

NOTICE: This is a paper that was accepted for publication in the **IEEE Transactions on Visualization and Computer Graphics**. It is not the final published version. The DOI of the definitive version available in IEEE Explore is

<https://doi.org/10.1109/TVCG.2020.3022340>

# Learning Safety through Public Serious Games: A Study of “Prepare for Impact” on a Very Large, International Sample of Players

Luca Chittaro and Fabio Buttussi

**Abstract**—Recent years have witnessed a growing interest in serious games (SGs), i.e. digital games for education and training. However, although the potential scalability of SGs to large player populations is often praised in the literature, available SG evaluations did not provide evidence of it because they did not study learning on large, varied, international samples in naturalistic conditions. This paper considers a SG that educates players about aircraft cabin safety. It presents the first study of learning in a SG intervention conducted in naturalistic conditions with a very large, worldwide sample, which includes 45,000 players who accepted to answer a knowledge questionnaire before and after playing the game, and more than 400,000 players whose in-game behavior was analyzed. Results show that the SG led to improvement in players’ knowledge, assessed with different metrics. Moreover, analysis of repeated play shows that participants improved their in-game safety behavior over time. We also focus on the role of making errors in the game, showing how they led to improvement in knowledge. Finally, we highlight the theoretical models, such as error-based learning and Protection Motivation Theory, that oriented the game design, and can be reused to create SGs for other domains.

**Index Terms**—Serious games, training, education, user study, research-in-the-large, aviation safety



## 1 INTRODUCTION

RECENT years have witnessed a growing interest in the use of digital games for education and training. The term “serious game” (SG) has become mainstream in the literature and tends to be used interchangeably with “games for learning” [1]. Such increase in interest is supported by the gradual build-up of empirical evidence about the general effectiveness of the SG approach to learning in a diversity of domains, as highlighted by different meta-analyses [1], [2], [3].

### 1.1 Advancing SG Studies through a Research-in-the-Large Methodology

Connolly et al. [3] analyzed 129 papers in regard to the potential positive impacts of computer games and SGs not only on learning and skill enhancement, but also on engagement. They found frequently occurring outcomes and impacts concerning knowledge acquisition/content understanding, affective outcomes, and motivational outcomes. Four years later, Boyle et al. [1] re-examined the literature and found 143 additional papers that confirmed the previous conclusions with higher quality evidence.

Clark et al. [2] examined 69 studies, concluding that game conditions significantly enhanced learning relative to nongame conditions, and effects varied across game design characteristics such as game mechanics, visual and narrative features. These findings suggest that SGs can be a novel, attractive option for public campaigns and interventions that could educate through video games distributed on-line. This approach could be especially interesting for topics, like appropriate safety behaviors, towards which public attention and knowledge is often scarce. As an example, different authors have shown how aircraft passengers’ knowledge about safety behaviors on-board is alarmingly low, and the safety education methods used today (safety cards, videos, and live briefings) are ineffective [4], [5], [6]. Reasons for such ineffectiveness include inability to engage the learner [4], lack of comprehension [5], and difficulty in recalling the information even immediately after attending to it [6].

SG interventions could not only help in increasing attention and engagement, but also improve learner’s knowledge and judgement concerning emergencies, in aviation [7] as well as in other safety domains of public concern such as road safety [8] or mass emergency preparedness [9]. Moreover, the same approach can be exploited also in vocational training concerning emergencies, for example trauma triage [10] or fire fighting [11].

However, although the potential scalability of SGs to large player populations is often praised in the literature [10], [12], available SG evaluations did not provide evidence of it because they did not test learning on large, varied samples in naturalistic conditions. Indeed, the

• The authors are with the Human-Computer Interaction Lab, Department of Mathematics, Computer Science, and Physics, University of Udine, Udine 33100, Italy. E-mail: {luca.chittaro, fabio.buttussi}@uniud.it.

hundreds of SG studies in the literature typically focus on lab evaluations or small local interventions, and concern modestly sized samples. Typical sample size in the studies analyzed by [3] included only tens or at best hundreds of participants, with only five studies reaching a thousand participants (the largest sample had 1817 participants), and no studies considering an international worldwide sample. Similarly, the studies analyzed in [1] involved only tens or at best hundreds of participants, with only six studies reaching a thousand participants (the largest sample had 1666 participants) and no studies considering an international worldwide sample. The analysis of a smaller number of studies provided by [2] is consistent with the previous ones.

To the best of our knowledge, the present paper describes the first study of learning in a SG intervention conducted in naturalistic conditions with a very large sample (more than 400,000 participants) and at a worldwide level. We also aim at showing how on-line deployment of SGs can be an effective strategy that allows an intervention to obtain a large-scale assessment of education needs in the public as well as actual learning on a practically relevant topic.

To conduct the study, we followed the research-in-the-large methodology, which is based on embedding a research apparatus into a mobile app and making it publicly available on app stores, such as Apple's App Store or Google Play, to attract a potentially large number of users. Mobile apps can collect data from users and devices, and send it to researchers through the Internet. In this way, researchers can obtain data for statistical analysis, run studies with heterogeneous samples of participants, and observe behavior in naturally occurring user contexts [13], [14]. Although proposed only recently, research-in-the-large has already been used for several studies in different areas such as mindfulness [15], cognitive science [16], and happiness studies [17].

Moreover, by following a research-in-the-large approach, there are no differences between the way players retrieve and play the studied game in their naturalistic contexts and the way they do it with any other game on their devices. This increases ecological validity with respect to lab studies, in which participants are observed by researchers after meeting them in person and the overall context can be very different from those in which people usually play video games.

## 1.2 SGs as Tools for Error-Based Learning

The field of SGs is wide and diverse, leading to numerous, open and unexplored issues that need to be studied. In particular, Mayer [18] pointed out that, as the field progresses, focus should shift from broad doctrines to specific learning mechanisms linked to research evidence. An additional goal of our study is to explore the role of making errors as a learning mechanism in a SG.

Designing a learning experience in a way that includes making errors and even encourages learners to make them is central to some non-SG learning approaches such as error training [19], error management training [20] and more generally exploratory active learning that includes

trial-and-error [21]. In these approaches, errors are seen as serving an informative function for the learner, pinpointing where his/her knowledge needs improvement, and prompting refinement of his/her mental models.

However, error-based approaches in education were avoided for decades, because prominent scholars depicted errors as having a negative effect on learning. Two emblematic examples are given by Skinner [22] who equated errors with punishment that can inhibit behavior but does not contribute to learning, and Bandura [23] who viewed errors as detrimental to learning and promoted a guided, error-free learning approach. These views fear that an exploratory learning strategy, with incorrect paths and errors entertained by the learner, would make learning of the correct procedures more difficult. They recommend instead error-free approaches with step-by-step guidance that should result in flawless behavior from the beginning, and a feedback that focuses only on positive social reinforcement of the correct execution of tasks [21].

Research about error-based approaches emerged more recently, showing that not only an error-based approach is effective, but it can actually be more effective than error-free approaches, provided that a crucial condition (corrective feedback) is met. The meta-analysis by Keith & Frese [20] contrasted error management training (EMT) with proceduralized training methods that have a negative attitude toward errors and seek to minimize them through step-by-step instructions on correct task execution. The meta-analysis considered 24 EMT studies, most of which in the area of software skills, and the results were highly favorable to EMT. The growing body of literature about learning based on errors has been surveyed and analyzed by Metcalfe [21], who provides further examples of controlled experimental studies, showing that, in comparison with error-free learning, the making of errors results in better memory for the correct response, as long as the error is followed by corrective feedback. The analyzed studies also shed light about proper corrective feedback: to be effective, feedback must not simply tell learners whether they were right or wrong, but also suggest the correct answer [21].

Thanks to evidence about effectiveness that removed the original stigma from error-based approaches, one would expect to see them more easily embraced in the design of learning interventions. However, a consideration that still makes some designers hesitant is a possible learner's negative attitude towards errors, because making errors can have negative emotional effects. Learners can find errors frustrating, interrupting, time-consuming, and possibly resulting in negative judgement of their behavior. Resort to error-free approaches might contribute to reinforce fear of errors in the learner. SGs have been specifically advocated as a solution to overcome these limitations. As highlighted by Plass, Homer, and Kinzer [12], one of the potential advantages of SGs over traditional education and training methods is "graceful failure", that is the possibility of designing failure as an expected and even necessary step in the SG learning experience rather than an undesirable outcome. This should encourage learner's exploration, and make it easier for de-

signers to follow the error-based learning approach [21].

However, the meta-analysis by Keith and Freese [20] and the survey by Metcalfe [21] did not analyze SGs or game-like approaches. The study conducted by Ivancic and Hesketh [19] concerns instead a car simulator, which is closer to the SG context considered in this paper. The results indicated that, compared with error-free learning where participants drove through a training run not designed to elicit errors, error-based training led to significantly better transfer to driving tests that were analogous to the situations encountered in training, and more effective strategies for coping with novel driving situation.

A SG can facilitate error-based training by stimulating players to make errors in an engaging, simulated environment. Showing the negative effects of the error on the learner's avatar in the game can make the error even more salient and at the same time help players in understanding the cause-effect relationships between actions and their effects. The current paper investigates whether the positive effect of errors that are followed by audiovisual simulation of their consequences as well as corrective feedback, extends to the context of SGs.

## 2 METHODS

### 2.1 The "Prepare for Impact" Game

"Prepare for Impact" [24] is a SG created for an educational intervention to improve aviation safety knowledge in the general public, and developed using the Unity 5.3 game engine. The SG is available for mobile devices running Android or iOS, and can be freely downloaded from Google Play and Apple App Store (see [24] for download links and a video trailer).

#### 2.1.1 Design Principles

The game is inspired to the survival genre: its levels allow players to virtually experience different, life-threatening aircraft accidents, with the goal of surviving the situation by choosing a course of actions that properly handles the different threats. The narrative of the different levels of "Prepare for Impact" is based on real accident reports. As the theory of narrative engagement points out, going through a story helps people achieve active mastery of decision principles that can be recalled when related situations arise [25].

In general, most video games can be seen as an operant conditioning mechanism [26]. In "Prepare for Impact", correct player's responses are reinforced through positive feedback, by receiving the ability to proceed further towards a positive outcome in the game narrative as well as receiving badges when a level is successfully completed. On the contrary, behavior that needs to be discouraged receives aversive feedback, through seeing the negative consequences of errors on the player's avatar or other characters, getting stuck at a given point in the game or receiving verbal criticism from other characters. The followed approach includes desirable features [26], [27] such as enabling players to explore hypothetical situations, instantly observe the link between cause and effect, providing players with immediate feedback, and showing them

the consequences of the chosen (right or wrong) behavior in vivid ways. Following error-based approaches to learning, the game does not simply provide negative feedback after an error, but adds corrective feedback explaining the error and suggesting how to prevent it.

Since seeing negative consequences of accidents in vivid ways could evoke fear in the player, we also followed Protection Motivation Theory [28], [29], a leading theoretical model of how individuals respond to fear-evoking information about risk. The theory recommends that a risk should be presented to the message recipient in ways that highlight both its severity and the recipient's vulnerability to the risk. This will likely evoke fear that could motivate the recipient to pay attention. However, for learning to be effective, a recommendation about how to avoid the depicted risk must follow, describing an action that is effective in averting the risk and that the individual is capable of.

A preliminary, small-size version of the game contained just one level with a narrative inspired by the real accident occurred to US Airways flight 1549 [30]. A lab evaluation on a small sample of players [31] contrasted learning with the game (experienced through a virtual reality headset) vs. learning with traditional printed safety cards. Results indicated that the SG was superior to the traditional educational material: knowledge retention after one week from exposure was significantly larger in those who tried the game rather than the traditional materials. Moreover, subjective as well as physiological measurements showed that the SG was more engaging than the safety card, a factor that can contribute to explain the obtained superior retention. These results encouraged us to extend the game with several accident scenarios and to port it to mobile devices in a way that could make it easily accessible to the general public.

#### 2.1.2 Knowledge Questionnaire and Game Levels

On the first run, the game informs players that it will collect data for research purposes, and then invites players to fill in a questionnaire to assess their knowledge about aircraft emergency procedures for passengers (hereinafter, *pre-test*) before trying the game levels. Players are free to accept or decline the invitation. If they decline, the game will never make further invitations to fill the questionnaire. The questionnaire consists of 12 questions, each of them with 4 to 6 possible answers, of which one or more are correct (Table A1 of the online supplemental materials shows all questions and answers, highlighting the correct ones). The questionnaire covers different cabin safety topics, concerning right and wrong passenger's behaviors during turbulence, cabin decompression, emergency landing, and ditching. The questions deal with all the required passenger's actions in aircraft emergencies, from fastening seat belts and assuming the brace position to using life vests and reaching the exits. They also test players' knowledge about situations in which fire, water, and smoke in the cabin make the evacuation more complex. The high number of possible answers and the multi-choice nature of the questions makes the questionnaire difficult to answer in a fully correct way without a

strong knowledge of passenger safety procedures, as also confirmed by feedback from two aviation safety experts. We used a difficult questionnaire to prevent ceiling effects: an easy questionnaire might have made it impossible to measure possible improvements by comparing results before and after playing the game.

