FoodChoices(Q): Exploring the design of a serious game proxy for Likert-style survey questionnaires

FoodChoices(Q)

Exploring the design of a serious game proxy for Likert-style survey questionnaires

KAMIL ROGODA
PIOTR DANISZEWSKI
KAMIL FLOROWSKI
RISHAB MATHUR
KOUROSH AMOUZGAR
JAMES MACKENZIE
KIM SAUVÉ
ABHIJIT KARNIK

Lancaster University, United Kingdom*

Likert-style questionnaires and surveys are commonly used tools for research. To alleviate survey-fatigue, researchers have explored gamification routes to increase engagement and lower drop-outs. However, these attempts still rely on direct use of questionnaire text and focus on creating engagement around the actual activity and do not fully alleviate the challenges of filling survey. In this paper, we explore an alternative approach involving the use of a serious game to capture user responses through in-game activities rather than direct questions. We chose the Food Choice Questionnaire (FCQ) and explored the design challenges of creating a serious game which deploys a sub-sample of the FCQ questions as four mini-game activities. The player actions and decisions are used to compute a result which is compared with their FCQ responses. We demonstrate the method to evaluate the equivalence of game results to the questionnaire responses. We discuss how future serious games can be designed and evaluated to generate similar outcomes while avoiding potential pitfalls through design and analysis.

CCS CONCEPTS • Human-centered computing~User studies • Applied computing~Computer games

Additional Keywords and Phrases: serious games, gamification, Food Choice Questionnaire, survey fatigue

ACM Reference Format:

Kamil Rogoda, Piotr Daniszewski, Kamil Florowski, Rishab Mathur, Kourosh Amouzgar, James Mackenzie, Kim Sauvé and Abhijit Karnik. 2022. FoodChoices(Q): Exploring the design of a serious game proxy for Likert-style survey questionnaires. Proc. ACM Hum.-Comput. Interact, 6, CHI PLAY, Article 236 (October 2022), 15 pages, https://doi.org/10.1145/ 3549499

1 INTRODUCTION

Likert-style questionnaires and surveys are an important tool in data-analysis driven research in academia and industry. These questionnaires are used to a variety of parameters like behavioural traits [33], cognitive load [16], perception [5] and even preferences [31]. However, surveys and questionnaires also present usage challenges. Survey fatigue is commonly reported [28] among participants due to exposure to large number of surveys from different sources. Respondents also tend to leave fields blank or simply not return them [6]. Researchers have explored various methods to overcome these

* This is the author-generated manuscript copy prior to publication. The definitive version of the article can be found on the ACM website via: https://doi.org/10.1145/ 3549499

challenges by modifying the scale interface [22] and through gamification [11, 15] while avoiding introduction of a strong bias sources into the survey results. