PEDAGOGICAL AGENTS: INFLUENCES OF ARTIFICIALLY GENERATED INSTRUCTOR PERSONAS ON TAKING CHANCES

Patrick Jost, Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway, patrick.jost@ntnu.no

ABSTRACT

Educational institutes are currently facing the new normality that an ongoing pandemic situation has brought to teaching and learning. Distributed learning with content that blends over several platforms and locations needs to be created with didactic expertise in a feasible manner. At the same time, the possibilities for creating and distributing digital content have developed rapidly. Advanced computing supports the creation of artificial images, natural speech, and even natural-looking but non-existent persons. Since such generative content is often also published under a Creative Commons license, it presents as viable option for designing learning content, assignments, or instructions for tasks. However, there is still limited evidence on how, for example, generated pedagogical agents (tutors) influence behaviour and decisions.

This study investigated the influences of artificially generated tutor personas in a decision-making task distributed internationally on the Google Play store. The field experiment extended the balloon analogue risk task (BART) with instructions from generated persona photographs to evaluate potential influences on risk-taking behaviour. In a between-subject design, either a female tutor, a male tutor, or no tutor picture at all was presented during the task. The results (N=74) show a higher risk propensity when displaying a male artificial instructor compared to a female instructor. Participants also proceed with greater caution when instructed by a female tutor as they reflect longer before initiating the next step to pump up the balloon. Further lines of research and experiences from the distribution of an investigative instruction app on Google Play are summarised in the conclusive implications.

Keywords: instructional design, artificial tutors, generative content, decision making, risk-taking, balloon analogue risk task, google play store, field experiment

1 Background and research objectives

Distributed mobile learning is becoming an important channel for educational institutes. As a possibility not only for transmitting content to learners but also for assigning tasks and monitoring progress, mobile apps also enable to reach those students that do not possess further devices. Although publishing apps in the app stores has become a fairly easy task, there are substantial concerns left for educators when creating courses, assignments or evaluations. Instructional design fitting to the distributed, mobile context and accurate evaluation of learner progress are two of those challenges. It is well known that influences from interaction and presentation are affecting the cognitive load of mobile learning apps (Jost et al., 2020; Mayer & Moreno, 1998; Sweller, 1994, 2016). Additionally, the reduced possibilities of personal interaction between teacher and student can contribute to the difficulties of instructions and assignments.

1.1 The application of pedagogical agents

On the other hand, the prospects of advanced computing are presenting opportunities to support the creation of learning instructions. Instructions could, for example, be supported by tutors that are introducing learning assignments. Additionally, they could also facilitate the evaluation of progress by having conversations with students that involve choices of correct/incorrect answers. Similar to a chatbot scenario (Berger et al., 2019) or virtual agents in Virtual Reality/Augmented Reality applications these guided dialogues could support focus and engagement (Wang et al., 2019). The concept of avatars has been explored since many years for underlying motivational effects but also regarding pedagogical and stereotypical gender effects

of such so-called pedagogical agents (Baylor, 2009; Gulz & Haake, 2006; Haake & Gulz, 2008). Visualisations of pedagogical agents (hereafter referred to as PA) have been widely investigated by researchers who reported effects regarding learning styles, learning impact and performance (Atorf et al., 2019; Laureano-Cruces et al., 2016; Veletsianos, 2010), influences on persuasion (Khan & Sutcliffe, 2014), student motivation/emotions (Bendou et al., 2017; Liew et al., 2016) and benefits concerning cognitive load (Schroeder, 2017). However, only recently advanced computing using neuronal networks (i.e. generative adversarial networks) is able to generate virtual portraits that are photorealistic and practically indistinguishable from representations of existing persons (Karras et al., 2019). Webservices with an API such as generated.photos or thispersondoesnotexist.com are thereby providing photograph generators with configurational options on emotion or ethnicity. These generated portrait images can, for example, be applied as pedagogical agent in a mobile learning app or as facilitating instructor in research-oriented apps.

Already, the creators of these services are researching realistic generative video personas which would then widen the application possibilities in tutoring even further. While the easy and feasible access would promote to use these portraits in learning content, limited knowledge exists about the impact they possess on human perception and behaviour. Yet, in instruction scenarios where progress is required to be monitored or is depending on exams, it is essential to know about influences that could distort results. On the other hand, specific influences from pedagogical instructors could also be used to personalise tutoring to the student. In this way, engagement could be supported by nudging theory concepts or pedagogical tutors could help to overcome biases in decision-making (Lieder et al., 2019). Awareness of influences on decisions is particularly important when evaluating learner progress, or when graded assignments are incorporated in the mobile learning environment. Equally significant is this recognition of effects when gathered insight is planned to be scientifically analysed, for example, in investigations with Serious Games (Jost, 2020).

1.2 Decision-making and the balloon analogue risk task

When looking at the educational context of remote mobile learning, specifically the evaluation of progress or assignments, are an essential application of instructional design of choices. However, typing considerable amounts of text is time intensive and tedious on most mobile devices. Therefore, mobile formative assessments are often designed as prompted dialogue choices (e.g. quizzes) and multiple-choice scenarios in primary (Mabruri et al., 2019), middle school (Miller & Cuevas, 2017) as well as in higher-education computing education (Chu et al., 2010). Concerning higher education, the main application areas for mobile learning apps are found to be language, health and computer science (Pimmer et al., 2016). Exemplary recent studies have implemented mobile courses for teaching digital electronics and computing/system analysis while integrating multiple-choice quizzes and instruction presentation by a female cartoon pedagogical agent (Oyelere et al., 2018; Rakhmawati & Firdha, 2018).

In this context, however, a substantial body of education research over the last 20 years has investigated multiple-choice decision-making and reported influences originating from individual risk propensity/aversion (Espinosa & Gardeazabal, 2005; Walker & Thompson, 2001; Yang & Tackie, 2016). The evidence described is indicating that multiple-choice evaluation is presenting a disadvantage for students with lower risk propensity. Studies thereby also imply a risk-related gender-bias resulting in a distorted assessment regarding male students that would have a higher risk propensity (Biria & Bahadoran, 2015). Accordingly, it is further proposed by some researchers to adjust multiple-choice evaluations and their scoring scheme to balance for such a bias (De Laet et al., 2015).

In consequence, it is important to learn about influences on risk behaviour from task designs with generated pedagogical agents. Pedagogical instructors could then contribute to balance learning experience and assessment biases or support personalised instructions. Similarly, known effects must be controlled when scientifically investigating an instructive task that has a presenter/instructor. Individual risk trait has been investigated with various psychological assessments. A systematic review conducted by Harrison et al. (2005) lists several instruments using Likert-scale measures and also the balloon analogue risk task (BART) (Lejuez et al., 2002). The BART differs as it is a task-oriented computer exercise that instructs persons to choose options with the underlying risk of losing all gains/points on an item comparable to a multiple-choice scenario. In the scenario simulated on-screen, the participants decide to further pump-up a balloon or collect an amount that has been summed up of each successful prior pump. Each balloon can explode,

however, at any given pump starting from the very first one which results in losing all gains/points from prior pumps.

Since the initial study by Lejuez et al. (2002), construct validity and applicability of the BART for assessing individual risk propensity has been demonstrated in numerous examinations in different contexts (e.g. Cazzell et al., 2012; Fukunaga et al., 2012; Hopko et al., 2006; Lauriola et al., 2014; Li et al., 2020; Rao et al., 2018). A recent study of MacLean et al. (2018) applied a mobile version of the simulated balloon pumping task to use in ambulatory settings and established the validity of measured indices to a laboratory variant. The BART, therefore, provides a proven option to assess risk propensity influences from generated instructor agents with an investigative mobile app. Additionally, as summarised by Evans (2007) and pointed out by Klein (2008), a shorter time of reasoning/reflection is indicating riskier decision making. Thus, a metric of reflection time before a decision step can provide further insight on influences from PA.

1.3 Research objectives

To get insight on influences from generated PA, this study extends the BART to create an investigative mobile app for a field experiment conducted on the Google Play Store. The research objective of this empirical evaluation was thereby twofold by investigating: (i) how generated instructor personas influence taking chances (i.e. risk propensity) in a decision exercise and by providing insight on (ii) how the design of evaluative mobile apps distributed on app stores can be improved.

After having introduced the background and research objectives (1), the following sections progress by describing the research design and the creation of the investigative mobile app (2). Thereby it is described how the study applies the Quest Game-Frame (Jost 2020) to create the balloon analogue risk task extended by generated instructor personas and analytical assessment (2). Subsequently, the findings are presented (3) and implications, limitations and further research are discussed (4) and summarised conclusively (5).

2 Methodology

For examining the influences of PA on taking decision risks a mobile research app closely following the conduct/constructs of the original BART experiment (Lejuez et al., 2002) was created with the research game framework suggested by Jost (2020). The mobile application thereby employs the same interaction buttons, interface phrasing and an equal instructional text that was used in the original experiment. Subsequently, the built research app, that represented a balloon game challenge was internationally published (in English) on the Google Play Store for *a two-month field experiment*. During the period, the mobile app was advertised by an equally international Google Ads campaign with a budget of € 0.50/day.

2.1 Research design and hypotheses

By assessing the risk propensity constructs of the BART and utilising the unobtrusive reflection metrics included in the research game framework of Jost (2020) the field experiment examines three experimental conditions in a between-subjects design (Figure 1). The exercise instructions were presented either by a female or male generated persona or by no persona at all (i.e. the original BART condition). Moreover, the instructor persona was pictured during the whole exercise as announcer of progress and stating the progress/collected money.

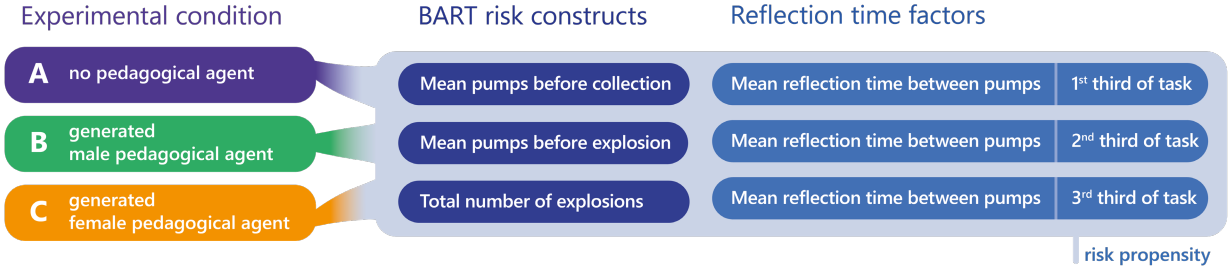

Figure 1. Experimental design to assess generated PA influences with the mobile balloon risk task

In the original BART, the construct of pumps before collection of the total accumulated money (per balloon) was mainly regarded as a measure of risk tendency. However, other dependent variables or the BART experiment were found to produce almost identical results (Lejuez et al., 2002). For this field test, it was thus chosen to also include the number of exploded balloons and the average of pumps before explosions to increase the validity of insight. Additionally, the chosen research game-framework has shown to reliably measure reflection times in decision tasks which was therefore implemented in the experimental design to allow for cross-validation of risk-taking influences of the PA with the BART results. A priori it was expected that there could be a trend to faster decision making later in the task due to getting used to the interaction and the exercise procedure. Therefore, the reflection time assessment was split in measuring the averages of the first, second and third segment of total balloon trials. It was further hypothesised that there would be an influence of the generated PA as cognitive/emotional effects have previously been reported and the generated portraits are of even higher quality (practically lifelike portraits) compared to those prior studies have applied.

Accordingly, for investigating the risk behaviour with the mobile app, the following null hypotheses were established:

> H_{0A}: 'There are no significant differences in measured risk-taking by the BART risk constructs when instructing participants with either a female, a male or no generated persona visualisation.'
>
> H_{0B}: 'There are no significant differences among the three instruction scenarios in reflection time between pumping decisions.'

2.2 Applying the Quest Game-Frame to design the PA experiment

For designing and publishing the investigative mobile app, the Quest Game-Frame (QGF) was applied (Jost, 2020) that is based on scripted components for the Unity 3D game engine (unity.com). While the framework is generally planned for multiple research point-of-interest mini-quests, it can also be used to create a single test scenario. The underlying unobtrusive assessment module is prepared for evaluation of decision settings as it protocols selections of dialogue options from the included dialogue system. It thus allows for quick creation of research designs with choice prompts, which are protocolled in a database with secure (https) connection regarding the reflection time between decisions. It further is designed for anonymous collection of data with a generated token that is regarding GDPR (Voigt & Von dem Bussche, 2017) rules by storing only data required for analysis and no data that would allow personal reference such as, for example, IP addresses. Data points were further saved in tables with table-column headings that permitted no conclusion on the column contents for people other than the data processors. In addition, research experiment components in the QGF are prepared to communicate with a private database server under the single authority of the data protection officer and thus are not openly accessible to other organisations as compared to general cloud storage solutions.

Another requirement for analytical mobile apps is to include an informed consent as well as a data privacy policy which was provided by the research game framework as introduction scene before app users could proceed to partake (see next section 2.3). Using these prepared data collection/analytical components, the BART research setting, as described in the original experiment (Lejuez et al., 2002) was designed and extended by the three experimental conditions. The instruction to the task which was stated in the original experiment was presented in *variant A* without any agent while in *variant B* a generated male portrait was used as pedagogical instructor and in *variant C* a generated female PA was instructing the app user.

The two generated portraits were randomly selected from a pre-rendered collection of 100.000 computed portraits that were generated from artificial intelligence trained on studio photographs in the process described by Karras et al. (2019). The company icons8 (icons8.com) made these available on Google Drive. The online random number generator random.org that creates real random numbers based on atmospheric noise was utilised to choose two random portraits from the collection. The inclusion criteria for this study were portraits that faced the camera, were visually appearing in the same age group (middle-aged adult) and had a friendly (i.e. slightly smiling) facial expression. Photographs with render flaws (Figure 2) such

as missing ears or other distorted facial features resorting from computing (e.g. holes in the skin) were also excluded.

Figure 2. Exemplary portraits from the pre-rendered collection not eligible for the study

The random selection from the numbered portrait collection took 4 rounds to select a suited male generated PA. The selected generated male picture was of white ethnicity; thus, the female PA was chosen to be also a portrait of a white person. After 9 next random rounds, a female portrait facing camera without rendering flaws and of a similar age appearance than the male picture was selected. The generated PA were included in the design of the mobile app as instructors of the BART and as announcer of progress. The resulting three experimental conditions as shown to the app users are illustrated in Figure 3, while Figure 4 shows the instructive balloon game app running on a smartphone with the female PA (condition C).

Figure 3. Experimental conditions: A – no PA; B – male generated PA; C – female generated PA

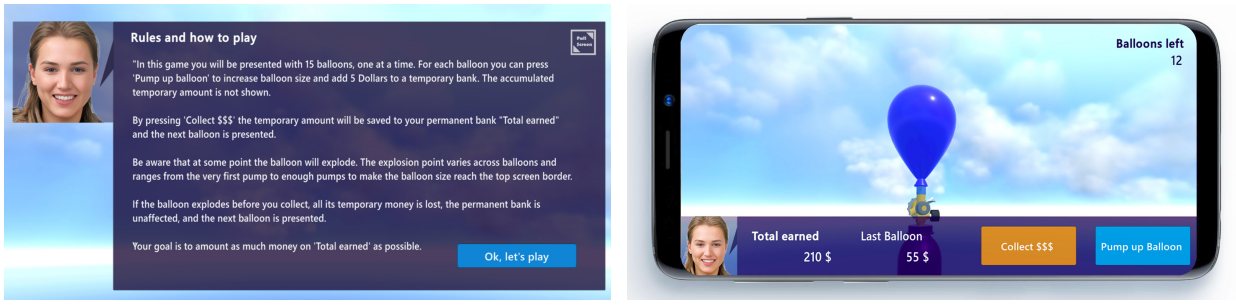

Figure 4. Mobile instruction app with female generated pedagogical agent visualisation (variant C)

As in this study is a field experiment utilising the Google Play store and applies a mobile app for assessing decisions risk-taking as opposed to the laboratory-experiment of Lejuene et al. (2002) two experiment settings were adapted to the context. First, the initial BART was applying balloons with different but precalculated average breaking (explosion) points. The balloons with higher breaking point averages were found as providing a more significant risk measure. 30 balloon trials, each with a breaking point average of 64 and a pump range between 1 and 128 were proven as a significant risk indicator. For this field experiment, the breaking point range was adapted to between 1 and 64, and the task featured 15 balloon trials which fits better to the limited time a casual app is played in a mobile scenario. Second, the breaking point average was fixed to 64 in the laboratory experiment and was the same for every test-run since participants only

took part in the experiment once. Conversely, the mobile app in the field was planned as an experience that can be played as often as pleased. Therefore, the breaking point average was randomised between 1 and 64 with a Knuth Shuffle algorithm (Knuth, 2014). For between-subject assessment, however, only the first trial identified by token was evaluated. Development tests with 20 app runs with 15 balloon trials each demonstrated a normal distribution of the runs average breaking point of $M = 31.65$, a standard error of $SE = 1.09$ and a 95% confidence interval of CI [29.36, 33.9].

2.3 Data collection and analysis

The mobile app was published internationally on the Google Play Store in English in May 2020 and was planned to run two months until the same day in July 2020. During this period, the app listing was supported by an international Google Ad campaign (0.5 €/day). The mobile store entry did promote the app as a research game. It was described in the store entry that the app would give some information if a participant is an above or below average risk-taker similar to the original BART experiment conduct. According to store analytics, the mobile research app was offered (i.e. displayed in the store) in 151 countries while being compatible with 13169 different android device models/systems. The apps' only requirement besides Android API 21+ was the permission to access the internet due to writing the collected data on a secure connection to the database.

A GDPR conform data privacy policy was made available to participants to read before proceeding to take part in the balloon exercise. Ethics approval was granted for the project by the NSD (Norwegian Centre for Research Data). Acquiring informed consent was provided by the introductory screen of the app. Participants were informed about all details of data collection/storage and had to give active consent before receiving a generated participation token and could proceed to the balloon task. Additionally, this introductory screen was asking some demographic information from participants including, gender, age-group, and country of origin. A log on the database was used to count saved results on each scenario and thus maintain equal participatory distribution between the three experimental conditions by providing the next participant with the variant having the lowest participation count. The metrics outlined in Figure 1 were only stored when participants advanced and finished the task, or in other words, when all 15 balloons were either collected or exploded. However, the analytics provided by the Google Play Store did provide additional data such as, for example, the total number of downloads/installations of the mobile app. The collected data after the two-month period was subsequently statistically analysed regarding the hypotheses and research objectives.

3 Results

3.1 Participation and field experiment metrics

During the two-month period, 1237 persons installed the mobile app on their mobile device (Table 1). However, from those only about 10% or 127 balloon task participations were recorded. Looking at the participation token that was stored on the device at first participation it became apparent that 28 records or 22% from recorded participations were second or more tries from the same participant which were subsequently excluded for the between-subject analysis. Further, from the remaining 99 entries, 25 were excluded that were no honest tries. There were mainly two combined characteristics that led to these exclusions. Either less than 20 pumps were registered during the whole task, or the pump/decision frequency was so fast that it could not be measured in any of the three thirds. In other words, the participants did not make a serious effort but just browsed through with tapping the pump button in "rapid-fire" style.

	N	Quota	A (no PA)	B (male PA)	C (female PA)
Installations from Google Play Store	1237	100 %	-	-	-
Users participating in the task	127	10 %	-	-	-
Unique participations	99	8 %	-	-	-
Valid participations	74	6 %	28	26	20

Table 1. Distribution of participation in the field experiment

Tests during development have shown that measuring very fast touch repetition series on smartphones that get below 40 milliseconds average was unreliable (e.g. skipping taps/measures). In the dataset, this was seen by some participants having non-measurable/undefined entries in the mean reflection calculation but still reasonable (>20) pump efforts and a plausible test score. Those entries were kept, but the average measure was set from undefined to 0. Table 2 shows the valid participations by age-group and gender.

	Age-group							Gender		
	16-19	20-29	30-39	40-49	50-59	60-69	70+	female	male	other
A	12	9	5	0	2	0	0	13	13	2
B	7	6	5	2	2	1	3	15	9	2
C	7	7	1	2	3	0	0	9	8	3

Table 2. Distribution of valid participations by age-group and gender

3.2 BART risk constructs

The collected data of valid participation entries ($N = 74$) was not normally distributed according to plots and Shapiro-Wilk tests (1965). Thus, non-parametric Kruskal-Wallis (1952) statistical analysis for independent samples with following pairwise Dunn-Bonferroni (1961) testing of the means was conducted for inferential analysis ($\alpha = 0.05$). While the risk constructs mean pumps before collection, $H(2) = 1.90$, $p = .388$ and total explosions showed no differences $H(2) = 0.95$, $p = .623$, the mean of pumps before explosions was significantly different between the groups $H(2) = 6.03$, $p = .049$. The pairwise evaluation showed that the difference was found in comparing average pumps before explosions in participants having the generated male instructor and persons who were instructed by the generated female instructor persona ($z = 2.41$, $p = .049$). Calculating the effect size revealed a medium effect in this difference of $r = .35$ (Cohen, 1988; Field, 2009).

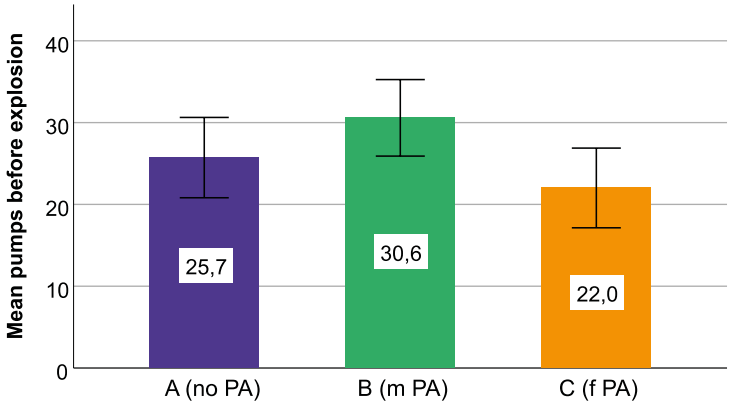

Figure 5. Mean pumps before exploding balloons displaying a significant mean difference between B and C [Error bars +/- 2 SE]

Figure 5 illustrates that participants who had been presented the male generated PA were reaching on average a 39% higher rate of pumps before explosions ($M = 30.6$) than their peers that were instructed with a female generated persona ($M = 22.0$). Notably, participants in group A that were presented no PA in the instructions were not showing significant differences regarding pumps before explosions to both other groups.

3.3 Reflection time factors

When turning to the reflection time statistical analysis of the time spent between pumps was displaying the suspected trend. While in the first third (i.e. the first five balloons) the pumps were executed slower, the reflection timespan decreased on average over the second to the last third (Table 3).

	A (no PA)		B (male PA)		C (female PA)	
	M	*SE*	*M*	*SE*	*M*	*SE*
First third of pump decisions	139*	35	177	38	264*	46
Second third of pump decisions	124	28	204	40	241	56
Third third of pump decisions	114	28	99	36	147	33
Total	196	27	193	29	258	43

*Table 3. Mean reflection time between pumping decisions in milliseconds; *indicating a significant mean difference between A and C*

Hypothesis testing further revealed a significant difference between the A, B and C groups in average reflection time within the first third of the instructed balloon exercise, $H(2) = 6.175$, $p = .046$. The subsequent second $H(2) = 4.056$, $p = .132$, and third, $H(2) = 2.755$, $p = .252$ parts of the participants balloon trials were not displaying significant differences. Pair comparisons revealed that within the first third of the exercise participants with the generated female PA were deciding/reflecting significantly longer before pumping the balloon a step further compared to the group without a PA ($z = -2.47$, $p = .041$, $r = .36$) (Figure 6; Table 3).

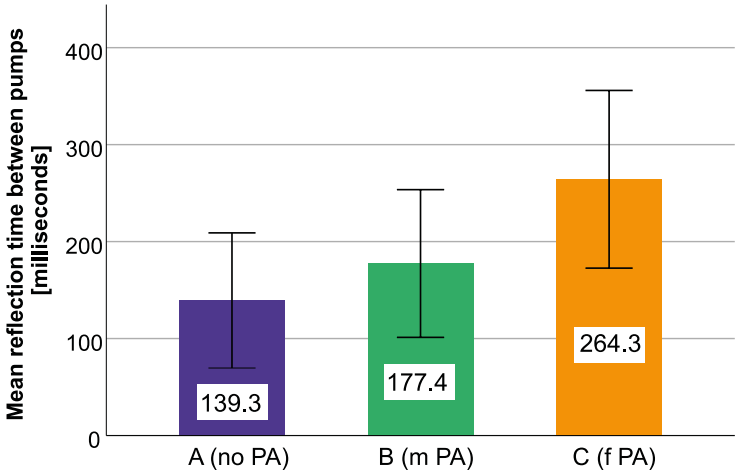

Figure 6. Mean reflection time between pumps displaying a significant mean difference between A and C [Error bars +/- 2 SE]

Statistical evaluation disclosed significant differences between the three experimental conditions. Thus, the analysis suggests rejection of H_{0A} considering the BART risk constructs and H_{0B} regarding reflection time between pumping decisions.

4 Discussion

4.1 Interpretation of the results for the design of instructions

When interpreting the analytical results regarding the first (i) research objective to investigate *the influence of generated instructor personas on taking chances,* the findings contribute to mainly three considerations.

First, instructions presented with a picture visualisation of an instructor are influencing risk tendency in decision-making. This influence is seen from the BART construct of average pumps until explosions of balloons which indicates a higher risk propensity when showing a generated male PA presenting instructions and progress as compared to a female instructor persona. However, a significant difference is only observed between the male/female setting (experiment condition) while the scenario with no PA shows as a middle ground in this regard. This is in line with previous studies that yielded no differences when investigating facilitation of learning or motivation by application of PA compared to a control group as reviewed by Heidig and Clarebout (2011). However, for distance or self-learning mobile learning concepts that are

implementing storylines and quests with characters as a concept of engagement (Dickey, 2020; Göbel et al., 2009), it is suggested considering the indicated PA gender bias. This holds specifically true for personalised/situational assessing of learner progress through decision-based tools that would be presented to one group by a male and to another by a female PA. One way to maintain engagement through a captivating story but control for a risk-taking bias could be to separate motivational support from instructional information as suggested by Baylor (2009). Consequently, however, such a segmentation impacts the story flow of a resulting story learning experience or quest (Jost, 2020). When looking at balancing learner progress/assessment with narrative experience, it would thus be required to consider the assessment design and timing already in the starting conception phase of a mobile learning app. Therefore, a possible balanced narrative experience with instructions from a PA could be integrated with story-related cut-scenes yet based on neutral instructional design in scenes with evaluation mechanisms. Importantly, in the case of scientific investigation/evaluation with quest-based experiences, a further way to control for PA/instructor risk-taking bias would be character-neutral control scenarios either during mini-quests or pre-test. In fact, this resembles interviewer bias in interpersonal research, however, with the considerable advantage of immediate and feasible procedural adaptability which is related to the following second implication for instructional design.

Second, this measurable risk propensity difference is demonstrated by displaying a static image of a male or a female *generated PA* to instructions. While influences from gendered tutoring agents on, for example, performance in engineering education (Johnson et al., 2013) have been documented before, it is worth noting that in this case the personas were generated by artificial intelligence. *The generative media approach is thus proving an ability to generate male/female photorealistic personas that create a measurable impact on taking chances*. In terms of utilisation for personalised education/instructions, this enables nigh unlimited application scenarios. Aside from effects of a gendered PA to support STEM education regarding female students (Krämer et al., 2016) other researchers have found attractiveness of PA as helping persuasion (Khan & Sutcliffe, 2014). The application of such detailed and adaptable visualisation traits in instruction scenarios enables prior impossible pedagogical opportunities. In particular, in distance learning scenarios where personalisation is essential to establish and maintain a motivational context for students or adaptive game-based learning (Göbel & Mehm, 2013; Seel, 2012) where taking opportunities can be a pedagogical objective. Similarly, it can contribute to a profound investigation of decision-making factors with narrative quests (Jost, 2020) by algorithmically employing detailed and subtle changes of characters that would otherwise be unfeasible. Importantly, generating adaptable PA for instructions is thereby not limited to academia or scholastic education but translates to organisational and industrial instruction scenarios. When remembering that the creators of the generated persona portraits (generated.photos) are further developing the algorithm to video persona generation, personalised scenarios for inhouse safety training or other human-factors skills concerning risky behaviour become feasible.

Related to this is the *third* consideration for instructional design. Given the implications of reflection times in decision-making (Evans, 2007), *instruction videos or safety training can be imagined to be personalised with generated personas that provide a focusing effect* (i.e. system 2 decision-making) when working through the learning content or following motor learning instructions (i.e. assembly guidance). As the experimental mobile app has shown, a generated female instruction PA can induce a more cautious progression through decision steps. With this in mind, a generated female PA instruction agent could help focusing between making decisions on the next assembly steps and thus support a more reflective progression through an assembly task. Similarly, a course planned to be studied in a mobile context could be improved by a generated female PA that enhances time spent on reflection, for example on true/false statements related to the learning content, before taking the next levels.

4.2 Implications for creating analytical mobile store apps

Turning to the second research objective (ii) on *practical findings for the design of evaluative mobile apps distributed internationally on an app store*, four key implications can be identified from practical conduct.

First, with the QGF, it is easy to implement a mobile research app investigating a single research scenario. Translating a valid psychological assessment in a mobile app and publish it internationally to 151 countries on the Google Play Store was done in two days with no technical difficulties. Reliable assessment of parameters such as reflection time measures and collecting data with a secure connection to the database of

the researchers was effortless and functional with the provided scripting components for the Unity engine environment. Also, by using the free assets from the Unity asset store and the generated PA, the instructional balloon test trial did not require special visual design skills or funds. Statistical analysis provided by the Google Play Store showed that during the period of the field experiment, no system crashes or other errors occurred. All participation efforts were reliably recorded in the database of the researchers.

Second, informed consent required for collecting data repels most users to engage with the app. Before any planned investigation, it should be considered that about 5% of downloads/installations of the mobile app convert to valid participation for evaluation. In the presented field experiment, 90% of Google Play Store users that installed the mobile research app did not participate. Ultimately only 6% of entries were honest and valid efforts. It is essential to monitor this by a participation token that allows identification of devices which engage in multiple trials when aiming for between-subjects evaluation. In general, it was possible to identify multiple entries without the token but by mobile device type/system combination and timestamp of participation.

Third, for acquiring potential participants, it is necessary to advertise the mobile app with a promotion campaign. Due to the immense number of existing apps on the Google Play Store, a mobile app for research purposes can normally only achieve sufficient participation within a reasonable time through advertising. However, the budget can be appropriately tailored to the targeted sample size. For the presented two-month experiment, 0.5 Euro per day converted to averagely 20.6 installations per day. Yet, due to the outlined repulsive effect of the informed consent, this eventually converted to only 1.24 valid participations per day (6%). After stopping the campaign, the installation rate dropped to 0.56 per day.

Fourth, technical and sample selection/identification constraints must be considered. Although no errors or crashes occurred, technical limitations such as the reported constraints by registering all touch inputs at high repetition rates can potentially be a source of unreliable measures. It is thus imperative to run repeated pre-experimental test trials with different mobile devices to reveal possible device limitations. Additionally, supplemental values from smartphone sensors could support the reliability of assessment by monitoring, for example, device orientation and movements to identify accidental inputs. In a field experiment on an international mobile app, store influences and insight on the study sample are obviously limited. However, with the mentioned ad campaign function, the Google Play Store provides country-specific advertising and has further configuration options as well as analytical options regarding sample origin. Nevertheless, it would be sensible to store country-specific data in the researchers' own database for processing in compliance with GDPR. At the same time, participants have shown to not truthfully self-report demographic data. For example, 28% of participants were stating Austria as a country of origin in the introductory informed consent form. However, according to Google Store metrics, not a single person located in Austria installed the app. The reason for this is likely that participants simply selected the first entry from the alphabetically sorted list instead of reporting truthfully. Unobtrusive assessment from device specifics/system should, therefore, be preferred to self-reported demographic data.

4.3 Limitations of the study and future research direction

The results provided valuable insight into the influences generated instructor personas might have on taking risks during instructive tasks. Nonetheless, as a field experiment with limited control over the study sample and how people are engaging with the exercise, there are some *limitations* to regard.

First, the Google Play Store was used to distribute the mobile instruction app internationally to 151 countries. Cultural differences were not considered. The experimental conditions were two PA of white ethnicity, and the interface language was English only. Still, the app was installed and played several times by people in Asia (e.g. India and Pakistan) according to Google Play Store analysis. In addition, when internationally advertising the app in all countries, there is no control over where and in which context precisely the app will be advertised by Google.

Second, the context in which the risk task exercise was played was not assessed. Being a mobile app, likely, the context of play differed substantially over the participants. As in other field experiments, contextual influences may have contributed to the group differences and influenced the internal validity. Possible assessment of mobile device parameters or exact location of participants could have helped in this regard but are also a privacy concern.

Third, technical constraints are limiting the precision of measures to some degree. The outlined limits in accessing fast repeated touch inputs could have some influence on the results. They naturally affected all groups; however, the accuracy of statistical inference thereby depends on the sample size, which was reduced by the factors outlined in section 3.1.

Lastly, as the research app was planned to be installed and played from a store as a casual mobile application, the BART maximum balloon trials and maximum pump trials have been adapted to a smaller range, and the breaking point was not set to a fixed number for each test run. Although the random algorithm was tested and showed the expected distribution and reliable mean breaking point (see 2.2), the results must be seen with these field test adaptions in mind.

Future research approaches should thus investigate the influences of generated PA further in more controlled laboratory settings and consider additional details in the relationship on decision-making. For example, eye-tracking studies during instruction/decision tasks could give insight on how and when people are referring to the pictured PA. Moreover, the duration of fixations could inform about features of attraction representing PA risk-taking influence while also a comparison to more abstract illustrations could inform instructional design regarding on how to control for influences in analytical scenarios.

5 Conclusion

The reported results from the field experiment on generated PA have provided insights on their effects on risk-taking in connection with instructional design and decision-making. The risk constructs of the balloon analogue risk task and reflection measures were assessed with an internationally distributed mobile exercise. Creation, distribution, and secure data collection of the mobile app was supported by a research quest framework that allowed comfortable configuration of the planned research design and publication to the Google Play Store.

The first, analytical research objective of the two-month field experiment addressed how photorealistic PA portraits generated by an artificial intelligence algorithm impact taking chances when applied in task instructions. Between the three groups where one [A] had no illustration of a PA, one [B] had a male, and another [C] a female instructor picture displayed, it was found that the visualisations influenced risk propensity. Taking chances regarding how far a balloon would be pumped before explosion differed significantly between the scenarios with people instructed by a male generated PA taking higher chances than their peers with a female generated PA. Complementary to this result, it was found that persons with the generated female PA did reflect longer between pump steps, or in other words, proceeded with more caution than persons that had no instructor visualisation.

The second, research objective on the practical implications for conducting field experiment research with an app on a mobile store revealed several considerations for designing evaluative mobile apps. The conversion rate (i.e. persons that actually engage in the task after installation) was shown as only about 10% when informed consent is implemented before allowing to participate according to GDPR. This fact should be included in the planning of an assessment specifically for research-oriented mobile apps. Moreover, objective metrics should be controlled with complementary measures (e.g. from sensors) and pre-tested regarding technical constraints of touch devices. Self-reported personal demographic data was highly unreliable in this field experiment with almost a third of actual participants not answering truthfully. Relatedly, a quarter of participants that were engaging in the actual exercise did not make an honest effort. Another factor that should be controlled when designing an evaluative mobile app with, for example, non-person related participation tokens and interaction metrics on the mobile device. Future research approaches should thus complement and extend the findings of this field experiment in more controlled settings to further learn about the indicated influences of generated instructor personas in instruction and learning scenarios.

ACKNOWLEDGEMENT

This research was supported by the Research Council of Norway (Norges Forskningsråd) by funding the IKTPLUSS project ALerT, #270969.

REFERENCES

Atorf, D., Kannegieser, E., & Roller, W. (2019). Study on Enhancing Learnability of a Serious Game by Implementing a Pedagogical Agent. *International Conference on Games and Learning Alliance*, 158–168.

Baylor, A. L. (2009). Promoting motivation with virtual agents and avatars: Role of visual presence and appearance. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3559–3565.

Bendou, A., Abrache, M.-A., & Cherkaoui, C. (2017). Contribution of pedagogical agents to motivate learners in online learning environments: The case of the PAOLE agent. *Proceedings of the Mediterranean Symposium on Smart City Applications*, 344–356.

Berger, E., Sæthre, T. H., & Divitini, M. (2019). PrivaCity—A Chatbot Game to Raise Privacy Awareness Among Teenagers. *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*, 293–304.

Biria, R., & Bahadoran, A. (2015). Exploring the role of risk-taking propensity and gender differences in EFL students' multiple-choice test performance. *Canadian Journal of Basic and Applied Sciences*, *3*(05), 144–154.

Cazzell, M., Li, L., Lin, Z.-J., Patel, S. J., & Liu, H. (2012). Comparison of neural correlates of risk decision making between genders: An exploratory fNIRS study of the Balloon Analogue Risk Task (BART). *Neuroimage*, *62*(3), 1896–1911.

Chu, H.-C., Hwang, G.-J., Tsai, C.-C., & Tseng, J. C. (2010). A two-tier test approach to developing location-aware mobile learning systems for natural science courses. *Computers & Education*, *55*, 1618–1627.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 Revised edition). Taylor & Francis Inc.

De Laet, T., Vanderoost, J., Callens, R., & Vandewalle, J. (2015). How to remove the gender bias in multiple choice assessments in engineering? Experimental validation and theoretical analysis using prospect theory. *Proceedings of the 43rd Annual SEFI Conference*, 1–8.

Dickey, M. (2020). *"The Quest" Narrative for K12 Game-based Learning: A Case Study of Using "The Quest" as a Model for Game-based Learning Design for K12 Teachers.* 8.

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Espinosa, M. P., & Gardeazabal, J. (2005). Do students behave rationally in multiple-choice tests? Evidence from a field experiment. *Evidence from a Field Experiment (December 22, 2005)*.

Evans, J. St. B. T. (2007). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Field, A. P. (2009). *Discovering statistics using SPSS: (And sex, drugs and rock 'n' roll)* (3rd ed). SAGE Publications.

Fukunaga, R., Brown, J. W., & Bogg, T. (2012). Decision making in the Balloon Analogue Risk Task (BART): Anterior cingulate cortex signals loss aversion but not the infrequency of risky choices. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(3), 479–490.

Göbel, S., & Mehm, F. (2013). Personalized, adaptive digital educational games using narrative game-based learning objects. In *Serious Games and Virtual Worlds in Education, Professional Development, and Healthcare* (pp. 74–84). IGI Global.

Göbel, S., Mehm, F., Radke, S., & Steinmetz, R. (2009). 80days: Adaptive digital storytelling for digital educational games. *Proceedings of the 2nd International Workshop on Story-Telling and Educational Games (STEG'09)*, *498*(498).

Gulz, A., & Haake, M. (2006). Visual design of virtual pedagogical agents: Naturalism versus stylization in static appearance. *Proceedings of the 3rd International Design and Engagability Conference@ NordiChi 2006*.

Haake, M., & Gulz, A. (2008). Visual stereotypes and virtual pedagogical agents. *Journal of Educational Technology & Society*, *11*(4), 1–15.

Harrison, J. D., Young, J. M., Butow, P., Salkeld, G., & Solomon, M. J. (2005). Is it worth the risk? A systematic review of instruments that measure risk propensity for use in the health setting. *Social Science & Medicine*, *60*(6), 1385–1396. https://doi.org/10.1016/j.socscimed.2004.07.006

Heidig, S., & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review*, *6*(1), 27–54. https://doi.org/10.1016/j.edurev.2010.07.004

Hopko, D. R., Lejuez, C. W., Daughters, S. B., Aklin, W. M., Osborne, A., Simmons, B. L., & Strong, D. R. (2006). Construct validity of the balloon analogue risk task (BART): Relationship with MDMA use by inner-city drug users in residential treatment. *Journal of Psychopathology and Behavioral Assessment*, *28*(2), 95–101.