After players have declined or completed the questionnaire, the game displays the main menu, from which they can start the tutorial and the different levels of the SG. Initially, only the tutorial can be selected. Players must complete the tutorial to unlock the possibility to play the first level, and then must complete a level to unlock the next one. In the following, we will use the term *session* to indicate each run of a game level or the tutorial. In the tutorial, players learn game controls by boarding an aircraft as passengers. They have to move in the cabin to reach their seat row, put their luggage in the overhead bin, reach the assigned seat, sit down, and fasten seat belts. The tutorial makes players familiarize with all game controls, i.e., controls to move in the 3D environment (virtual joystick) and controls to perform actions (buttons). At any instant of each level, the SG displays 0 to 3 buttons, one for each (correct or wrong) action that players can perform in that instant. The game levels allow players to experience different emergency scenarios on commercial aircrafts. The first and third levels are set on a Boeing 777-300 [32], a widely used twin-aisle aircraft, while the second and fourth levels are set on an Airbus 320 [33], a widely used single-aisle aircraft. The first game level (hereinafter, *L1*) is partially inspired by the accident occurred at Toronto International Airport [34], in which the aircraft crashed into a field after overshooting the runway in inclement weather. The second level (*L2*) is instead inspired by the well-known accident that occurred to US Airways flight 1549 [30], which struck a flock of large birds a few minutes after take-off, lost thrust in both engines, and was forced to ditch on a river. The third level (*L3*) concerns a ground collision between two aircraft on a misty day. The fourth level (*L4*) reproduces a cabin decompression, followed by a crash landing on a field at night. In the different game levels, the player is exposed to additional threats that make evacuation more complex: fire, water, debris, smoke, which can also make some exits unusable. Fig. 1 illustrates the type of computer graphics employed in all the levels.

During each game level, if players choose correct actions, they progress in the evacuation of the aircraft. On the contrary, if they choose wrong actions or omit right ones, the SG first provides negative feedback and then a

recommendation about proper behavior. More precisely, if the errors are irreversible in the real world (e.g., jumping from wings instead of using the wing slides), the SG vividly shows the negative consequences of the errors (e.g., the passenger getting hurt), then interrupts the level, displays a brief textual recommendation about proper behavior, and finally restarts the level from the instant in which players took the wrong decision. On the contrary, if the errors are reversible in the real world (e.g., taking luggage during the evacuation), then the SG uses characters (passengers or flight attendants) to address the user with verbal feedback and recommendations (e.g., “leave your luggage, you’re slowing down everyone”). However, if players ignore such recommendations and persist in the error (e.g., keeping the luggage), then the game treats the error in the same way as irreversible errors.

Players can quit a game level at any time after invoking a pause menu. If players reach the end of the level, the SG assigns them a score, which is calculated as the sum of the time it took to complete the level and a penalty time for each error made during the session. In this way, higher time scores indicate worse performance, and players must evacuate the aircraft quickly and without errors to improve their score. The SG also assigns a bronze, silver, or gold badge when players complete game levels making only reversible errors, making only one reversible error, or making no errors at all, respectively.

When players who had filled the pre-test complete *L4* for the first time, the SG invites them to fill in the same questionnaire again (hereinafter, *post-test*). This is done to assess their knowledge about emergency procedures after trying the game levels. Players can accept or decline to fill in the post-test. Regardless of players’ choice, the SG will not invite them to fill the post-test anymore, and they can continue playing the game.

## 2.2 Hypotheses and Experimental Design

The study focuses on the validity of SGs as tools to educate the general public, and is based on a worldwide intervention that involved a very large sample of players in naturalistic conditions. To evaluate whether the SG could lead to improvements in players’ knowledge, we (i) used the knowledge questionnaire described in Section 2.1.2, and (ii) studied participant’s behavior while playing the game (hereinafter, *in-game behavior*).

The study aimed at answering the following research questions:

- RQ1: Does playing the SG lead to improvement in knowledge about the taught topics?



Fig. 1. Examples of graphics from the SG: a) passengers assuming the brace position during runway overrun (*L1*), b) water flooding the aircraft through a door that must not be opened (*L2*), c) passengers escaping from fire after ground collision (*L3*), d) passengers wearing oxygen masks after cabin decompression (*L4*).

- RQ2: Does learning show up also in a reduction of wrong in-game behavior? Is this possible reduction larger when the game levels are played more than once?
- RQ3: Does making an error in the game lead to an improvement in player's safety knowledge related to that error?

Both RQ1 and RQ2 address the research goal of evaluating improvement in players' knowledge, but using different measuring instruments, respectively a knowledge questionnaire (RQ1) and in-game behavior (RQ2). RQ2 also extends consideration to the possible effects of repetitive play of the game levels. RQ3 is meant to shed more light about how playing SGs might contribute to improve knowledge, focusing on the role of making errors in the game. This investigation of errors can be particularly interesting for the reasons discussed in Section 1.2.

Regarding RQ1, we hypothesized that the SG could lead to an improvement in the knowledge questionnaire as suggested by previous studies about the learning effects of SGs [1], [2], [3]. However, unlike previous studies, our paper evaluates the hypothesis on a very large, international sample of players in naturalistic settings.

Regarding RQ2, we hypothesized that playing a game level could lead to a decrease of wrong in-game behaviors on a subsequent session of the same level, because the SG is designed to make players vividly experience the negative consequences of their errors on their avatar (or on other characters) and at the same time provide players with the appropriate recommendation to avoid those errors in the future, so players are likely to learn how to behave correctly and avoid that type of error when playing the level again. For the same reasons, we hypothesized that playing a level more than once could reinforce learning, leading to further reduction of wrong in-game behaviors. This outcome should not be considered obvious in video games. In traditional education based on repetitive testing with written tests, the learner is motivated to make less errors in order to not have to repeat the test. On the contrary, a video game player can intentionally repeat errors just because (s)he was impressed by the audiovisual representation of the error consequences and wants to experience them again or see them in a different place of the game world. Moreover, (s)he could also intentionally look for new errors to make when playing again because (s)he wants to see what surprising things could happen as a result of making those new errors in the game. This sort of player's behaviors might not significantly decrease the number of errors in a second or third game session even if the player has actually learned what is the right way to complete the level without errors.

Finally, regarding RQ3, we hypothesized that making in-game errors could lead to an improvement in related knowledge questions between pre-test and post-test for the reasons mentioned above, and because the SG can allow players to restart from the instant in which they made the error and correct it. This could help players to quickly revise their mental models, focusing especially on the faulty parts that need change or refinement [21]. For example, if players open a door that is under water level,

they see their avatar drown as a result of the action, and also receive corrective feedback in the form of a textual recommendation. This should enable them to quickly update the faulty part of their mental model of aircraft doors in relation to when they can be opened. This change should show up not only in future in-game behavior but also in the knowledge questionnaire.

The design of the study was within-subjects when we assessed improvement in players' knowledge (RQ1 and RQ2), while it was between-subjects when we considered the effects of making an error and we thus compared players who made the specific in-game error vs. players who did not make it (RQ3).

Ethical committee approval to carry out the study was not required because the study concerned a publicly available educational service. Terms of use displayed by the SG informed players that data about app usage was recorded and was going to be used only in anonymous form for scientific purposes. Since we reasoned that some players might perceive the SG as scary, we informed the app stores about the presence of some fearful/violent content. As a consequence, they assigned a 12+ rating to the SG. The rating was displayed on the store game page.

### 2.3 Data Collection

The SG collects data about the employed mobile devices, the pre-test and post-test answers, the played sessions (tutorial and game levels), and the errors made by players during the sessions. Collected data are temporarily stored in the device and sent to a secure server when a network connection is available. The database on the server organizes data in four types of entries:

- *Device profile* contains (i) a unique device identifier (UDI) generated by the SG, (ii) type of platform (Android or iOS) and version of the operating system, (iii) device model, CPU, GPU, RAM, and graphics API, (iv) width (in pixels), height (in pixels), and dpi of the screen, (v) language and time zone settings at first usage.
- *Knowledge test* contains (i) the UDI, (ii) a flag to distinguish pre-test and post-test, (iii) the times at which players started and stopped filling the questionnaire, (iv) the number of questions completed, (v) the number of correctly answered questions, (vi) a flag that indicates if the player checked or not an answer to a question, for each of the 62 answers contained in the questionnaire.
- *Session* contains (i) the UDI, (ii) the game level played (tutorial or level number), (iii) the times at which players started and stopped the session, (iv) the score and the badge (bronze, silver, or gold; no badge; session not completed). The combination of time at which a player started the session and the UDI is used as unique session identifier.
- *Error* contains (i) the UDI, (ii) the time at which players started the session in which the error was made, (iii) the time at which players made the error, (iv) the type of error (all the 22 in-game error types are listed in Table A2 of the online supplemental materials).

Data collection started on March 17th, 2016, when we made the SG freely available on the on-line stores. On May 15th, 2018, we made a full copy of the database for the analysis purposes of this study. The analyzed database contained 3,514,947 device profiles; 3,800,856 pre-tests, of which 2,194,490 completed; 919,110 post-tests, of which 531,996 completed; 38,115,359 sessions, and 86,670,827 in-game errors. The larger number of pre-tests with respect to device profiles is explained by the fact that some users re-installed the app or cleaned app data. The next section explains how this was managed.

## 2.4 Data Cleaning

Before analyzing the data, we performed a data cleaning process to remove data coming from players who completed the pre-test or post-test more than once. The issue of users repeating tests multiple times is well-known in Web-based studies [35], [36]. As described in Section 2.1.2, our SG offers the possibility to complete the tests only at specific times, preventing players from filling them again. However, some players uninstall the game after playing it (for example, to free space on the device for other games) and later reinstall it because they want to play it again. Moreover, users can force a clean of the game data (an advanced option offered by the operating system of the mobile device for any app). In both cases, the game will lose its data on the device, and will thus propose the test again when re-started. In addition, network errors can sometimes lead to the reception of incomplete data. For example, it may happen that the database receives the data about an in-game error, but not the session the error belonged to. If network errors occur, the SG tries to send the data again later, but in a few cases, resending of data is impossible, for example when players uninstall the SG or clean its data, leaving the data incomplete. Keeping track of the UDI made it possible to detect the devices from which multiple pre-tests or post-tests were submitted. We discarded from the analysis all data coming from those devices as well as all data coming from devices that sent incomplete data. After this cleaning process, the database contained 3,238,302 device profiles; 3,059,775 pre-tests, of which 1,790,568 were completed; 724,788 post-tests, of which 397,016 were completed; 26,427,481 sessions, and 65,454,296 in-game errors.

We calculated the completion time of each completed test as the difference in seconds between the time at which players started the questionnaire and the time at which they submitted the last answer. This highlighted that some players took a very long or a very short time to fill the questionnaire. Malhotra [37] explains long answer times in on-line surveys in terms of interruptions during filling out of the questionnaire, while he considers extremely quick completion times as suggestive of insufficient respondent's consideration that could introduce lower quality data in the dataset. In our specific case, players who took very long times might have paused because they received a notification or a phone call, or temporarily abandoned the mobile device to perform other tasks, while players who completed the test very quickly (e.g., by selecting answers randomly) might have been

eager to start playing the game. In any case, tests filled in very long or very short completion times cannot be considered as reliable, because they are indicative of respondents who did not pay proper attention to the test. To determine the range of acceptable completion times, we took into consideration literature about reading times, and also visually analyzed the distribution of completion times, as described in detail in Section A3 of the online supplemental materials. After excluding pre-tests that were completed too quickly or too slowly, the pre-tests of 320,908 players remained. Of these players, 45,164 devoted an acceptable time also to complete the post-test and were thus considered for the analysis of possible improvement in the knowledge questionnaire.

## 2.5 Measures

To assess possible improvement in the knowledge questionnaire (RQ1), we considered the following measures:

- *Question score.* For each question in the knowledge questionnaire, we computed a question score for each participant. Question score is 0 if the respondent checked at least one wrong answer. Otherwise, question score is equal to the number of correct answers checked by the respondent divided by the number of correct answers for the question (for example, if the participant checked two correct answers and there were three correct answers, question score is 0.67). Therefore, question score ranges between 0 and 1. Comparison of this measure between pre-test and post-test for each of the 12 questions provides information about the topics for which playing the SG can lead to an improvement.
- *Overall knowledge questionnaire score.* This measure is the sum of the 12 question scores. Therefore, it can range between 0 and 12. Comparing this measure between pre-test and post-test assesses possible overall improvement in players' knowledge.

