning that high passive risk scores were associated with low adherence to safe cybersecurity behavior. It was also shown that behavioral choices related to cybersecurity are highly correlated with a tendency to take passive risks. Overall, these studies established that passive risk tendencies are an important factor in the context of cyber behavior.

3 Related Work

Every gamified approach for mitigating the risk posed by SE attacks consists of two essential components: measuring ISA and improving it.

Questionnaires are the most common means of measuring ISA, with the vast majority of prior studies on this topic relying on them [6, 12, 15, 29–32]. Despite their widespread use, they tend to be an unreliable measurement tool

for behavior because of their subjective nature. Additionally, they are prone to bias, as they rely solely on self-reported behavior [25].

A more advanced method of evaluating ISA is to use attack simulations (also referred to as challenges). Despite their inability to consider users' context, the employment of challenges to assess ISA during security awareness training is extremely valuable, as it provides important insights into authentic user behavior. Their application in the literature, however, has been limited [22]. Our framework utilizes three different types of challenges: phishing, permission attacks, and impersonation. When using our framework, users do not know when or how they are presented with these challenges; this ensures that the challenges are as natural and objective as possible. Additionally, the framework collects data from sensors in users' everyday environments to examine aspects of their ISA in real-life settings, outside of controlled laboratory conditions.

When it comes to improving ISA using gamification, two core elements distinguish current gamified solutions: training duration and personalization of the content. Most of the gamified training mentioned in the literature was performed for a single brief session and utilized a physical board/card game, which is a difficult requirement for long-term training (over a period of weeks) [12, 15]. We only identified one paper with a longer training process – that of Alahmari *et al.*, where the training took place for two weeks [28]. Our framework is designed to achieve long-term behavioral change through continuous learning over several weeks or months, without requiring physical attendance at training sessions.

Böckle *et al.* [27] highlighted the problem of the “one size fits all” approach in gamified solutions for improving ISA and suggested the use of personalization. Heid *et al.* [33] created a gamified prototype that poses questions related to security and privacy issues associated with apps installed on the user's smartphone. A quiz engine providing multiple choice questions regarding known vulnerabilities and app properties was implemented using Appcaptor, a mobile application analysis platform that performs static and dynamic app tests. The question engine automatically generates questions from Appcaptor's database content, which are personalized for the users based on the apps installed on their smartphones. However, this work is limited, because only one sensor served as a source of information, the proposed method was only tested within the research group, and it relies on an external data source that is not publicly available, preventing its reproducibility.

Our literature review failed to identify any other papers utilizing personalization besides the work of Heid *et al.* mentioned above. Our framework generates scores for each user based on their weaknesses, as measured using multiple sensors. We evaluated the framework's impact in a comprehensive experiment spanning several weeks. All the materials used are publicly available and presented in the appendix. Furthermore, our gamified solution is the only method that demonstrates how passive risks can be transformed into active ones, which is a key contribution of our research.

A summary of the related work is provided in Table 1.

Table 1. Summary of related work

Paper	Platform	Personalization	Considers PRT	Continuous Learning	Questionnaires	Attack Simulations	Sensors
Newbould <i>et al.</i> [12], 2009	Board game	✗	✗	✗	✓	✗	✗
Denning <i>et al.</i> [15], 2013	Tabletop card game	✗	✗	✗	✓	✗	✗
Gjertsen <i>et al.</i> [6], 2017	Exercises	✗	✗	✗	✓	✗	✗
Scholefield <i>et al.</i> [30], 2019	Mobile (Android) game	✗	✗	✗	✓	✗	✗
Dincelli <i>et al.</i> [29], 2020	Interactive storytelling	✗	✗	✗	✓	✗	✗
Heid <i>et al.</i> [33], 2020	Multiple choice quizzes	✓	✗	✓	✗	✗	✓
Omar <i>et al.</i> [31], 2021	Educational quizzes	✗	✗	✗	✓	✗	✗
Wu <i>et al.</i> [32], 2021	Multiple choice quizzes	✗	✗	✗	✓	✗	✗
Alahmari <i>et al.</i> [28], 2022	Mobile app	✗	✗	✓	✓	✗	✗
Canham <i>et al.</i> [22], 2022	Phishing simulations	✗	✗	✗	✓	✓	✗
Our method, 2023	Mobile (Android) game	✓	✓	✓	✗	✓	✓

4 Proposed Method

In this section, we present ConGISATA. First we provide a high-level description of the framework (illustrated in Fig. 1a), and then we elaborate on each component.

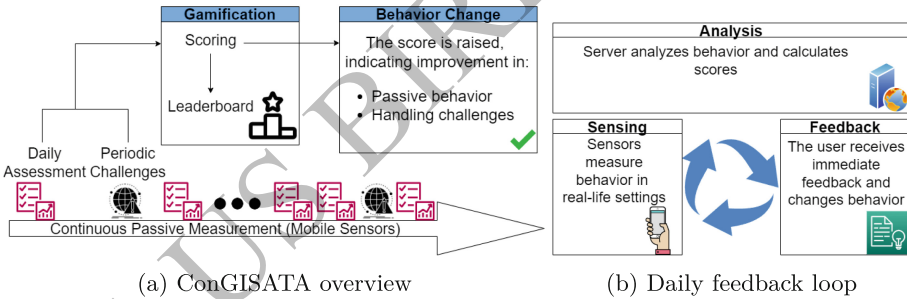


Fig. 1. The ConGISATA security awareness training and assessment framework

In the framework, the following steps are performed in a process aimed at raising ISA:

Calibration Period: For each user, the game starts with a calibration period in which the user’s initial security awareness score is assessed for each criterion in the taxonomy. This assessment does not require the user to interact with the game, as it is performed using the mobile sensors and challenges described in Sect. 4.1. Following this evaluation, the initial overall ISA score is presented to the user on the game’s home screen, and a score for each criterion is presented on the learning screen.