Johnson, A. M., Ozogul, G., Moreno, R., & Reisslein, M. (2013). Pedagogical agent signaling of multiple visual engineering representations: The case of the young female agent. *Journal of Engineering Education*, *102*(2), 319–337.

Jost, P. (2020). The Quest Game-Frame: Balancing Serious Games for Investigating Privacy Decisions. In *Proceedings of the 11th Scandinavian Conference on Information Systems (SCIS2020)* (pp. 1–17). Association for Information Systems (AIS). https://aisel.aisnet.org/scis2020/5/

Jost, P., Cobb, S., & Hämmerle, I. (2020). Reality-based interaction affecting mental workload in virtual reality mental arithmetic training. *Behaviour & Information Technology*, *39*(10), 1062–1078. https://doi.org/10.1080/0144929X.2019.1641228

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Khan, R. F., & Sutcliffe, A. (2014). Attractive agents are more persuasive. *International Journal of Human-Computer Interaction*, *30*(2), 142–150.

Klein, G. (2008). Naturalistic Decision Making. *Human Factors*, *50*(3), 456–460. https://doi.org/10.1518/001872008X288385

Knuth, D. E. (2014). *Art of computer programming, volume 2: Seminumerical algorithms*. Addison-Wesley Professional.

Krämer, N. C., Karacora, B., Lucas, G., Dehghani, M., Rüther, G., & Gratch, J. (2016). Closing the gender gap in STEM with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers & Education*, *99*, 1–13. https://doi.org/10.1016/j.compedu.2016.04.002

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Laureano-Cruces, A. L., Sánchez-Guerrero, L., Velasco-Santos, P., Mora-Torres, M., & Ramírez-Rodríguez, J. (2016). Design of Pedagogical agents: The learning styles, the emotions and the color. *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 421–431.

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2014). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task. *Journal of Behavioral Decision Making*, *27*(1), 20–36.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75.

Li, X., Pan, Y., Fang, Z., Lei, H., Zhang, X., Shi, H., Ma, N., Raine, P., Wetherill, R., & Kim, J. J. (2020). Test-retest reliability of brain responses to risk-taking during the balloon analogue risk task. *NeuroImage*, *209*, 116495.

Lieder, F., Callaway, F., Jain, Y. R., Krueger, P. M., Das, P., Gul, S., & Griffiths, T. L. (2019). A cognitive tutor for helping people overcome present bias. *The Fourth Multidisciplinary Conference on Reinforcement Learning and Decision Making.*

Liew, T. W., Zin, N. A. M., Sahari, N., & Tan, S.-M. (2016). The effects of a pedagogical agent's smiling expression on the learner's emotions and motivation in a virtual learning environment. *The International Review of Research in Open and Distributed Learning*, *17*(5).

Mabruri, H., Ahmadi, F., & Suminar, T. (2019). The Development of Science Mobile Learning Media to Improve Primary Students Learning Achievements. *Journal of Primary Education*, *8*(1), 108–116.

MacLean, R. R., Pincus, A. L., Smyth, J. M., Geier, C. F., & Wilson, S. J. (2018). Extending the Balloon Analogue Risk Task to Assess Naturalistic Risk Taking via a Mobile Platform. *Journal of Psychopathology and Behavioral Assessment*, *40*(1), 107–116. https://doi.org/10.1007/s10862-017-9628-4

Mayer, R. E., & Moreno, R. (1998). A cognitive theory of multimedia learning: Implications for design principles. *Journal of Educational Psychology*, *91*(2), 358–368.

Miller, H. B., & Cuevas, J. A. (2017). Mobile learning and its effects on academic achievement and student motivation in middle grades students. *International Journal for the Scholarship of Technology Enhanced Learning*, *1*(2), 91–110.

Oyelere, S. S., Suhonen, J., Wajiga, G. M., & Sutinen, E. (2018). Design, development, and evaluation of a mobile learning application for computing education. *Education and Information Technologies*, *23*(1), 467–495. https://doi.org/10.1007/s10639-017-9613-2

Pimmer, C., Mateescu, M., & Gröhbiel, U. (2016). Mobile and ubiquitous learning in higher education settings. A systematic review of empirical studies. *Computers in Human Behavior*, *63*, 490e501.

Rakhmawati, L., & Firdha, A. (2018). The use of mobile learning application to the fundament of digital electronics course. *IOP Conference Series: Materials Science and Engineering*, *296*, 012015. https://doi.org/10.1088/1757-899X/296/1/012015

Rao, L.-L., Zhou, Y., Zheng, D., Yang, L.-Q., & Li, S. (2018). Genetic contribution to variation in risk taking: A functional MRI twin study of the balloon analogue risk task. *Psychological Science*, *29*(10), 1679–1691.

Schroeder, N. L. (2017). The influence of a pedagogical agent on learners' cognitive load. *Journal of Educational Technology & Society*, *20*(4), 138–147.

Seel, N. M. (Ed.). (2012). *Encyclopedia of the sciences of learning*. Springer.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Sweller, J. (2016). Working memory, long-term memory, and instructional design. *Journal of Applied Research in Memory and Cognition*, *5*(4), 360–367.

Veletsianos, G. (2010). Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education*, *55*(2), 576–585.

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*.

Walker, D. M., & Thompson, J. S. (2001). *A note on multiple choice exams, with respect to students' risk preference and confidence*.

Wang, I., Smith, J., & Ruiz, J. (2019). Exploring Virtual Agents for Augmented Reality. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300511

Yang, Z., & Tackie, M. (2016). Risk Preference and Student Behavior on Multiple-Choice Exams. *Economics Bulletin*, *36*(1), 58–67.