To assess possible effects of playing on the reduction of wrong behaviors in the game (RQ2), we considered the following measures of in-game errors made by players (for brevity, we indicate *in-game error type* as IGET):

- *Occurrence of IGET.* For each game level and each IGET that can be made in that level, this measure indicates if players made that IGET in that level (0=IGET not made, 1=IGET made). The average of this measure is equivalent to the percentage of participants who made the IGET in the game level. For example, a 0.25 average means that 25% of participants made the IGET in the game level. We compared this measure between the first and the second session of each game level to assess whether playing the level once could lead to a decrease in the occurrence of the IGET in a subsequent session of that level. Moreover, we compared the measure between the first, the second, and the third session of each game level to assess whether playing the game twice could lead to a further decrease in the measure in a subsequent session. It is

important to note that the number of participants in the second comparison is smaller than the first comparison, because participants who completed two sessions of the same game level are many more than those who completed three sessions of the same game level (see Table 1 and Section 2.6).

- *Total number of occurred IGETs.* For each game level, this measure is the sum of occurrence of IGET, considering all IGETs that can be made in that level. We measured it also for the four game levels combined by summing the value of the measure of each level. The measure can range from 0 to 11 for L1; from 0 to 13 for L2; from 0 to 10 for L3; from 0 to 13 for L4; from 0 to 47 for the four levels combined, since some IGETs can be made in more than one level. Comparing this measure between the first and the second session as well as between the first, the second, and the third session assesses its possible decrease (indicating an overall improvement in players' knowledge) or increase (indicating deterioration) over time.

To assess whether making an error in the game leads to an improvement in the knowledge related to that error (RQ3), we considered the questions related to IGETs in the questionnaire, and computed the following measure:

- *Pre-post difference in IGET-related question score.* It is the difference between post-test question score and pre-test question score in a question related to an IGET. A positive value of this measure indicates knowledge improvement on the topic covered by the question. For each question and each IGET related to the question, we compared the measure between the group of players who made that IGET at least once in the game before answering the post-test and the group who never made that IGET before the post-test. For each IGET, Table A2 of the online supplemental materials indicates which questions of the knowledge questionnaire are related to the IGET. If making an IGET leads to improvement in a related question, the measure for that question should be higher in the group of players who made the IGET rather than the group who did not make the IGET.

## 2.6 Participants

After the cleaning process (see Section 2.4), collected data concerned 3,238,302 players. The most frequent languages set on their devices were English (33.9%), Spanish (12.9%), Russian (11.7%), Portuguese (7.3%), and Indonesian (5.1%). Other 37 languages followed, with percentages of players descending from 3.3% (French) to values close to zero (Table A3 of the online supplemental materials provides all the details). This is a first indication of an international, worldwide sample. We also observed the distribution of players' time zones as set on their devices at the first run of the SG (Fig. A2 of the online supplemental materials provides all the details). For most time zones, the percentage of players in a time zone (number of players in the time zone / 3,238,302) was close to the percentage of Internet users in that time zone

(number of Internet users in the time zone / total number of Internet users in the world), providing additional evidence of the international, worldwide nature of the sample. If we focus on the 320,908 participants who completed the pre-test in an acceptable time, the percentage of devices with language set to English predictably increases (73.5%), because the knowledge questionnaire was in English. Nevertheless, this sample still includes 41 different languages (Table A4 of the online supplemental materials provides all the details), and the distribution of time zones still indicates an international, worldwide nature of the sample (Fig. A3 of the online supplemental materials provides all the details). The sample of participants who completed both pre-test and post-test in an acceptable time (45,164 players) shows similar distributions and includes 39 different languages.

To answer the research questions that concerned the knowledge questionnaire (RQ1, RQ3), we considered the players who completed the pre-test as well as the post-test questionnaires within an acceptable time range, and the resulting sample size was 45,164. To answer the research question (RQ2) that concerned in-game behavior, for each game level, we considered the players who played it more than once: sample size varied with the game level as well as the number of times players completed the level, and ranged from 112,752 to 425,021. Ta-

TABLE 1  
MEASURES, COMPARISONS, STATISTICAL TESTS, AND PARTICIPANTS FOR EACH RESEARCH QUESTION.

RQ	Measure	Comparison	Statistical test	Participants
RQ1	Overall knowledge questionnaire score	Comparison of the measure between pre-test and post-test	Wilcoxon signed-ranks	45,164
	Question score	For each of the 12 questions, comparison of the measure between pre-test and post-test	Wilcoxon signed-ranks	45,164
RQ2	Total number of occurred IGETs	For all four game levels combined and for each game level individually, comparison of the measure between the first and the second session	Wilcoxon signed-ranks	L1: 418,860 L2: 425,021 L3: 275,327 L4: 249,178 All levels: 108,855
		For all four game levels combined and for each game level individually, comparison of the measure between the first, the second, and the third session	Friedman (post-hoc: Wilcoxon signed-ranks with Bonferroni correction)	L1: 171,187 L2: 213,283 L3: 119,459 L4: 112,752 All levels: 37,783
	Occurrence of IGET	For each IGET and game level, comparison of the measure between the first and the second session	McNemar's	L1: 418,860 L2: 425,021 L3: 275,327 L4: 249,178
		For each IGET and game level, comparison of the measure between the first, the second, and the third session	Cochran Q (post-hoc: McNemar's with Bonferroni correction)	L1: 171,187 L2: 213,283 L3: 119,459 L4: 112,752
RQ3	Pre-post difference in IGET-related question score	For each question and each IGET related to the question, comparison of the measure between (i) the group of players who made that IGET at least once and (ii) the group of players who never made that IGET	Mann-Whitney	45,164



ble 1 lists in detail the exact number of participants involved in each analysis. To analyze the possible improvement between the first and the second session of a game level (resp. all game levels) in terms of IGETs, the participants were necessarily the players who completed a first and a second session with that level (resp. all game levels). Similarly, in the analyses that considered the first, second, and third session of a game level (resp. all game levels), the participants were the players who completed a first, a second, and a third session with that level (resp. all game levels). We carried out an analysis with the first two sessions and a separate analysis with the first three sessions, because the first analysis could show the possible improvement after a single session on a very large sample (ranging from 249,178 to 425,021 participants), while the second analysis could provide information about possible further improvement after an additional session on the sample of players (ranging from 112,752 to 213,283 participants) who completed three sessions of the same game level. We did not impose any constraint on the time between sessions of the same game level or on the order in which participants played game levels in the second and third sessions. We made this decision because we were interested in observing the overall in-game behavior of players in a naturalistic setting, in which players play the game levels based on their actual time availability and preferences.

## 2.7 Statistical analysis

Statistical analysis was carried out using SPSS 25. Table 1 lists the statistical tests performed in the study (Section A5 of the online supplemental materials provides the details about execution of each analysis). It is worth noting that a statistical test run on a large sample will almost always find a significant difference (unless there is no difference at all), but very small differences, even if statistically significant, can often be meaningless [38]. This is typical of studies conducted on thousands of participants, where p-values go towards zero even when the results have no practical significance [39]. For this reason, effect size analysis has a fundamental role in large-sample studies. As summarized by Sullivan & Feinn [38], both p-value and effect size are essential to understand the impact of the results: statistical significance examines whether the findings are likely to be due to chance, whereas effect size helps to understand the magnitude of the differences found. As explained by Cohen [40], a medium effect size represents an effect likely to be “visible to the naked eye of a careful observer”, a small effect size is “noticeably smaller than medium but not so small as to be trivial”, and a large effect size is “the same distance above medium as small was below it”. On the contrary, if an effect size is below the small threshold, then it should be considered as negligible. Therefore, we report p-values and effect sizes to highlight which of our results have practical significance. To consider a hypothesis confirmed, we do not simply look at p-value significance, but also consider if the effect size is above the threshold for small effect. For Wilcoxon signed-ranks tests and Mann-

Whitney tests, effect size is reported as the absolute value of the  $r$  coefficient (effect thresholds: small 0.1, medium 0.3, large 0.5). For McNemar’s tests, it is reported as the odds ratio (effect thresholds: small 1.22, medium 1.86, large 3.00). For Friedman and Cochran Q tests, we report the effect size of the post-hoc tests, using respectively the absolute value of the  $r$  coefficient of Wilcoxon signed-ranks test and the odds ratio of McNemar’s test with the effect thresholds given above. It is important to remark that all effect sizes above the small threshold are not trivial: in particular, a small effect might not be “visible to the naked eye”, but has a practical significance [40].

## 3 RESULTS

Considering RQ1, Wilcoxon signed-ranks test (two-tailed) showed that the difference in the overall knowledge questionnaire score between pre-test and post-test was statistically significant ( $Z=-158.94$ ,  $p<0.001$ ): the overall knowledge questionnaire score in the post-test ( $M=8.04$ ,  $SD=2.33$ ) was higher than in the pre-test ( $M=5.81$ ,  $SD=2.36$ ). The effect size was large ( $|r|=0.53$ ). Therefore, our hypothesis about the positive effects of the SG on players’ knowledge measured by the questionnaire was confirmed. Considering each specific question, Wilcoxon signed-rank test showed a statistically significant improvement ( $p<0.001$ ) in question scores for all 12 questions except one, for which the difference in question score was negligible and not statistically significant. The effect size for the statistically significant increases in question scores ranged from 0.11 (small) to 0.42 (medium). Therefore, for all but one question, results reached both statistical and practical significance, and our hypothesis about the positive effects of the SG on knowledge was confirmed. Details for each question are provided in Table B1 of the online supplemental materials.

Considering RQ2, Wilcoxon signed-ranks test (two-tailed) showed that the difference in the total number of occurred IGETs between the first and the second session was statistically significant for each game level as well as for the four game levels combined ( $p<0.001$  in all cases), as reported in Table 2. The total number of occurred IGETs in the second session was lower than in the first session for all game levels (individual as well as combined). The effect size was always of practical significance: small for L3; medium for L1, L2, and L4; large for all levels combined (Table 2). Therefore, our hypothesis that learning would show up also in terms of reduction in the number of IGETs was confirmed, because all results were both statistically and practically significant. Analyzing the occurrence of each specific IGET in the first and the second session for each game level, statistical significance was reached for 45 (IGET, game level) pairs ( $p<0.001$  for all) out of 47. Considering those 45 pairs, effect size was negligible for 3 pairs, small for 13 pairs, medium for 14 pairs, and large for the remaining 15 pairs. Among the 42 pairs for which the difference was practically significant, 37 show a decrease in occurrence of IGET as hypothesized, and 5 an increase. Such increase concerned only three IGETs (two of them in L2 only, and

the other in L1, L3, and L4), all with a small odd ratio. In summary, learning after the first session was supported by statistically and practically significant results obtained with 37 (IGET, game level) pairs (15 large effects, 14 medium effects, 8 small effects). Details for each IGET and game level are reported in Table B2 of the online supplemental materials.

Considering the analysis of players who completed also a third session, Friedman test showed that the difference in the total number of occurred IGETs between the first, the second, and the third session was statistically significant for each game level as well as all four game levels combined ( $p < 0.001$  in all cases), as reported in Table 3. The total number of occurred IGETs in the second session was always lower than the first session, and the total number of occurred IGETs in the third session was always lower than the second session. Wilcoxon signed-ranks post-hoc comparisons (i.e., first vs. second, second vs. third, and first vs. third session) were all statistically significant ( $p < 0.001$ ), as shown in Table 4. Effect sizes of the comparisons between the first and the second session were all of practical significance: small size for L2 and L3, medium for L1 and L4, and large for the four game levels combined. The effect sizes of the comparisons between the second and the third session were also all practically significant (small size). The effect sizes of the comparisons between the first and the third session were small for L3, medium for L2 and L4, and large for L1 and for all game levels combined. Since all these results were both statistically and practically significant, our hypothesis about playing the SG more than once was confirmed. More precisely, playing the first session results in a considerable knowledge improvement: occurrence of IGETs in the four levels combined dropped substantially from the first ( $M=12.92$ ,  $SD=6.10$ ) to the second session ( $M=7.54$ ,  $SD=4.37$ ). After playing the second session, an additional, smaller but statistically and practically significant, improvement was observed in the third session for the four levels combined ( $M=6.21$ ,  $SD=3.91$ ), and for each individual level. Considering the occurrence of each specific IGET in the first three sessions for each game level, statistical significance was reached for all 47 (IGET, game level) pairs ( $p < 0.005$  for one pair, and  $p < 0.001$  for all others). Post-hoc tests reached statistical significance for all except four comparisons. Considering the 137 comparisons for which statistical significance was reached, practical significance was achieved for 116 of them (the effect size was large for 33, medium for 30, small for 53). Details for each IGET and game level are reported in Table B3 and B4 of the online supplemental materials.