Training: In this step, the user starts to interact with the framework in an attempt to gradually improve their behavior and raise their ISA scores. Sensor measurement and challenges still occur in the background, resulting in daily changes to the user's ISA scores, which are presented to them. Training is performed cyclically, through a daily feedback loop, as follows:

Sensing - Each day, the different aspects of the users' behavior are measured by obtaining the sensor values.

Updating Scores - The application's learning screen is always accessible and displays the users' scores for each criterion in the ISA taxonomy, along with their overall ISA score. At midnight, these scores are updated to reflect the previous day's performance. The learning screen also presents the score delta for each criterion, which is the change in the score between two consecutive days. The score deltas enable the user to identify specific behavioral weaknesses (criteria with a negative score delta) and take corrective action.

Articles and Blog Posts - When the user is faced with a low score or a negative score delta for a criterion, they can obtain additional information about that specific focus area via the learning screen, which provides a link to an external, predetermined, and comprehensive article or blog post (for convenience, we refer to them as articles in the rest of the paper) on that subject.

Behavior Change - Upon reading the articles, the user will modify their behavior accordingly, improving their score over time, and climb the leaderboard.

Figure 1b illustrates the daily feedback loop of passive ISA.

4.1 Assessing Mobile Users' ISA

To generate an overall ISA score for each user, we measure two aspects of their behavior: active and passive. The active side refers to the user's ability to handle situations in which immediate action is required, as when facing an attack. Our framework measures this aspect using SE challenges. The passive side refers to ongoing elements of the user's behavior that do not result in an immediate punishment if not performed, such as using a lock screen or deleting unused apps to avoid malware. In our framework, we adapt the method proposed by Bitton *et al.* [8,11] to generate a passive ISA score. Instead, we generate an overall ISA score, which reflects both aspects, active and passive, as follows:

Assessing ISA Using Attack Simulations (Challenges): Each user is regularly presented with challenges derived from three attack vectors. These challenges assess the user's ability to handle real-life attack scenarios and ensure that this capability is also reflected in their overall ISA score. The challenges are presented in a randomized manner (in terms of both time and order) throughout the training process, to prevent detectable patterns. The active score denotes the user's performance on SE challenges and is based on a scale of zero to 100. The score is derived from a moving window of the last X challenges, where X is determined based on the training duration. Each challenge is individually scored between

zero, assigned for a failure to make the correct decision, and $100/X$, assigned for successful decision making. For example, for $X = 5$, each challenge can contribute at most $100/5 = 20$ points to the overall active score. Some challenges may involve two decision points, in which case the score assigned is $(100/X)/2$ if the user makes the correct decision at just one of the decision points. For example, a phishing challenge may include two decision points; the first is clicking on the unknown link to enter the phishing website, and the second is providing sensitive details such as login credentials. In such a case, if a user only clicks on the unknown link but does not provide any details, $(100/5)/2 = 10$ points will be added to the overall active score.

Assessing ISA Using Sensor Measurements: Bitton *et al.* [8] developed a taxonomy to measure mobile users' ISA that classifies criteria by technological focus areas and psychological dimensions. Each focus area is further divided into sub-focus areas, and each of these sub-focus areas encompasses several security topics. For instance, the "Applications" focus area is bifurcated into "Application Installation" and "Application Handling" sub-focus areas, with "Untrusted Sources" as a security topic under "Application Installation". The intersection of this security topic with the "Confronting Behavior" psychological dimension leads to a specific criterion: "*Installs applications solely from trusted sources*".

Bitton *et al.* [11] also proposed a framework for evaluating ISA, which employs a mobile agent with embedded sensors, a network traffic monitor, and cybersecurity challenges. Their framework, which is based on the ISA taxonomy, enables the computation of ISA scores at any given point. The study found that there was a difference between users' self-reported behavior and their actual behavior, highlighting the significance of monitoring real-life user behavior instead of relying solely on questionnaires.

In order to gain a deeper understanding of users' behavior in real-life scenarios, we included sensors based on the ISA taxonomy in our mobile application. These sensors periodically perform a thorough scan of users' devices and actions. By analyzing the resulting data, we can compute a user's passive ISA score and an individual score for each criterion, and identify their potential weaknesses. This knowledge allows us to provide the user with personalized feedback about their ISA scores and offer guidance on how to improve their security practices.

In the paper, we describe ConGISATA's use in training a group of users, which we believe is the more common scenario. The framework can be easily adapted to train individuals, however we do not discuss that in the paper. The process of computing the passive score begins with a calibration period, during which each user's initial passive ISA score is obtained, without any prior training. After this period, the mean and standard deviation of the entire user group are calculated for each criterion in the taxonomy. During training, a new z-score (standard score) is computed daily for each user for each criterion, using the mean and standard deviation derived in the calibration period. The new z-scores are then averaged for each of the taxonomy's focus areas and are subsequently averaged again to obtain a final passive score for each user. Since the z-score is not meaningful to users, the cumulative probability function of the normal distribution is used to transform the z-score to a 0–100 scale.

Computing the Overall Score: The overall ISA score is the average of the active and passive scores.