Finally, we consider the pre-post difference in IGET-related question score for each question and each related IGET, comparing the group of players who made that IGET at least once in the game and the group who never made it (RQ3). We found a statistically significant difference for 18 (question, IGET) pairs out of 20 ( $p < 0.001$  for 17 pairs,  $p < 0.05$  for 1 pair). The effect size was negligible for 10 of the statistically significant differences and was practically significant for the remaining 8 pairs (small size). For all these 8 pairs, the pre-post difference in question

TABLE 2

MEAN (M) AND STANDARD DEVIATION (SD) OF TOTAL NUMBER OF OCCURRED IGETs, WILCOXON SIGNED-RANK TEST STATISTICS (Z), TWO-TAILED SIGNIFICANCE (P), AND EFFECT SIZE (|R|) OF THE COMPARISON BETWEEN FIRST AND SECOND SESSION.

Game level	1st session		2nd session		Z	p	r
	M	SD	M	SD			
All levels	12.88	5.86	7.26	4.46	-243.19	< 0.001	0.52
L1	4.01	2.39	1.78	1.74	-435.31	< 0.001	0.48
L2	3.44	1.75	2.45	1.70	-285.39	< 0.001	0.31
L3	2.06	1.49	1.51	1.42	-160.96	< 0.001	0.22
L4	3.83	2.36	2.04	1.86	-304.51	< 0.001	0.43

TABLE 3

MEAN (M) AND STANDARD DEVIATION (SD) OF TOTAL NUMBER OF OCCURRED IGETs, FRIEDMAN TEST STATISTICS ( $\chi^2$ ), AND SIGNIFICANCE (P) OF THE COMPARISON BETWEEN FIRST, SECOND, AND THIRD SESSION.

Game level	1st session M	1st session SD	2nd session M	2nd session SD	3rd session M	3rd session SD	$\chi^2$	p
All levels	12.92	6.10	7.54	4.37	6.21	3.91	32023.43	< 0.001
L1	3.98	2.47	1.89	1.71	1.50	1.54	118058.21	< 0.001
L2	3.42	1.78	2.53	1.65	2.15	1.66	70644.79	< 0.001
L3	2.11	1.51	1.64	1.38	1.35	1.33	22214.68	< 0.001
L4	3.85	2.29	2.19	1.78	1.61	1.65	73181.95	< 0.001

TABLE 4

WILCOXON SIGNED-RANKS TEST STATISTICS (Z), SIGNIFICANCE (P), AND EFFECT SIZE (|R|) OF POST-HOC TESTS COMPARING TOTAL NUMBER OF OCCURRED IGETs IN FIRST VS. SECOND, SECOND VS. THIRD, AND FIRST VS. THIRD SESSION.

Game level	First vs. second session			Second vs. third session			First vs. third session		
	Z	p	r	Z	p	r	Z	p	r
All levels	-140.26	< 0.001	0.51	-65.90	< 0.001	0.24	-149.90	< 0.001	0.55
L1	-267.36	< 0.001	0.46	-88.10	< 0.001	0.15	-290.82	< 0.001	0.50
L2	-183.39	< 0.001	0.28	-96.64	< 0.001	0.15	-239.45	< 0.001	0.37
L3	-90.90	< 0.001	0.19	-64.94	< 0.001	0.13	-138.75	< 0.001	0.28
L4	-195.14	< 0.001	0.41	-100.05	< 0.001	0.21	-231.06	< 0.001	0.49

score was higher in the group of players who made the IGET, supporting the hypothesis that making in-game errors improves players' knowledge about the topics related to the error. Table B5 of the online supplemental materials reports details for the 20 (question, IGET) pairs.

## 4 DISCUSSION

Results of the study show that playing the serious game led to improvement in players' knowledge about the taught topics, measured with a knowledge questionnaire (RQ1). This confirms the positive results of previous studies about the learning effects of SGs [1], [2], [3], but with an unprecedented large and international sample of more than 45,000 players. Moreover, players used the game in their naturalistic settings and contexts. The way users got to know about, downloaded, installed and played the

game on their own mobile devices was no different from what happens with other free games available in the on-line stores, and is representative of the conditions that public education campaigns and interventions inevitably encounter when they want to exploit mobile games distributed through on-line stores. From this point of view, our evaluation did not simply concern a game in isolation as in lab studies, but a broader, public game-based intervention in the real world. The difference in the score of the knowledge questionnaire (RQ1) was statistically significant and practically significant, with effect size surpassing the large threshold. Average questionnaire score increased of 38% after playing the four game levels. Analysis of each of the 12 questions showed statistical as well as practical significance of the improvement for all questions except one. We thus discuss in more detail that question, which concerned the purpose and behavior of floor lights in the aircraft cabin. Two factors that probably made the SG unable to improve knowledge on this specific question were that the IGET related to the question appeared only in one game level and that, unlike other safety knowledge taught by the game, understanding the behavior of the floor lights in the simulated accident portrayed by such level required subtle consideration of the aircraft environment on the player's side. More salient details like threats (fire, smoke, water,...) or objects on which the player had to act directly (doors, life vest, seat belt,...) likely attracted attention away from less salient background details like the small lights on the floor. Moreover, the game did not highlight a direct cause-effect link between visual features of the floor lights and negative effects on the player's avatar or progression in the game, while improper consideration of threats or wrong direct actions on objects produced such immediate negative feedback. More generally, what happened with this question also highlights the importance of evaluating learning effects of SGs in detail, assessing on which topics learning is obtained or not, and then pinpoint the specific parts of the game that have not led to the expected improvement and need thus to be changed in future releases of the SG. In our case, a change that would likely improve learning on the floor light question would be to create a situation in one of the game levels (or in a new, dedicated level) in which understanding the behavior of the floor lights is essential to survive the level, and the game clearly highlights the cause-effect relation between not using the information from the floor lights and negative effects suffered by the player's avatar.

A limitation of the study is that it did not compare learning with the SG to learning with traditional cabin safety education materials (safety briefings and cards), because of the enormous difficulties of recruiting a traditional materials sample that could match the treatment sample in size and geographical distribution. However, as described in Section 2.1.1, an evaluation in the lab had already compared a group playing a preliminary version of the SG with a group of users who learned instead from traditional safety education materials [31].

The results concerning RQ2 show learning also as a reduction of wrong behaviors (in-game error types) on very

large samples up to 425,021 players. The difference in the total number of occurred IGETs between the first and the second session was statistically and practically significant for each game level as well as for the four game levels combined. Therefore, our hypothesis that learning would show up also in terms of reduction in the number of IGETs was confirmed. The fine-grained analysis of each IGET for each game level showed a contribution to these result by 37 out of 47 (IGET, game level) pairs. It is interesting to note that one IGET (staying too close to the aircraft after evacuation) followed instead an opposite, increasing trend in all three game levels in which it was present. This is a good example of a repetition of an error that could happen in SGs even when players know that they should not perform that behavior. In particular, when players play a level for the first time, they are focused on completing it successfully and they run away from the aircraft until they reach a safe distance and the level is complete. Some of them stay too close to the aircraft to watch the events happening: in this second case, the SG detects the IGET, creates an explosion that hurts the player's avatar, and explains that players should reach the safe area where a crowd of passengers is standing. When playing the level a second time, players are more confident about their capability of completing it successfully and can decide to try to stay at a distance that allows them to watch the details of the simulation (for example, external aircraft damage or fire, crowd of animated passengers coming out and escaping from the aircraft). This can lead them to stay too close to the aircraft, unintentionally triggering the IGET even if they were aware of it.

In general, our findings show that a SG that combines an error-based approach with a rich feedback that includes both a vivid audiovisual portrait of the negative consequences of player's errors and corrective feedback about the right action, promotes learning about correct behavior and error avoidance when playing the level again. It is worth noting that, while some IGETs could be made only in one game level, other IGETs could occur in different levels, and some learning effects could be observed also between a level and the next one. For example, results suggest that the occurrences of IGETs concerning luggage were much higher in L1 than in the subsequent game levels. On the contrary, the IGET concerning fall from wings was smaller in the first session of L1 than L4. This could be explained by the very different context in which the IGET could occur. More precisely, in L1 the accident occurred during the day, the aircraft had wide wings, and a flight attendant near the exit gave instructions, making a jump from the wing less likely, while in L4 the accident occurred during the night, the aircraft wings were smaller, and no flight attendant was present near wings. The analysis of learning over three sessions for each game level confirmed our hypothesis that repetitive play of the SG is beneficial, because it results in larger reductions in wrong behaviors (IGETs). The total number of occurred IGETs in the second session was lower than the first session and the total number of occurred IGETs in the third session was lower than the second session for

the four game levels combined and for each game level individually. All these differences were statistically and practically significant. Overall, a considerable knowledge improvement occurs after playing the first session: occurrence of IGETs in the four levels combined dropped considerably from the first session to the second session. After playing the second session, there is a smaller, but statistically and practically significant, improvement for each level and for the four levels combined. Considering the fine-grained analysis of each specific IGET, post-hoc tests reached both statistical significance and practical significance for 116 of the 141 comparisons.

The results concerning RQ3 support the hypothesis that making in-game errors leads to an improvement in the knowledge related to those errors. For 18 out of 20 (IGET, game level) pairs, the pre-post difference in IGET-related question score was statistically significant. For eight pairs, the difference reached also practical significance (effect size above the small threshold), and the group of players who made the error improved more than the group who did not make the error. Of the remaining 10 pairs that reached statistical significance, 8 were consistent with our hypothesis (with effect sizes between 0.02 and 0.06), and only 2 were not (with effect sizes  $\leq 0.02$ ). The only two pairs that produced inconsistent results concerned the IGET related to the question for which no improvement was found, and the IGET that increased in second sessions as we discussed above. The IGETs in the two pairs that did not reach statistical significance were the IGET related to seat belts, and the IGET related to brace position. For both IGETs, the virtual crew gave clear, verbal commands that probably affected players regardless of making the two IGETs. Overall, the results concerning 16 out of 20 pairs provide new evidence about the effectiveness of making errors for learning, in a context and with an educational tool that is different from those studied in the literature on learning from errors [19], [20], [21]. In the specific context, our results support the validity of exploiting "graceful failure" in SGs [12], i.e. designing failure in the game as an expected and even necessary step in the learning process rather than an undesirable outcome. This makes SGs attractive as a new digital tool in error-based learning approaches, to allow players to make errors in an engaging, simulated environment. The SG design we considered also allows players to restart from the instant at which they made the error and correct it, making it easier for players to revise their mental models as soon as the error occurs.

## 5 CONCLUSIONS

This paper presented the results of a naturalistic study of a multi-level, mobile SG used in a public intervention by a very large, international sample of players. Results showed the effectiveness of the SG in educating the general public, even about topics about which people are not usually motivated to seek information or find it difficult to learn using traditional educational materials. Moreover, the paper has isolated and highlighted design principles and theoretical models (error-based learning, Protec-

tion Motivation Theory, operant conditioning) that have been used in creating the game, and due to their generality can be reused to create similar SGs for other contexts and domains.

A possible limitation of "Prepare for Impact", and any SG that is going to follow Protection Motivation Theory, is that the game could be perceived as too scary by some potential users. For this reason, we have started experimenting with different game designs with the aim of teaching the same knowledge without appealing to fear. In particular, we designed an SG based on arcade game elements, including humorous situations with cartoon characters, which was tested on a small sample of users both in the lab and in a naturalistic setting [41]. The SG was found to be engaging as well as capable to increase knowledge about correct and wrong behaviors in aircraft emergencies. We also designed a new, publicly available app called "Air Safety World" [42], which gamifies lessons with a virtual instructor. We included rewards for successfully completing the lessons: users receive virtual coins that allow them to unlock mini games as well as acquire aircraft and customizations for a virtual airport they can manage in the app. All these approaches are complementary: a public intervention can release different kinds of SGs that appeal to different types of players in such a way that any user can choose the game genre that maximizes his/her engagement in the educational experience.

Finally, we are broadening our research focus to investigate the design and evaluation approach of this paper in other safety domains. In particular, we are focusing on the topic of helping disabled persons in the emergency evacuation of buildings and public places.

## ACKNOWLEDGMENT

Our research was partially supported by a grant of the Federal Aviation Administration (FAA). Nicola Zangrando (HCI Lab, University of Udine) carried out 3D modeling activities for the development of the game.