4.2 Gamification

To increase user engagement and optimize the effectiveness of training, we have incorporated essential gamification elements into our framework. Table 2 lists some of these key elements, along with their rationale, and explains how they have been implemented in ConGISATA.

Table 2. ConGISATA’s gamification elements

Element	Explanation
Continuous Learning	Dunlosky <i>et al.</i> [21] provided a comprehensive review of study techniques and assessed their effectiveness. One of the techniques covered is continuous learning, which was termed <i>distributed practice</i> and defined as “implementing a schedule of practice that spreads out study activities over time”. Based on prior research, distributed practice was one of just two techniques to be rated by the authors as having high utility. It was assessed that distributed practice “works across students of different ages, with a wide variety of materials, on the majority of standard laboratory measures, and over long delays”. Focusing on the cybersecurity domain, the findings of Kumaraguru <i>et al.</i> [7] align with those of Dunlosky <i>et al.</i> , demonstrating the benefits of extended security training over condensed single sessions. Based on these findings, we designed our game as a continuous learning process, unlike the common approach found in the literature of a single-session game
Considers PRT	Following the research presented in Sect. 2, we implemented a penalty mechanism to discourage PRT, whereby users that fail to take preventive measures will face penalties, resulting in point deductions. This approach transforms PRT into active risk-taking, where users are held accountable for their inaction shortly after it occurs, regardless of whether or not any damage was incurred. For instance, if our sensors detect that certain users have not installed anti-malware software, they will have points deducted, even if no malware has exploited this vulnerability on their devices. Furthermore, users will continue to face daily penalties until they address and fix the issue by installing anti-malware software, further discouraging avoidance behavior
Levels/Progression	It is crucial to provide players with a clear indication that they are acquiring knowledge and advancing through the training process. We achieve this through a ranking system comprised of two elements: points and levels. Players earn points (reflected in their ISA score) by exhibiting good security practices, and as they accumulate more points, they move up to higher levels. Our framework assigns users to one of three levels based on their ISA score: “beginner”, “intermediate”, and “pro”. These levels do not change the difficulty of training and are only used to give the users the feeling that they are advancing
Competition	Competition is a fundamental aspect of nearly every game, in contexts including security. Healthy competition can significantly enhance engagement and enjoyment among players and encourage individuals to surpass their previous performance. Our game incorporates competition through (1) a leaderboard that ranks players by points, providing insight into their standing relative to others; and (2) the points and levels mentioned above, promoting competition among players
Adaptive Gamification Through Personalized Feedback/Guidance	Immediate personalized feedback is important to prevent player confusion and maintain their engagement in the game. Further guidance helps players progress and improve as the game continues. Immediate feedback in our game is in the form of the learning screen. Each event that causes points to be earned or deducted, such as a sensor discovering poor application handling behavior, is presented on the learning screen on the day on which the event occurred. Additional guidance is possible through a dedicated article on the event’s topic. In addition, each user’s scores appear on their learning screen, highlighting the areas requiring improvement
Conciseness	The game’s exercises should be brief and not take much of the players’ time. In our game, the feedback is succinct and highlights the topics pertinent to each player. This approach reduces the time commitment for players and avoids redundant review of familiar material

5 Evaluation

To evaluate the proposed framework, we performed a long-term experiment involving 70 undergraduate and graduate students who use their smartphones regularly. The subjects' ages ranged from 21 to 31, with a mean age of 25 and a median age of 26. The experiment involved the collection of sensitive personal information from subjects for a long period of time, including their browsing patterns. We took measures to preserve the subjects' privacy and reduce any privacy risks associated with participating in the experiment. The experiment was approved by the Institutional Review Board (IRB), provided that: (1) The subjects participated in the experiment, freely, at their own will. The subjects received course credit in exchange for their participation. The subjects were fully aware of the type of data that would be collected and were allowed to withdraw from the study at any time. (2) The data was encrypted before being transmitted between the server and the mobile app. (3) The server was within the university domain, with restricted access and organizational defenses. (4) When possible, the sensitive data itself was not transmitted to the server (such as SMS contents), only the meta-data was (such as the number of SMS messages containing URLs). During the experiment, we measured the subjects' behavior while operating their smartphones and exposed them to three types of SE attacks in 15 attack simulations. We then compared each subject's initial and final ISA scores, measuring the improvement achieved. We also examined how the participants' performance in responding to the challenges evolved during the training process. This section provides a detailed description of the evaluation process and results.

5.1 Mobile Sensors

To evaluate the passive aspects of subjects' behavior, we implemented multiple sensors using Android APIs and used them to assess various criteria from the taxonomy of Bitton *et al.* We did not assess all of the criteria for reasons of simplicity and privacy. The criteria and the way they were assessed are presented in Table 3. In some cases, we found that the corresponding sensor did not work well for a large number of subjects or the sensor had no influence on the score; for example, for criterion OS2, we found that all of our subjects did not root their device before or during the experiment. In such cases, we omitted these sensors and the criteria that correspond to them, and they are not included in our analysis of the results. In addition, 10 out of the 70 subjects either had a technical problem with their smartphone which prevented them from participating, did not use the app, or decided to withdraw from the study. These subjects were omitted from the results analysis as well. Finally, while a higher z-score usually indicates better performance, some of the criteria represent bad behaviors (such as criterion AI1). In such cases, indicated in Table 3 by having "(lower is better)" in their means of assessment, we multiplied their z-score by -1 , changing positive numbers into negative progression indicators.

Table 3. List of criteria assessed for the experiment

Criterion	Means of assessment
AI1: Downloads apps from trusted sources	An app was considered trusted if it was downloaded from an official app store (such as Google Play). The score for this criterion is the number of untrusted apps found on the subject's device (lower is better)
AI2: Does not install apps that require dangerous permissions	The score for this criterion is the number of apps on the subject's device which require dangerous permissions, as classified by Android (lower is better)
AI3: Does not install apps with a low rating	We considered a low rating to be less than three and a half stars (out of five) in the Google Play store. The score for this criterion is the number of apps with a low rating found on the subject's device (lower is better)
AH1: Regularly updates apps	Google Play features the last date on which an app was updated. The score for this criterion is the number of apps that are not up-to-date found on the subject's device (lower is better)
AH3: Properly manages running/installed apps	An app is considered unused if the subject did not use the app for more than two weeks. The score for this criterion is the number of unused apps found on the subject's device (lower is better)
B1: Does not enter malicious domains	A domain is considered malicious if Google's safebrowsing API has classified it as such. The score for this criterion is the number of malicious domains the subject has entered in the last seven days (lower is better)
VC1: Does not open messages received from unknown senders	We monitored two message inboxes for each subject - SMS and Gmail's spam inbox. An SMS is considered to be from an unknown sender if the sender of the SMS is not in the subject's contact list. The SMS score is the percentage of how many unknown SMSs the subject has opened in the last seven days. Likewise, the Gmail score is the percentage of emails classified as spam by Gmail that were opened in the last 30 days. The final score for this criterion is the average of the SMS and Gmail scores (lower is better)
VC2: Does not click on links received from unknown senders	We considered an event to be of the 'clicking on links received from unknown senders' type if the following three conditions were met: (1) the subject opened a message from an unknown sender, as defined in VC1, (2) the message that was opened contained a URL, and (3) we also identified a transition between the SMS/Gmail apps and the browser app (Google Chrome), suggesting the subject has clicked on that URL. The score for this criterion is the number of times a subject has clicked on URLs from unknown senders in the last seven days (lower is better)
A2: Uses two-factor authentication mechanisms	A subject was considered to be using two-factor mechanisms if either a two-factor authentication app or an SMS (from the last seven days) indicating two-factor use was found on their device. The score for this criterion is one if the subject uses two-factor mechanisms and otherwise zero (higher is better)
A3: Uses password management services	The subject was considered to be using password management services if a password-managing app was found on their device. The score for this criterion is one if the subject uses password management services and otherwise zero (higher is better)
OS2: Does not root or jailbreak the device	We used a dedicated Android package (rootBeer) that implements various heuristics to determine whether or not a device is rooted. The score for this criterion is one if the subject has not rooted the device and otherwise zero (higher is better)
SS2: Uses anti-virus application regularly to scan the device	The score for this criterion is one if the subject has an anti-virus app installed on the device and otherwise zero (higher is better)
SS5: Uses PIN code, pattern, or fingerprint	A device was considered secured if a lock-screen was enabled. The score for this criterion is one if the subject's device is secured and otherwise zero (higher is better)
N1: Does not connect to unencrypted networks	A network was considered encrypted if a security protocol was enabled (such as WPS, WPA2). The score for this criterion is the number of unencrypted networks the subject has connected to in the last seven days (lower is better)
N3: Uses VPN services on public networks	The subject was considered to be using VPN services if a VPN app was found on their device. The score for this criterion is one if the subject uses VPN services and otherwise zero (higher is better)
PC1: Disables connectivity when not in use	The score for this criterion is the number of times in the last seven days that a connectivity channel (i.e., Bluetooth, Wi-Fi, NFC) was enabled for more than five minutes, without being connected (lower is better)

5.2 Social Engineering Challenges

To evaluate ConGISATA's influence on behavior in active risk situations, we implemented three types of challenges: phishing, impersonation, and permission attacks. The challenges were presented weekly, with one challenge of each type per week, resulting in three challenges every week and a total of 15 challenges. The order of the challenges presented during the week, as well as the day and hour in which they were presented, was randomized. Examples of the challenges are provided in Fig. 2. The challenges were designed as follows.

Phishing: Phishing is the most prevalent SE attack vector. In our experiment, this attack involved creating a web page that emulates a login page from a pre-designed template, typically for student services. The attack was initiated by emailing the subjects and enticing them to click on an attached link to authenticate themselves for a supposed university-related event. The link directed them to one of three domains that we purchased for the experiment, which resemble the actual university domain. The email was sent by a familiar sender, like 'student administration,' who is known to the subjects as a legitimate email source for university administration. The email was sent during the academic semester when administrative emails from the university are expected. Although the phishing email appears genuine, there are several indications that it was a phishing attack. First, it was not sent from the university's mail system; second, the link provided was not associated with the university's domain; and third, the phishing web page did not employ the HTTPS protocol. To safeguard the subjects' privacy, authentication information was not transmitted to the server. In this challenge, we evaluated the subject twice. First, we determined whether they clicked on the link and accessed the website. If they did, we then determined if they entered login details. The following phishing templates were used:

(1) *Facebook security:* An email was sent, informing subjects that they violated Facebook's code of conduct and their profile was at risk of deletion. Subjects were urged to log in to their account and appeal, via a link provided in the email.

(2) *Moodle - new grade:* Moodle is a learning platform that the university uses to upload course materials and students use to submit assignments. An email was sent to subjects telling them a grade was assigned to them on the Moodle platform, providing a link to log in and view it.

(3) *Organizational password change:* Students are required to change their organizational password periodically. An email was sent to students asking them to change their password via the link provided or their account would be locked.

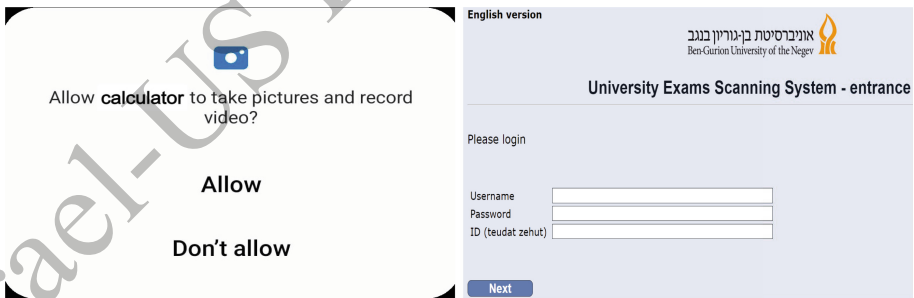
(4) *New appeal response:* In the subjects' university, students can make an appeal about the way in which their test was reviewed and graded. During the exam period, an email was sent to subjects informing them of a response to an appeal they made regarding a specific exam, followed by a link to the appeal system.

(5) *New exam scan*: In the exam period, an email was sent to subjects telling them that an exam they took had been graded and the results were published. A link to the exam website was provided, enabling the subject to see the grade.

Permission Attack: Malicious applications can trick unaware subjects into granting dangerous permissions during runtime. In each variant of this challenge, the device requested the granting of a dangerous permission to an app that does not need that permission. The mobile agent triggered the attack scenario when the subject used the phone and appeared on the screen using the Android permission requests' UI. The subject could reject or approve the request; a subject who granted privileges to the app was considered vulnerable to the attack.

The experiment included the following permission request templates: The Calculator requests camera permissions, WhatsApp requests calendar permissions, Camera requests SMS permissions, and Gmail requests SMS permissions.

Impersonation: Fraudulent apps can deceive people in order to gain possession of their credentials. In this challenge, we simulated a malicious application that sends a push notification while impersonating a legitimate service. The user interface of the notification exhibited a characteristic indicative of a phishing attack, which is the appearance of our mobile agent's name, along with the impersonated app name. Upon clicking the notification, our application launched, presenting a replica of the login screen of a well-known and trusted app. To assess the subjects' performance in this attack, we classified them into two categories: half-vulnerable if they clicked the notification but did not complete the login process, and fully vulnerable if they both clicked the notification and completed the login process. To ensure the privacy of the subjects, the authentication information was not transmitted to the server. The experiment included an app impersonating Facebook, Instagram, and the university's official app.



(a) Permission attack

(b) Phishing

(c) Impersonation

Fig. 2. Illustration of the different challenges

5.3 Articles and Blog Posts

Prior to the experiment, we searched the web for publicly available relevant educational articles and blog posts. We looked for two types of items: items about each focus area in the ISA taxonomy (meaning only about passive aspects) for the ConGISATA group and items about each of the three types of SE challenges (meaning only about active aspects) for the baseline group. After a thorough review, we found 32 items (16 per group) and labeled them by topic. Additionally, for the baseline group, each item was manually assigned a comprehensiveness grade, reflecting its depth and complexity on a scale from one (denoting basic and intuitive content) to five (indicating comprehensive and technical material). This grade determined the order in which the items were provided to subjects in the baseline group, as described in Sect. 5.4. In the ConGISATA group, the order of the items was predetermined and fixed for the entire training process. There was one item about each focus area in the ISA taxonomy. The list of articles and blog posts is presented in Table 5 in the appendix.

5.4 Experiment Setup

Each subject was assigned randomly to one of two groups, ConGISATA and baseline, each of which initially had 35 subjects. All subjects were asked to install our mobile app on their smartphones and keep it for the next five weeks. As mentioned in Sect. 4, we first needed to calculate an initial score for each subject in a calibration period. All subsequent scores in the training process were relative to this initial score and used for personalization and later analysis. For both groups, the calibration period consisted of the first week of the experiment. During this period, the mobile sensors monitored the subjects' behavior, and they were presented with three challenges (one of each type). Afterward, the sensor monitoring and three weekly challenges continued until the end of the experiment. In addition, as mentioned in Sect. 4.1, to compute the active score, we used a moving window of the last X challenges. We set X to be five for both groups. The training process began at the end of the calibration period and continued for four weeks. Each group was trained using one of two different methods; a comparison of the groups' training processes is provided in Table 4.

5.5 Results

In this study we address the following three research questions:

RQ1: Can our framework improve users’ passive ISA score, as measured by the mobile ISA taxonomy? If so, how does it compare to the baseline method? First, we analyzed the passive score deltas and examined each criterion individually. Figure 6 (in the appendix) shows the delta in the score for each of the criteria as a function of the number of days since the experiment started. An increase

in the score was observed for all but one criterion. Furthermore, our framework resulted in a more notable improvement in the group’s performance relative to that of the baseline group. We also examined the total passive ISA score for each group, calculated as the average across the focus areas of the various criteria. As seen in Fig 3, the use of our framework improved the passive ISA score, whereas no improvement was observed for the baseline group.