## REFERENCES

- [1] E. A. Boyle et al., "An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games," *Comput. Educ.*, vol. 94, pp. 178–192, 2016.
- [2] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth, "Digital games, design, and learning a systematic review and meta-analysis," *Rev. Educ. Res.*, vol. 86, pp. 79–122, 2016.
- [3] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Comput. Educ.*, vol. 59, no. 2, pp. 661–686, 2012.
- [4] C. L. Corbett and G. A. McLean, "Passenger safety awareness Reprise 2007: Still ignorant after all these years," in *Procs. Fifth Triennial Intl. Aircraft Fire and Cabin Safety Research Conf.*, 2007.
- [5] C. L. Corbett, G. A. McLean, and D. K. Cosper, "Effective Presentation Media for Passenger Safety I: Comprehension of Briefing Card Pictorials and Pictograms. Final Report DOT/FAA/AM-08/20," Washington, DC, USA, 2008.
- [6] D. Seneviratne and B. R. C. Molesworth, "Employing humour and celebrities to manipulate passengers' attention to pre-flight safety briefing videos in commercial aviation," *Saf. Sci.*, vol. 75, pp. 130–135, 2015.

- [7] L. Chittaro, "Designing Serious Games for Safety Education : 'Learn to Brace' versus Traditional Pictorials for Aircraft Passengers," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 5, pp. 1527–1539, 2016.
- [8] Q. Li and R. Tay, "Improving drivers' knowledge of road rules using digital games," *Accid. Anal. Prev.*, vol. 65, pp. 8–10, 2014.
- [9] L. Chittaro and R. Sioni, "Serious games for emergency preparedness: Evaluation of an interactive vs. a non-interactive simulation of a terror attack," *Comput. Human Behav.*, vol. 50, pp. 508–519, 2015.
- [10] D. Mohan et al., "Serious games may improve physician heuristics in trauma triage," *Proc. Natl. Acad. Sci. United States Am.*, vol. 115, no. 37, pp. 9204–9209, 2018.
- [11] F. M. Williams-Bell, B. Kapralos, A. Hogue, B. M. Murphy, and E. J. Weckman, "Using Serious Games and Virtual Simulation for Training in the Fire Service: A Review," *Fire Technol.*, vol. 51, no. 3, pp. 553–584, 2015.
- [12] J. L. Plass, B. D. Homer, and C. K. Kinzer, "Foundations of Game-Based Learning," *Educ. Psychol.*, vol. 50, no. 4, pp. 258–283, 2015.
- [13] G. Miller, "The smartphone psychology manifesto," *Perspect. Psychol. Sci.*, vol. 7, pp. 221–237, 2012.
- [14] N. Henze and M. Pielot, "App stores: External validity for mobile HCI," *Interactions*, vol. 20, pp. 33–38, 2013.
- [15] L. Chittaro, A. Vianello, "Evaluation of a mobile mindfulness app distributed through on-line stores: a 4-week study," *Int. J. Hum. Comput. Stud.*, vol. 86, pp. 63–80, 2016.
- [16] S. Dufau et al., "Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science," *PLoS One*, vol. 6, 2011, Art. no. e24974.
- [17] M. A. Killingsworth and D. T. Gilbert, "A wandering mind is an unhappy mind," *Science*, vol. 330, p. 932, 2010.
- [18] R. E. Mayer, "On the Need for Research Evidence to Guide the Design of Computer Games for Learning," *Educ. Psychol.*, vol. 50, pp. 349–353, 2015.
- [19] K. Ivancic and B. Hesketh, "Learning from errors in a driving simulation: effects on driving skill and self-confidence," *Ergonomics*, vol. 43, pp. 1966–1984, 2000.
- [20] N. Keith and M. Frese, "Effectiveness of error management training: a meta-analysis," *J. Appl. Psychol.*, vol. 93, pp. 59–69, 2008.
- [21] J. Metcalfe, "Learning from Errors," *Annu. Rev. Psychol.*, vol. 68, pp. 465–489, 2017.
- [22] B. F. Skinner, *Science and Human Behavior*. New York, NY, USA: Macmillan, 1953.
- [23] A. Bandura, *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1986.
- [24] HCI Lab - University of Udine, "Prepare for Impact website," 2019. [Online]. Available: <http://hclab.uniud.it/impact/>. [Accessed: 27-Jul-2020].
- [25] M. Miller-Day and M. L. Hecht, "Narrative means to preventative ends: A narrative engagement framework for designing prevention interventions," *Health Commun.*, vol. 28, pp. 657–670, 2013.
- [26] B. J. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do*. San Francisco, CA, USA: Morgan Kaufmann, 2003.
- [27] A. C. Graesser, P. Chipman, F. Leeming, and S. Biedenback, "Deep learning and emotion in serious games," in *Serious games: Mechanisms and effects*, U. Ritterfeld, M. Cody, and P. Vorderer, Eds. Mahwah, NJ, USA: Routledge, 2009, pp. 81–100.
- [28] R. W. Rogers, "Cognitive and physiological processes in fear appeals and attitude change: A revised theory of Protection Motivation," in *Social Psychophysiology: A sourcebook*, J. T. Cacioppo and R. E. Petty, Eds. New York, NY, USA: Guilford Press, 1983, pp. 153–176.
- [29] D. L. Floyd, S. Prentice-Dunn, and R. W. Rogers, "A meta-analysis of research on Protection Motivation Theory," *J. Appl. Soc. Psychol.*, vol. 30, pp. 407–429, 2000.
- [30] National Transportation Safety Board, "Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River, US Airways Flight 1549, Airbus A320-214, N106US, Weehawken, New Jersey, January 15, 2009. Aircraft Accident Report NTSB/AAR-10/03," Washington, DC, USA, 2010.
- [31] L. Chittaro and F. Buttussi, "Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 4, pp. 529–538, 2015.
- [32] Boeing, "Boeing 777," 2019. [Online]. Available: <http://www.boeing.com/commercial/777/>. [Accessed: 27-Jul-2020].
- [33] Airbus, "Airbus A320," 2019. [Online]. Available: <https://www.airbus.com/aircraft/passenger-aircraft/a320-family.html>. [Accessed: 27-Jul-2020].
- [34] Transportation Safety Board of Canada, "Aviation Investigation Report A05H0002," 2005.
- [35] A. L. Bartell and J. H. Spyridakis, "Managing risk in internet-based survey research," in *2012 IEEE International Professional Communication Conference*, 2012, pp. 1–6.
- [36] S. D. Gosling, S. Vazire, S. Srivastava, and O. P. John, "Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires," *Am. Psychol.*, vol. 59, no. 2, pp. 93–104, 2004.
- [37] N. Malhotra, "Completion Time and Response Order Effects in Web Surveys," *Public Opin. Q.*, vol. 72, no. 5, pp. 914–934, 2008.
- [38] G. Sullivan and R. Feinn, "Using Effect Size—or Why the P Value Is Not Enough," *J. Grad. Med. Educ.*, vol. 4, no. 3, pp. 279–282, 2012.
- [39] M. Lin, H. C. Lucas, and G. Shmueli, "Too Big to Fail: Large Samples and the p-Value Problem," *Inf. Syst. Res.*, vol. 24, no. 4, pp. 906–917, 2013.
- [40] Cohen J, "A Power Primer," *Psychol. Bull.*, vol. 112, no. 1, pp. 155–159, 1992.
- [41] L. Chittaro and F. Buttussi, "Exploring the use of arcade game elements for attitude change: Two studies in the aviation safety domain," *Int. J. Hum. Comput. Stud.*, vol. 127, pp. 112–123, 2019.
- [42] HCI Lab - University of Udine, "Air Safety World website," 2019. [Online]. Available: <http://hclab.uniud.it/asw/>. [Accessed: 27-Jul-2020].



**Luca Chittaro** is full professor of Human Computer Interaction (HCI) in the Department of Mathematics, Computer Science, and Physics of the University of Udine, Italy, where he heads the HCI Lab (<http://hclab.uniud.it>). He has authored or co-authored more than 200 international academic publications, and he is an ACM Distinguished Speaker. His major research interests are in virtual reality, serious games, persuasive technology, mobile HCI, and their applications in health and safety. He has received research grants from a wide range of organizations, including the US Federal Aviation Administration (FAA), the European Union (EU), the Italian Ministry of University and Research (MIUR), and companies such as the Benetton Group and the Intesa Sanpaolo Bank group.



**Fabio Buttussi** received the PhD degree in computer science from the University of Udine. He is an assistant professor in the Department of Mathematics, Computer Science, and Physics of the University of Udine, Italy. His research interests are in virtual reality, HCI, serious games, and their applications in health and safety.



## APPENDIX A – ADDITIONAL MATERIALS




### A.1 In-game Questionnaire

Table A1 shows the 12 questions of the in-game questionnaire as well as their possible answers. Correct answers are highlighted.

### A.2 Error Types

Table A2 lists all the names of the 22 in-game error types (IGETs) and the game levels in which they can be made. The table also reports the questions of the in-game questionnaire related to the IGETs. Related wrong answers in those questions describe behaviors that lead to the IGET. For example, if players “run upright as fast as possible outside the smoke” (Answer 7d), they end up standing in toxic smoke and inhale it (i.e., they make the StandInsideSmoke IGET). On the contrary, related correct answers describe behaviors that prevent the IGET. For example, if players “leave everything on the plane” (Answer 8e), they do not take their luggage in the evacuation (i.e., they prevent the TakeLuggage IGET).

**TABLE A1**  
QUESTIONS AND MULTIPLE-CHOICE ANSWERS IN THE KNOWLEDGE QUESTIONNAIRE. CORRECT ANSWERS ARE HIGHLIGHTED IN GREEN.

Tick ALL the boxes that you think are correct (there can be ONE or MORE correct answers)	
<b>1. When do you have to wear the seat belts?</b>	
a.	When the plane is stationary at the terminal
b.	When the plane is moving from the terminal to the runway
c.	When the plane is landing
d.	If a possible turbulence is announced
e.	If you see dark clouds outside the window
<b>2. During an emergency landing, which of these positions are correct to prepare for impact?</b>	
a.	
b.	
c.	
d.	
<b>3. When do you have to inflate the life vest?</b>	
a.	Before impact
b.	Before leaving your seat
c.	While you are running in the aisle towards the exit
d.	When you are going through the emergency exit
e.	After you find yourself in water
<b>4. In case of water landing, where can you find your life vest?</b>	
a.	In the bin over your seat
b.	In the seat pocket in front of you, together with the safety card
c.	Flight attendants will give it to you during the water landing
d.	Flight attendants will give it to you after the water landing
e.	Flight attendants will give it to you just before exiting the plane
f.	Under your seat

**TABLE A1**  
(CONTINUED)

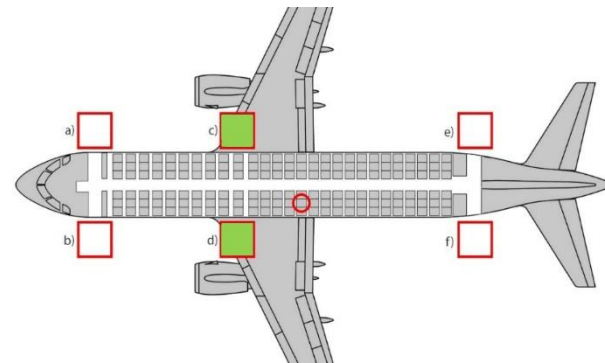
<b>5. You are sitting in the position highlighted by the red circle. Which exits do you have to use to evacuate?</b>	
	
<b>6. If the oxygen masks drop down at high altitude, what do you have to do immediately?</b>	
a.	Try to hold your breath until the plane descends to a low altitude
b.	Immediately help children and elderly people nearby, and then wear the oxygen mask
c.	Wait for flight assistants' detailed instructions before doing anything
d.	Wear the life vest
e.	Immediately wear the oxygen mask, and then help people nearby
f.	Assume the brace position
<b>7. If there is smoke in the cabin, what do you have to do while going to the exit?</b>	
a.	Keep wearing the oxygen mask
b.	Bend over or crawl
c.	Follow floor lights
d.	Run upright as fast as possible outside the smoke
e.	None of the previous answers
<b>8. During an emergency, what do you have to do with your luggage?</b>	
a.	Be sure not to leave it on the plane
b.	Take all and only the luggage you can carry
c.	Take all and only the most valuable items
d.	Find and take only the documents and the mobile phone inside your luggage
e.	Leave everything on the plane
<b>9. What is the purpose of the floor lights?</b>	
a.	They show the dangerous zones
b.	They show the dangerous zones, but only if they blink
c.	They show the safe zones, but only if they blink
d.	They show the path to the exits
e.	None of the previous answers
<b>10. During an evacuation on land, what do you have to do after you exit through a door on a wing?</b>	
a.	Jump down from the wing immediately
b.	Go down using the slides
c.	Wait until flight attendants give precise instructions
d.	Wait on the wing until the arrival of the rescue team
e.	None of the previous answers
<b>11. During an evacuation, when do you have to open a closed exit door?</b>	
a.	Always, as soon as you reach it
b.	If the exit is not on wings and there is no danger outside
c.	If there is no fire, debris, or other danger beyond the door
d.	If the door is over the water level, and there is no danger outside
e.	Never, only authorized flight attendants can open the exit doors
<b>12. During a water landing, what do you have to do as soon as you reach an emergency raft?</b>	
a.	Inflate the life vest, if you have not inflated it previously
b.	Sit down as close to the plane as possible
c.	Among the available positions on the raft, sit on the farthest from the plane
d.	If you can swim, dive into the sea and reach the dry land
e.	Stand and move your arms to call for rescue

TABLE A2

TYPES OF ERRORS THAT CAN BE MADE IN THE GAME (IGETs): NAME OF EACH IGET, GAME LEVELS IN WHICH THE IGET CAN BE MADE, DESCRIPTION OF THE IGET, AND QUESTIONS OF THE KNOWLEDGE QUESTIONNAIRE THAT ARE RELATED TO THE IGET.