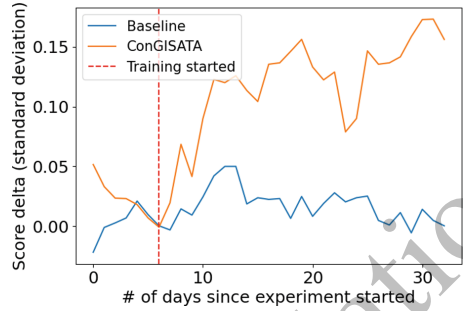


Fig. 3. Average passive score deltas per group over time

Table 4. Comparison of the groups’ training processes

Group	Subject of Articles	# Articles	Gamification	Personalization	Timing
ConGISATA	Passive risk related	16	✓	The collection of articles is fixed. Low scores or negative score deltas direct subjects to articles related to focus areas that need improvement.	All articles were available from the second week.
Baseline	Active risk related	8 are chosen personally, from a pool of 16	✓	Articles are selected based on the subject’s performance in challenges. Their comprehensiveness increases with repeated failures in the same attack vector.	Starting from the second week, articles were incrementally provided twice a week and remained accessible until the experiment’s conclusion, with notifications sent to subjects’ devices.

RQ2: Does ConGISATA help users improve their active ISA score, as measured using the challenges? If so, how does it compare to the baseline method? We analyzed the active score over time. Similar to other ISA training methods, the baseline method uses articles related to active risk situations thereby emphasizing active aspects. Thus, we anticipate that the active score of the baseline group will improve over time. Figure 4 shows the change in score throughout the experiment. Initially, both groups experienced a decrease in their scores for two reasons: Firstly, during the first week (to the left of the red dotted line), the groups received no training. Secondly, the initial active score was calculated after day 13 (indicated by the green dotted line), after a sufficient number of challenges were presented – a minimum of five challenges with at least one challenge from each one of the three attack vectors (see Sect. 5.4). After day 13 both groups demonstrated notable improvement, with the ConGISATA group exhibiting slightly better performance. This result emphasizes that the training for secure passive behavior received by the ConGISATA group also reinforces active behavior.

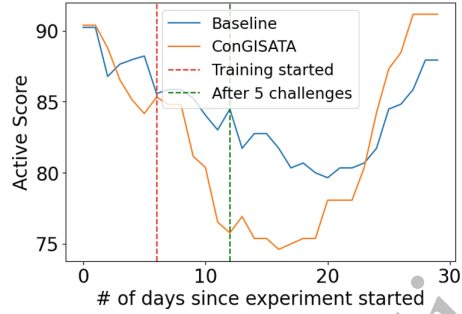


Fig. 4. Active score over time

RQ3: Does increased use of our framework correlate with greater improvement in passive behavior?

We logged every view of each of the app’s screens and looked for a correlation between views and behavioral change. As expected, the most significant Pearson correlation was found between the number of views of the learning screen and the total delta in the passive score ($r = 0.72$, $p = 3.41e-5$), as seen in Fig. 5. A similar result was obtained when checking for a correlation between the number of days in which a subject viewed the learning screen and the passive score delta. However, one of our learning screen’s main advantages is its continuous nature, allowing users to see up-to-date details on each focus area with respect to their passive behavior. Going through the entire screen thoroughly may require more than one view per day so we chose to report the number of views and not the number of days, to differentiate subjects who viewed the learning screen multiple times a day from those who did not.

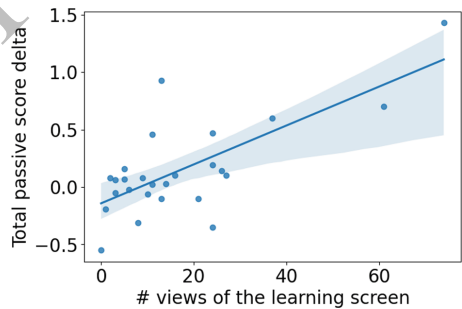


Fig. 5. Correlation between the number of views of the learning screen and passive score delta

6 Conclusion

This study introduces ConGISATA - a continuous gamified ISA training and assessment framework that collects data from various sensors in users' everyday environments to examine aspects of ISA in real-life settings. The sensor readings are integrated into the framework, which generates a feedback loop. This continuous feedback mechanism helps users learn from their mistakes and improve their resilience against prevalent security risks. The use of sensors and challenges also provides a more reliable ISA assessment than the commonly used self-reported questionnaires. Our results confirm that ConGISATA improves passive behavior, while the baseline method does not. Moreover, although ConGISATA only provides articles on passive behavior, it helps users improve their ability to handle active attack scenarios. ConGISATA can be used in a corporate environment, in new employee training or as a regularly performed periodic procedure. Adapting the framework to new threats should be relatively easy, and may include these steps: (1) adding a new type of challenge simulating the new threat; (2) implementing additional sensors to measure related real-life behaviors; and (3) collecting (or creating) educational articles about the new threat. The number of subjects in this study does not allow meaningful analysis of the contribution of timing and personalization to ConGISATA's ability to improve ISA. This limitation can be addressed in more extensive experiments, including an ablation study performed with a large group of users, which we plan for future work.

Appendix

List of Articles and Blog Posts

As described in Sect. 5.3, we collected 32 publicly available relevant educational articles and blog posts to use in the experiment (the blog posts and articles are listed in Table 5). The items for the ConGISATA group are listed first, with their corresponding ISA taxonomy criterion ID, and do not include a comprehensiveness grade. The items for the baseline group, which include a comprehensiveness grade, are listed after the bold horizontal line.