Name	Level	Description	Related questions
AllowOther ToKeep Luggage	3	Players did not tell to leave luggage to a passenger with luggage who was slowing them down, and were thus reached by fire	8
Block Passengers OnRaft	2	Players sat at the beginning of the raft, blocking the other passengers	12
External UnsafeArea	1, 3, 4	After evacuating the aircraft, players wasted time in an unsafe area outside	10
FallFrom Wings	1, 4	Players fell from the aircraft wings	10
GoInWrong Direction	1, 2	Players went towards the front of the aircraft, but the closest exit was in the opposite direction	
InflateLife VestEarly	2	Players inflated the life vest when they were still inside the aircraft after a water landing	3
KeepLuggage	1, 2, 3, 4	Players kept their luggage, slowing down the evacuation	8
LongWay InsideSmoke	4	Players followed a long path inside smoke instead of the shortest path to the exit	5, 9
NoBrace	1, 2, 4	Players did not assume the brace position during the emergency landing	2
NoInflate LifeVest	2	Players did not inflate the life vest when leaving the aircraft after a water landing	3, 12
NoOxygen Mask	4	Players did not wear the oxygen mask when it dropped down	6
NoSeatBelt	1, 2, 3, 4	Players did not fasten the seat belt during turbulence, landing, or taxiing	
NoTake LifeVest	2	Players did not take the life vest after a water landing	4
OpenDoor UnderWater	2	Players opened a door that was under water level	11
OpenDoor WithDebris	3	Players opened a door behind which there was debris	11
OpenDoor WithFire	4	Players opened a door behind which there was fire	11
StandInside Smoke	1, 3, 4	Players did not bend down inside smoke, and thus suffocated	7
StandUp Without LifeVest	2	After a water landing, players stood up to leave their seat before taking the life vest	4
TakeLifeVest OnLand	1, 3, 4	Players wasted time to take the life vest when the landing was not on water	
TakeLuggage	1, 2, 3, 4	Players took their luggage, slowing down the evacuation	8
WasteTimeOn Aisle	1, 2, 3, 4	Players wasted time when they were in the aisle, and were reached by fire or by water	
WasteTimeOn Seat	1, 2, 3, 4	Players wasted time when they were in their seat, and were reached by fire or by water	

### A.3 Acceptable Completion Times

To determine the range of acceptable completion times, we first visually observed that the distributions of pre-test and post-test completion times in Fig. A1 appear as the overlap of two distributions: for pre-tests, the distribution of times appears as the overlap of a normal distribution with mean at about 190 s, and a skewed distribution with mean at about 45 s. For post-tests, the distribution appears as the overlap of a normal distribution with mean at about 145 s and a skewed distribution with mean at about 30 s. Fig. A1 illustrates completion times in the 0-900 s range; 13,088 players in the pre-test and 3,402 in the post-test took even more than 900 s. A recent meta-analysis [43], based on 190 studies, estimated that the average silent reading rate for adults in English is 238 words per minute and that most adults fall in the range of 175-300 words per minute for non-fiction. Since the twelve questions with all their possible answers contain about 700 words, expected reading times for the knowledge questionnaire should range between 140 s and 240 s, averaging at 176 s. These values are in line with the normal distribution for pre-test times with mean at about 190 s. Very high reading rates (e.g., 700 words per minute) are unlikely without severe loss of text understanding [44], so the skewed distributions for pre-test and post-test times, whose means are below 60 s, do not represent reliable tests. Therefore, the lower range for reliable pre-test completion times was set at 140 s, the minimum expected time for a reader at the upper range of reading rate (300 words per minute) estimated in [43]. The value also coincides with the intersection of the two curves in Fig. A1. The upper range for reliable test completion times was instead set at 600 s. We decided to add extra time to the 240 s expected for the lower range of reading rate (175 words per minute) estimated in [43] after visually observing the distribution of pre-test times, and also to take into account that: i) several players worldwide were not native English speakers, and the questionnaire was available only in English, ii) completion times include reading times as well as decision times and motor times to move the finger to select and confirm the answers. We noticed that the normal distribution with reliable times for post-tests is shifted towards lower values with respect to the one observed for pre-tests, but this was expected because the reading rates from the literature concern texts that players had never read before, while in our post-test players read the same questionnaire they had already read in the pre-test. They have also been exposed to the terminology and the described situations by playing the game levels. Therefore, by visually observing the post-test completion time distribution, we set again the lower range for reliable post-test completion times at the intersection of the two curves (105 s).

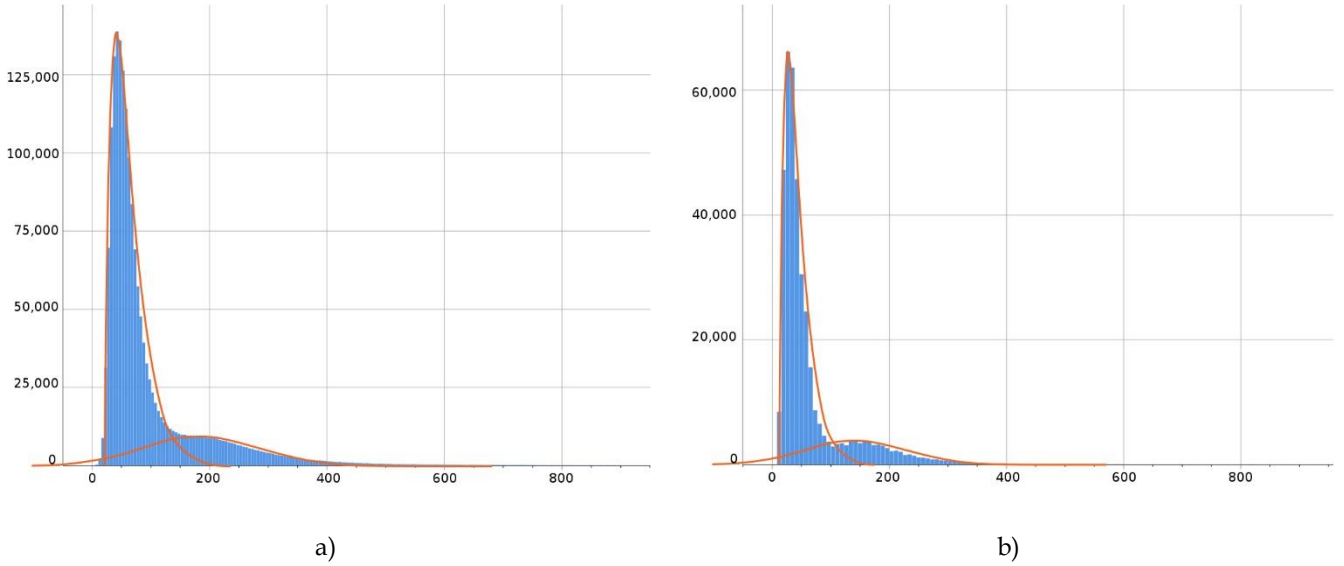


Fig. A1. Distributions of completion times (in blue) for (a) pre-test and (b) post-test, with normal and skewed distribution curves (in orange) that approximate the two parts of the distributions.

#### A.4 Device Languages and Time Zones

Table A3 lists the languages set on players' devices. For each language, the table reports the number and the percentage of devices on which that language was set. Fig. A2 instead shows the distribution of players' time zones set on their devices at the first run of the SG. To keep users' time zones consistent all over the year, we disregarded daylight saving time and always used UTC standard time. Table A4 and Fig. A3 show, respectively, the languages and the time zones set on the devices of the players who completed the pre-test in an acceptable time range.

#### A.5 Statistical Tests

To analyze the pre-post difference in overall knowledge questionnaire score and in each question score (RQ1), we used Wilcoxon signed-ranks test because the measure was repeated twice over the same participants and we could not assume that the population was normally distributed (Kolmogorov-Smirnov test  $p < 0.001$ , for all question scores and overall knowledge score).

To check if playing the game levels more than once led to a change in wrong in-game behaviors (RQ2), we compared the total number of occurred IGETs between the

first and the second session by using Wilcoxon signed-ranks test, because the measure was repeated twice over the same participants and we could not assume that the population was normally distributed (Kolmogorov-Smirnov test  $p < 0.001$  for each game level and all game levels combined). Then, for each IGET and game level, we compared the occurrence of the IGET between the first and the second session of the level by using McNemar's test, because the measure was repeated twice with the same participants and it was dichotomous. To extend the analyses to the third session of each game level, we used Friedman and Cochran Q tests on, respectively, the total number of occurred IGETs and the occurrence of IGET, because the first measure could not be assumed to be normally distributed (Kolmogorov-Smirnov test  $p < 0.001$ ), the second was dichotomous, and both were repeated three times. Wilcoxon signed-ranks and McNemar's tests were respectively used for post-hoc, with Bonferroni correction.

Finally, Mann-Whitney test was used to compare the pre-post difference in IGET-related question score between the group of players who made an IGET related to the question at least once in the game and the group who never made it (RQ3), because the samples were independent and we could not assume that the population was normally distributed (Kolmogorov-Smirnov test  $p < 0.001$  for all questions).



**TABLE A3**  
DEVICE LANGUAGES AND THEIR FREQUENCY CONSIDERING ALL PLAYERS.

Language	Frequency	Percentage
Afrikaans	84	0.003%
Arabic	79,100	2.443%
Basque	15	<0.001%
Belarusian	2	<0.001%
Bulgarian	9,611	0.297%
Catalan	854	0.026%
Chinese	72,845	2.249%
ChineseSimplified	1,224	0.038%
ChineseTraditional	578	0.018%
Czech	31,398	0.970%
Danish	3,464	0.107%
Dutch	24,305	0.751%
English	1,099,290	33.946%
Estonian	2,681	0.083%
Faroese	1	<0.001%
Finnish	9,839	0.304%
French	107,593	3.323%
German	85,964	2.655%
Greek	13,577	0.419%
Hebrew	6,089	0.188%
Hungarian	18,426	0.569%
Icelandic	259	0.008%
Indonesian	165,382	5.107%
Italian	53,309	1.646%
Japanese	15,983	0.494%
Korean	31,261	0.965%
Latvian	3,387	0.105%
Lithuanian	7,188	0.222%
Norwegian	776	0.024%
Polish	62,550	1.932%
Portuguese	237,234	7.326%
Romanian	38,257	1.181%
Russian	378,115	11.676%
SerboCroatian	136	0.004%
Slovak	10,213	0.315%
Slovenian	2,365	0.073%
Spanish	418,106	12.911%
Swedish	10,932	0.338%
Thai	55,062	1.700%
Turkish	60,892	1.880%
Ukrainian	12,988	0.401%
Vietnamese	56,971	1.759%
Unknown	49,996	1.544%

**TABLE A4**  
DEVICE LANGUAGES AND THEIR FREQUENCY CONSIDERING PLAYERS WHO COMPLETED THE PRE-TEST IN AN ACCEPTABLE TIME.