Table 5. The articles and blog posts used in the experiment

	Topic	Links	Comprehensiveness Grade
ConGISATA	Account (A2)	link	–
	Account (A3)	link	–
	Browser (B1)	link	–
	Virtual Communication (VC1)	link	–
	Virtual Communication (VC2)	link	–
	Network (N1)	link	–
	Network (N3)	link	–
	Application Installation (AI1)	link	–
	Application Installation (AI2)	link	–
	Application Installation (AI3)	link	–
	Application Handling (AH1)	link	–
	Application Handling (AH3)	link	–
	Security Systems (SS2)	link	–
	Security Systems (SS5)	link	–
	Physical Connectivity (PC1)	link	–
Operating System (OS2)	link	–	
Baseline	Impersonation Attacks	link	2
	Impersonation Attacks	link , link , link , link	3
	Impersonation Attacks	link	5
	Permission Attacks	link , link	2
	Permission Attacks	link , link	3
	Permission Attacks	link	5
	Phishing Attacks	link , link , link , link , link	1

Passive Score Delta by Criterion

Figure 6 shows the average score deltas for the groups per criterion, as a function of the number of days since the experiment started.

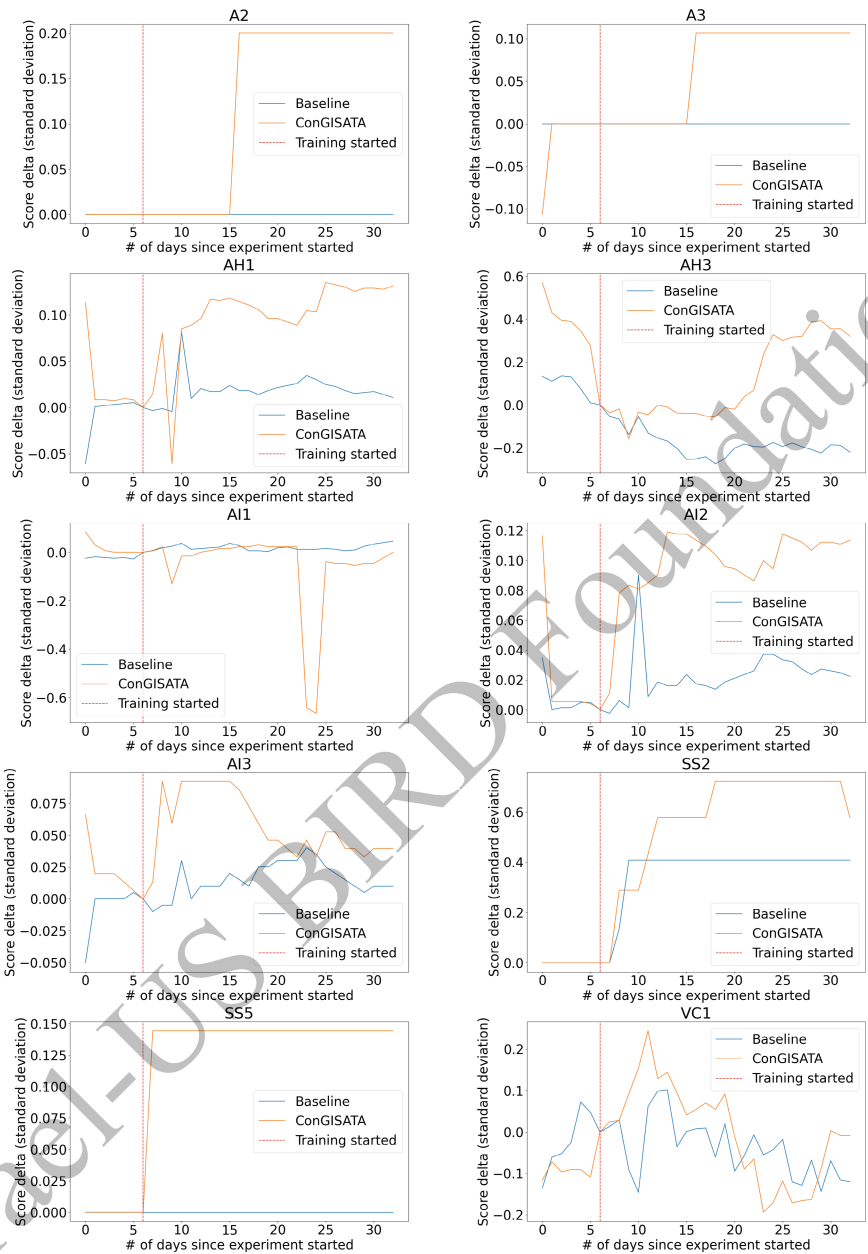


Fig. 6. Average score deltas for the groups per criterion, as a function of the number of days since the experiment started.

References

1. Kumar, A., Chaudhary, M., Kumar, N.: Social engineering threats and awareness: a survey. *Eur. J. Adv. Eng. Technol.* **2**, 15–19 (2015)

2. Kelly, R.: Almost 90% of cyber attacks are caused by human error or behavior. ChiefExecutive. Net (2017)
3. Bada, M., Sasse, A., Nurse, J.: Cyber security awareness campaigns: why do they fail to change behaviour? arXiv Preprint [arXiv:1901.02672](https://arxiv.org/abs/1901.02672) (2019)
4. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining “gamification”. In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp. 9–15 (2011)
5. Hamari, J., Koivisto, J., Sarsa, H.: Does gamification work?—a literature review of empirical studies on gamification. In: 2014 47th Hawaii International Conference on System Sciences, pp. 3025–3034 (2014)
6. Gjertsen, E., Gjære, E., Bartnes, M., Flores, W.: Gamification of information security awareness and training. In: ICISSP, pp. 59–70 (2017)
7. Kumaraguru, P., et al.: School of phish: a real-world evaluation of anti-phishing training. In: Proceedings of the 5th Symposium on Usable Privacy and Security, pp. 1–12 (2009)
8. Bitton, R., Finkelshtein, A., Sidi, L., Puzis, R., Rokach, L., Shabtai, A.: Taxonomy of mobile users’ security awareness. *Comput. Secur.* **73**, 266–293 (2018)
9. Keinan, R., Bereby-Meyer, Y.: “Leaving it to chance”—passive risk taking in everyday life. *Judgment Decis. Making* **7** (2012)
10. Keinan, R., Bereby-Meyer, Y.: Perceptions of active versus passive risks, and the effect of personal responsibility. *Pers. Soc. Psychol. Bull.* **43**, 999–1007 (2017)
11. Bitton, R., Boymgold, K., Puzis, R., Shabtai, A.: Evaluating the information security awareness of smartphone users. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2020)
12. Newbould, M., Furnell, S.: Playing safe: a prototype game for raising awareness of social engineering. In: Australian Information Security Management Conference, p. 4 (2009)
13. Hart, S., Margheri, A., Paci, F., Sassone, V.: Riskio: a serious game for cyber security awareness and education. *Comput. Secur.* 101827 (2020)
14. Chapman, P., Burket, J., Brumley, D.: PicoCTF: a game-based computer security competition for high school students. In: 2014 USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 2014) (2014)
15. Denning, T., Lerner, A., Shostack, A., Kohno, T.: Control-Alt-Hack: the design and evaluation of a card game for computer security awareness and education. In: Proceedings of the 2013 ACM SIGSAC Conference On Computer & Communications Security, pp. 915–928 (2013)
16. Alqahtani, H., Kavakli-Thorne, M.: Design and evaluation of an augmented reality game for cybersecurity awareness (CybAR). *Information* **11**, 121 (2020)
17. Luh, R., Temper, M., Tjoa, S., Schrittwieser, S., Janicke, H.: PenQuest: a gamified attacker/defender meta model for cyber security assessment and education. *J. Comput. Virol. Hacking Tech.* **16**, 19–61 (2020)
18. Yasin, A., Liu, L., Li, T., Fatima, R., Jianmin, W.: Improving software security awareness using a serious game. *IET Softw.* **13**, 159–169 (2018)
19. Arend, I., Shabtai, A., Idan, T., Keinan, R., Bereby-Meyer, Y.: Passive-and not active-risk tendencies predict cyber security behavior. *Comput. Secur.* 101929 (2020)
20. Selvam, V.: Human error in IT security. arXiv Preprint [arXiv:2005.04163](https://arxiv.org/abs/2005.04163) (2020)
21. Dunlosky, J., Rawson, K., Marsh, E., Nathan, M., Willingham, D.: Improving students’ learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* **14**, 4–58 (2013)

22. Canham, M., Posey, C., Constantino, M.: Phish derby: shoring the human shield through gamified phishing attacks. *Front. Educ.* **6**, 536 (2022)
23. Jaffray, A., Finn, C., Nurse, J.: SherLOCKED: a detective-themed serious game for cyber security education. In: *International Symposium on Human Aspects of Information Security and Assurance*, pp. 35–45 (2021)
24. Sophos Sophos 2023 Threat Report (2022). <https://assets.sophos.com/X24WTUEQ/at/b5n9ntjqmbkb8fg5rn25g4fc/sophos-2023-threat-report.pdf>
25. Redmiles, E., Zhu, Z., Kross, S., Kuchhal, D., Dumitras, T., Mazurek, M.: Asking for a friend: evaluating response biases in security user studies. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1238–1255 (2018)
26. Solomon, A., et al.: Contextual security awareness: a context-based approach for assessing the security awareness of users. *Knowl.-Based Syst.* **246**, 108709 (2022)
27. Böckle, M., Novak, J., Bick, M.: Towards adaptive gamification: a synthesis of current developments (2017)
28. Alahmari, S., Renaud, K., Omoronyia, I.: Moving beyond cyber security awareness and training to engendering security knowledge sharing. *Inf. Syst. E-Bus. Manag.* 1–36 (2022)
29. Dincelli, E., Chengalur-Smith, I.: Choose your own training adventure: designing a gamified SETA artefact for improving information security and privacy through interactive storytelling. *Eur. J. Inf. Syst.* **29**, 669–687 (2020)
30. Scholefield, S., Shepherd, L.A.: Gamification techniques for raising cyber security awareness. In: Moallem, A. (ed.) *HCHI 2019*. LNCS, vol. 11594, pp. 191–203. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22351-9_13
31. Omar, N., Foozy, C., Hamid, I., Hafit, H., Arbain, A., Shamala, P.: Malware awareness tool for internet safety using gamification techniques. In: *Journal of Physics: Conference Series*, vol. 1874, p. 012023 (2021)
32. Wu, T., Tien, K., Hsu, W., Wen, F.: Assessing the effects of gamification on enhancing information security awareness knowledge. *Appl. Sci.* **11**, 9266 (2021)
33. Heid, K., Heider, J., Qasempour, K.: Raising security awareness on mobile systems through gamification. In: *Proceedings of the European Interdisciplinary Cybersecurity Conference*, pp. 1–6 (2020)