Language	Frequency	Percentage
Afrikaans	9	0.003%
Arabic	1,783	0.556%
Basque	2	0.001%
Bulgarian	376	0.117%
Catalan	89	0.028%
Chinese	2,312	0.720%
ChineseSimplified	143	0.045%
ChineseTraditional	95	0.030%
Czech	1,922	0.599%
Danish	588	0.183%
Dutch	3,923	1.222%
English	235,958	73.528%
Estonian	223	0.069%
Faroese	1	<0.001%
Finnish	1,162	0.362%
French	6,155	1.918%
German	7,220	2.250%
Greek	1,523	0.475%
Hebrew	355	0.111%
Hungarian	923	0.288%
Icelandic	63	0.020%
Indonesian	5,791	1.805%
Italian	3,410	1.063%
Japanese	1,154	0.360%
Korean	913	0.285%
Latvian	219	0.068%
Lithuanian	455	0.142%
Norwegian	188	0.059%
Polish	2,627	0.819%
Portuguese	6,737	2.099%
Romanian	2,173	0.677%
Russian	6,829	2.128%
SerboCroatian	19	0.006%
Slovak	554	0.173%
Slovenian	217	0.068%
Spanish	16,217	5.053%
Swedish	1,763	0.549%
Thai	1,156	0.360%
Turkish	1,647	0.513%
Ukrainian	245	0.076%
Vietnamese	1,253	0.390%
Unknown	2,516	0.784%

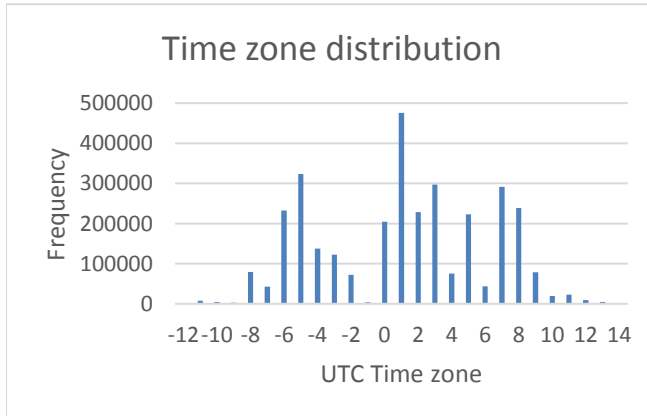


Fig. A2. Distribution of players' time zone as set on their devices at the first run of the SG (all 3,238,302 players).

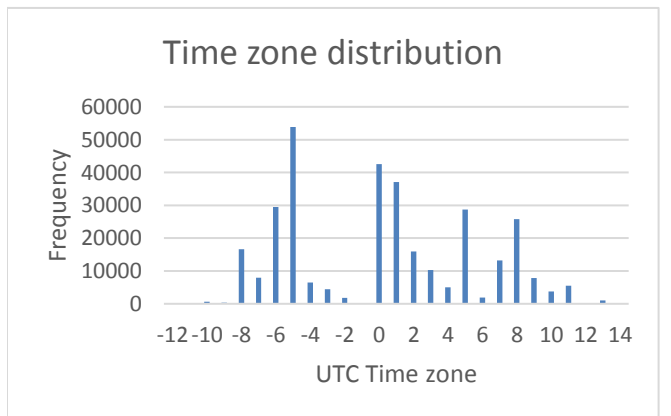


Fig. A3. Distribution of players' time zone as set on their devices at the first run of the SG (320,908 participants who completed the pre-test in an acceptable time).

## APPENDIX B – ADDITIONAL RESULTS

Considering RQ1 and each specific topic in the knowledge questionnaire, Table B1 shows mean and standard deviation of question score (pre-test and post-test) for each of the 12 questions as well as Wilcoxon signed-rank test statistics ( $Z$ ), two-tailed significance ( $p$ ), and effect size ( $|r|$ ). Wilcoxon signed-rank test showed a statistically significant improvement ( $p < 0.001$ ) in question scores for all questions except Question 9, for which the difference in question score was negligible and not statistically significant.

Considering RQ2 and each specific IGET, Table B2 reports the occurrence of IGET in the first and the second session for each game level for players who completed those two sessions. To show if the difference in the occurrence of IGET between the two sessions was significant, the table reports McNemar's test statistics ( $\chi^2$ ), significance ( $p$ ), and odds ratio. Statistical significance was reached for 45 (IGET, game level) pairs ( $p < 0.001$  for all), and the IGETs in the only two pairs that did not reach significance were NoTakeLifeVest and StandUpWithoutLifeVest. Among the 42 (IGET, game level) pairs where the difference was practically significant, 37 show a decrease in occurrence of IGET as hypothesized, and 5 an increase. Such increase concerned only ExternalUnsafeArea (in L1, L3, and L4), GoInWrongDirection (in L2 only) and WasteTimeOnAisle (in L2 only), all with a small odd ratio.

Considering the analysis of players who completed also a third session of the considered game level, Table B3 reports occurrence of each IGET for each game level in the first three sessions as well as Cochran Q test statistics (Cochran's  $Q$ ) and significance ( $p$ ) of the difference in the measure between the three sessions. Table B4 reports the McNemar's test statistics ( $\chi^2$ ), significance ( $p$ ), and odds ratio for all the post-hoc tests. In the overall comparisons, statistical significance was reached for all the 47 (IGET, game level) pairs ( $p < 0.005$  for NoTakeLifeVest, and  $p < 0.001$  for all the other pairs). Post-hoc tests reached statistical significance for all except four comparisons (first vs. second session for NoTakeLifeVest, first vs. third session for NoTakeLifeVest, StandUpWithoutLifeVest, and OpenDoorWithDebris).

Considering RQ3, Table B5 reports the pre-post difference in IGET-related question score for each question and each IGET related to the question in the group of players who made the IGET at least once in the game and in the group who never made the IGET. More precisely, for each (question, IGET) pair, Table B5 reports the mean and standard deviation of the measure for both groups. Moreover, the table shows if the difference in the measure between the two groups of players was significant by reporting Mann-Whitney test statistics ( $Z$ ), two-tailed significance ( $p$ ), and effect size ( $r$ ).

TABLE B1

MEAN (M) AND STANDARD DEVIATION (SD) OF QUESTION SCORE (PRE-TEST AND POST-TEST) FOR EACH OF THE 12 QUESTIONS, WILCOXON SIGNED-RANK TEST STATISTICS ( $Z$ ), TWO-TAILED SIGNIFICANCE ( $P$ ), AND EFFECT SIZE ( $|r|$ ) OF THE COMPARISON BETWEEN PRE-TEST AND POST-TEST.

Question	Pre-test		Post-test		$Z$	$p$	$ r $
	M	SD	M	SD			
1	0.41	0.42	0.51	0.43	-46.60	< 0.001	0.16
2	0.21	0.41	0.57	0.49	-113.87	< 0.001	0.38
3	0.24	0.43	0.67	0.47	-127.56	< 0.001	0.42
4	0.68	0.47	0.77	0.42	-41.74	< 0.001	0.14
5	0.58	0.35	0.64	0.35	-32.26	< 0.001	0.11
6	0.60	0.49	0.71	0.46	-43.25	< 0.001	0.14
7	0.38	0.39	0.58	0.39	-83.56	< 0.001	0.28
8	0.75	0.43	0.92	0.28	-75.63	< 0.001	0.25
9	0.79	0.41	0.78	0.41	-1.71	0.087	0.01
10	0.53	0.50	0.84	0.36	-101.39	< 0.001	0.34
11	0.22	0.35	0.41	0.40	-80.48	< 0.001	0.27
12	0.41	0.33	0.63	0.34	-99.89	< 0.001	0.33

**TABLE B2**  
**OCCURRENCE OF IGET IN THE FIRST AND THE SECOND SESSION FOR EACH GAME LEVEL AND IGET, MCNEMAR'S TEST STATISTICS ( $\chi^2$ ), SIGNIFICANCE (P), AND ODDS RATIO OF THE DIFFERENCE BETWEEN THE TWO SESSIONS.**

Game Level	In-game error type (IGET)	Occurrence of IGET in the first session	Occurrence of IGET in the second session	$\chi^2$	p	odds ratio
1	ExternalUnsafeArea	7%	11%	3969.67	< 0.001	1.63
1	FallFromWings	18%	7%	25119.16	< 0.001	3.13
1	GoInWrongDirection	56%	41%	19882.27	< 0.001	1.92
1	KeepLuggage	31%	9%	60060.40	< 0.001	4.74
1	NoBrace	23%	9%	30657.99	< 0.001	3.17
1	NoSeatBelt	44%	17%	70683.46	< 0.001	4.31
1	StandInsideSmoke	49%	20%	77204.24	< 0.001	4.40
1	TakeLifeVestOnLand	67%	26%	140187.40	< 0.001	9.63
1	TakeLuggage	47%	18%	75750.69	< 0.001	4.40
1	WasteTimeOnAisle	34%	15%	44944.13	< 0.001	3.62
1	WasteTimeOnSeat	23%	5%	57268.10	< 0.001	7.16
2	BlockPassengersOnRaft	75%	45%	80099.15	< 0.001	4.48
2	GoInWrongDirection	34%	44%	10903.57	< 0.001	1.65
2	InflateLifeVestEarly	36%	29%	3724.85	< 0.001	1.31
2	KeepLuggage	10%	7%	2808.31	< 0.001	1.56
2	NoBrace	8%	6%	1568.55	< 0.001	1.43
2	NoInflateLifeVest	67%	35%	83928.67	< 0.001	4.12
2	NoSeatBelt	28%	16%	21983.70	< 0.001	2.43
2	NoTakeLifeVest	2%	2%	0.34	0.562	1.01
2	OpenDoorUnderWater	34%	21%	20938.05	< 0.001	2.10
2	StandUpWithoutLifeVest	17%	17%	0.74	0.389	1.01
2	TakeLuggage	23%	16%	8219.40	< 0.001	1.73
2	WasteTimeOnAisle	3%	3%	307.29	< 0.001	1.26
2	WasteTimeOnSeat	8%	4%	4880.21	< 0.001	1.99
3	AllowOtherToKeepLuggage	43%	19%	37186.61	< 0.001	3.65
3	ExternalUnsafeArea	19%	23%	2027.79	< 0.001	1.35
3	KeepLuggage	4%	2%	1422.25	< 0.001	1.91
3	NoSeatBelt	10%	9%	282.84	< 0.001	1.17
3	OpenDoorWithDebris	15%	14%	32.91	< 0.001	1.05
3	StandInsideSmoke	21%	13%	7378.70	< 0.001	1.99
3	TakeLifeVestOnLand	34%	22%	12369.74	< 0.001	2.28
3	TakeLuggage	7%	5%	1076.71	< 0.001	1.51
3	WasteTimeOnAisle	22%	23%	83.17	< 0.001	1.06
3	WasteTimeOnSeat	32%	20%	12082.10	< 0.001	2.15
4	ExternalUnsafeArea	8%	12%	2261.12	< 0.001	1.60
4	FallFromWings	52%	15%	72303.51	< 0.001	7.76
4	KeepLuggage	6%	3%	4618.92	< 0.001	2.92
4	LongWayInsideSmoke	61%	37%	28078.25	< 0.001	2.84
4	NoBrace	15%	12%	1027.98	< 0.001	1.32
4	NoOxygenMask	13%	8%	3887.78	< 0.001	1.82
4	NoSeatBelt	34%	24%	7082.01	< 0.001	1.79
4	OpenDoorWithFire	52%	28%	25526.18	< 0.001	2.50
4	StandInsideSmoke	38%	20%	22336.84	< 0.001	3.01
4	TakeLifeVestOnLand	27%	18%	6887.40	< 0.001	1.95
4	TakeLuggage	13%	7%	5676.23	< 0.001	2.25
4	WasteTimeOnAisle	46%	12%	61071.76	< 0.001	5.99
4	WasteTimeOnSeat	17%	8%	9872.12	< 0.001	2.67

**TABLE B3**  
**OCCURRENCE OF IGET IN THE FIRST, THE SECOND, AND THE THIRD SESSION FOR EACH GAME LEVEL AND IGET, COCHRAN Q**  
**TEST STATISTICS (COCHRAN'S Q), AND SIGNIFICANCE (P).**

Game level	In-game error type (IGET)	First session	Second session	Third session	Cochran's Q	p
1	ExternalUnsafeArea	8%	14%	12%	3322.38	< 0.001
1	FallFromWings	19%	8%	6%	16540.43	< 0.001
1	GoInWrongDirection	56%	42%	39%	12228.77	< 0.001
1	KeepLuggage	30%	9%	6%	46817.23	< 0.001
1	NoBrace	23%	10%	9%	18566.79	< 0.001
1	NoSeatBelt	44%	19%	14%	49387.36	< 0.001
1	StandInsideSmoke	48%	22%	16%	52182.35	< 0.001
1	TakeLifeVestOnLand	66%	26%	19%	103323.83	< 0.001
1	TakeLuggage	45%	18%	13%	54529.80	< 0.001
1	WasteTimeOnAisle	35%	17%	13%	30582.74	< 0.001
1	WasteTimeOnSeat	25%	5%	3%	48741.03	< 0.001
2	BlockPassengersOnRaft	74%	45%	36%	72525.32	< 0.001
2	GoInWrongDirection	36%	45%	46%	6192.91	< 0.001
2	InflateLifeVestEarly	33%	31%	24%	4129.42	< 0.001
2	KeepLuggage	10%	7%	6%	3065.68	< 0.001
2	NoBrace	8%	6%	5%	1682.77	< 0.001
2	NoInflateLifeVest	67%	37%	25%	80740.53	< 0.001
2	NoSeatBelt	29%	16%	12%	24865.65	< 0.001
2	NoTakeLifeVest	2%	2%	2%	11.43	0.003
2	OpenDoorUnderWater	32%	22%	19%	12207.69	< 0.001
2	StandUpWithoutLifeVest	17%	18%	16%	148.54	< 0.001
2	TakeLuggage	23%	16%	15%	7050.78	< 0.001
2	WasteTimeOnAisle	3%	4%	4%	663.84	< 0.001
2	WasteTimeOnSeat	8%	4%	4%	4352.00	< 0.001
3	AllowOtherToKeepLuggage	45%	21%	15%	32564.42	< 0.001
3	ExternalUnsafeArea	19%	26%	21%	2286.68	< 0.001
3	KeepLuggage	4%	2%	1%	1849.00	< 0.001
3	NoSeatBelt	11%	10%	8%	714.29	< 0.001
3	OpenDoorWithDebris	13%	16%	14%	341.21	< 0.001
3	StandInsideSmoke	22%	14%	11%	6255.51	< 0.001
3	TakeLifeVestOnLand	34%	23%	19%	10919.24	< 0.001
3	TakeLuggage	7%	5%	4%	1141.09	< 0.001
3	WasteTimeOnAisle	23%	26%	25%	319.36	< 0.001
3	WasteTimeOnSeat	33%	22%	17%	10756.13	< 0.001
4	ExternalUnsafeArea	9%	14%	12%	1557.49	< 0.001
4	FallFromWings	51%	15%	10%	58512.02	< 0.001
4	KeepLuggage	7%	3%	2%	5040.02	< 0.001
4	LongWayInsideSmoke	60%	41%	31%	21766.08	< 0.001
4	NoBrace	16%	13%	11%	1080.07	< 0.001
4	NoOxygenMask	14%	8%	6%	3868.17	< 0.001
4	NoSeatBelt	36%	26%	20%	8265.36	< 0.001
4	OpenDoorWithFire	50%	31%	21%	20453.78	< 0.001
4	StandInsideSmoke	39%	21%	15%	21193.55	< 0.001
4	TakeLifeVestOnLand	27%	18%	15%	7092.61	< 0.001
4	TakeLuggage	14%	8%	7%	5104.11	< 0.001
4	WasteTimeOnAisle	46%	12%	7%	53683.16	< 0.001
4	WasteTimeOnSeat	17%	8%	5%	10674.86	< 0.001

TABLE B4

MCNEMAR'S TEST STATISTICS ( $\chi^2$ ), SIGNIFICANCE (P), AND ODDS RATIO OF POST-HOC TESTS COMPARING OCCURRENCE OF IGET IN FIRST VS. SECOND, SECOND VS. THIRD, AND FIRST VS. THIRD SESSION.

Ga-me le-vel	In-game error type (IGET)	First vs. second session			Second vs. third session			First vs. third session		
		$\chi^2$	p	odds ratio	$\chi^2$	p	odds ratio	$\chi^2$	p	odds ratio
1	ExternalUnsafeArea	3167.26	< 0.001	1.90	311.44	< 0.001	1.21	1601.37	< 0.001	1.61
1	FallFromWings	9282.48	< 0.001	2.84	169.69	< 0.001	1.20	11466.65	< 0.001	3.35
1	GoInWrongDirection	7297.96	< 0.001	1.85	247.33	< 0.001	1.13	10008.79	< 0.001	2.08
1	KeepLuggage	22612.54	< 0.001	4.56	1906.69	< 0.001	1.87	32662.27	< 0.001	8.31
1	NoBrace	11120.06	< 0.001	2.87	86.86	< 0.001	1.12	12823.53	< 0.001	3.18
1	NoSeatBelt	25815.13	< 0.001	3.87	1651.93	< 0.001	1.51	35945.86	< 0.001	5.49
1	StandInsideSmoke	25686.81	< 0.001	3.77	2850.88	< 0.001	1.70	39275.20	< 0.001	5.91
1	TakeLifeVestOnLand	55583.28	< 0.001	9.55	3249.34	< 0.001	1.79	69122.83	< 0.001	13.68
1	TakeLuggage	27462.09	< 0.001	4.09	2098.02	< 0.001	1.63	39404.09	< 0.001	6.43
1	WasteTimeOnAisle	15944.70	< 0.001	3.17	1238.50	< 0.001	1.45	23185.76	< 0.001	4.27
1	WasteTimeOnSeat	25327.41	< 0.001	7.43	727.21	< 0.001	1.64	30228.80	< 0.001	11.08
2	BlockPassengersOnRaft	37698.31	< 0.001	4.23	4740.93	< 0.001	1.63	58018.08	< 0.001	6.10
2	GoInWrongDirection	3883.78	< 0.001	1.52	116.29	< 0.001	1.08	5145.76	< 0.001	1.61
2	InflateLifeVestEarly	225.38	< 0.001	1.10	2303.40	< 0.001	1.39	3885.67	< 0.001	1.52
2	KeepLuggage	1288.29	< 0.001	1.52	311.83	< 0.001	1.27	2743.74	< 0.001	1.90
2	NoBrace	800.81	< 0.001	1.42	118.73	< 0.001	1.16	1474.18	< 0.001	1.63
2	NoInflateLifeVest	35832.48	< 0.001	3.61	8376.34	< 0.001	1.98	65043.55	< 0.001	6.46
2	NoSeatBelt	11296.13	< 0.001	2.43	2081.24	< 0.001	1.57	20160.12	< 0.001	3.46
2	NoTakeLifeVest	2.47	0.348	1.03	11.64	0.002	1.08	3.12	0.232	1.04
2	OpenDoorUnderWater	5721.37	< 0.001	1.72	837.06	< 0.001	1.28	10225.94	< 0.001	2.13
2	StandUpWithoutLifeVest	95.65	< 0.001	1.09	129.20	< 0.001	1.11	1.64	0.600	1.01
2	TakeLuggage	3659.44	< 0.001	1.67	317.70	< 0.001	1.19	5748.32	< 0.001	1.91
2	WasteTimeOnAisle	207.65	< 0.001	1.29	134.03	< 0.001	1.21	648.43	< 0.001	1.55
2	WasteTimeOnSeat	2548.46	< 0.001	2.00	55.64	< 0.001	1.13	3216.79	< 0.001	2.21
3	AllowOtherToKeepLuggage	15706.71	< 0.001	3.49	1985.55	< 0.001	1.72	24627.86	< 0.001	5.39
3	ExternalUnsafeArea	2086.38	< 0.001	1.57	1001.84	< 0.001	1.37	231.46	< 0.001	1.17
3	KeepLuggage	857.16	< 0.001	2.16	113.30	< 0.001	1.43	1481.31	< 0.001	2.98
3	NoSeatBelt	106.63	< 0.001	1.15	274.38	< 0.001	1.28	696.80	< 0.001	1.46
3	OpenDoorWithDebris	280.76	< 0.001	1.22	222.90	< 0.001	1.21	5.73	0.050	1.03
3	StandInsideSmoke	2848.70	< 0.001	1.88	501.59	< 0.001	1.35	5390.45	< 0.001	2.48
3	TakeLifeVestOnLand	5269.14	< 0.001	2.25	852.07	< 0.001	1.44	8824.29	< 0.001	2.84
3	TakeLuggage	551.53	< 0.001	1.55	68.36	< 0.001	1.20	945.24	< 0.001	1.81
3	WasteTimeOnAisle	300.02	< 0.001	1.19	23.99	< 0.001	1.05	152.60	< 0.001	1.13
3	WasteTimeOnSeat	4351.11	< 0.001	1.98	1330.02	< 0.001	1.52	9456.80	< 0.001	2.86
4	ExternalUnsafeArea	1513.89	< 0.001	1.73	263.27	< 0.001	1.25	553.00	< 0.001	1.41
4	FallFromWings	30380.81	< 0.001	7.00	1762.38	< 0.001	1.80	38788.41	< 0.001	10.82
4	KeepLuggage	2466.07	< 0.001	3.20	212.22	< 0.001	1.58	3692.86	< 0.001	4.76
4	LongWayInsideSmoke	8275.89	< 0.001	2.31	3382.96	< 0.001	1.76	19035.12	< 0.001	3.69
4	NoBrace	305.01	< 0.001	1.24	236.54	< 0.001	1.23	1063.92	< 0.001	1.52
4	NoOxygenMask	1665.79	< 0.001	1.78	356.49	< 0.001	1.37	3355.17	< 0.001	2.38
4	NoSeatBelt	2632.40	< 0.001	1.67	1625.69	< 0.001	1.57	7646.26	< 0.001	2.51
4	OpenDoorWithFire	7718.50	< 0.001	2.09	2750.12	< 0.001	1.70	18181.52	< 0.001	3.52
4	StandInsideSmoke	9446.58	< 0.001	2.82	1906.40	< 0.001	1.72	17394.63	< 0.001	4.59
4	TakeLifeVestOnLand	3066.25	< 0.001	1.94	730.38	< 0.001	1.44	6084.16	< 0.001	2.62
4	TakeLuggage	2852.29	< 0.001	2.30	106.15	< 0.001	1.21	3790.21	< 0.001	2.67
4	WasteTimeOnAisle	26941.58	< 0.001	5.78	1460.52	< 0.001	1.78	35703.61	< 0.001	10.24
4	WasteTimeOnSeat	4753.35	< 0.001	2.72	863.82	< 0.001	1.72	8472.11	< 0.001	4.33

TABLE B5

MEAN (M) AND STANDARD DEVIATION (SD) OF THE PRE-POST DIFFERENCE IN IGET-RELATED QUESTION SCORE FOR EACH QUESTION AND EACH RELATED IGET FOR THE GROUP OF PLAYERS WHO MADE THE IGET IN THE GAME AND THE GROUP WHO NEVER MADE IT, MANN-WHITNEY TEST STATISTICS (Z), TWO-TAILED SIGNIFICANCE (P), AND EFFECT SIZE (|r|) OF THE DIFFERENCE IN THE MEASURE BETWEEN THE GROUPS.

Question	In-game error type (IGET)	Pre-post difference in question score among players who made the IGET		Pre-post difference in question score among players who did NOT make the IGET		Z	p	r
		M	SD	M	SD			
1	NoSeatBelt	0.10	0.45	0.11	0.47	-0.28	0.780	0.00
2	NoBrace	0.35	0.56	0.36	0.57	-1.62	0.106	0.01
3	InflateLifeVestEarly	0.52	0.56	0.35	0.57	-31.56	< 0.001	0.15
	NoInflateLifeVest	0.50	0.57	0.36	0.57	-25.73	< 0.001	0.12
4	NoTakeLifeVest	0.20	0.52	0.09	0.45	-6.47	< 0.001	0.03
	StandUpWithoutVest	0.11	0.46	0.09	0.45	-3.69	< 0.001	0.02
5	LongWayInsideSmoke	0.08	0.40	0.04	0.42	-9.91	< 0.001	0.05
6	NoOxygenMask	0.19	0.52	0.09	0.49	-13.81	< 0.001	0.06
7	StandInsideSmoke	0.24	0.43	0.14	0.45	-24.14	< 0.001	0.11
8	AllowOtherToKeepLuggage	0.21	0.49	0.15	0.41	-11.57	< 0.001	0.05
	KeepLuggage	0.36	0.54	0.13	0.39	-43.88	< 0.001	0.21
	TakeLuggage	0.32	0.53	0.08	0.34	-57.25	< 0.001	0.27
9	LongWayInsideSmoke	-0.01	0.50	0.01	0.46	-4.27	< 0.001	0.02
10	ExternalUnsafeArea	0.30	0.58	0.31	0.57	-2.11	0.035	0.01
	FallFromWings	0.38	0.57	0.24	0.56	-26.46	< 0.001	0.12
11	OpenDoorUnderWater	0.25	0.47	0.14	0.42	-24.37	< 0.001	0.11
	OpenDoorWithDebris	0.23	0.46	0.18	0.44	-9.52	< 0.001	0.04
	OpenDoorWithFire	0.24	0.44	0.15	0.44	-24.19	< 0.001	0.11
12	BlockPassengersOnRaft	0.23	0.41	0.20	0.41	-7.61	< 0.001	0.04
	NoLifeVest	0.24	0.41	0.20	0.41	-9.62	< 0.001	0.05

## ADDITIONAL REFERENCES

- [43] M. Brysbaert, "How many words do we read per minute? A review and meta-analysis of reading rate," *J. Mem. Lang.*, vol. 109, 2019, Art. no. 104047.
- [44] K. Rayner, E. R. Schotter, M. E. J. Masson, M. C. Potter, and R. Treiman, "So Much to Read, So Little Time," *Psychol. Sci. Public Interes.*, vol. 17, no. 1, pp. 4-34, 2016.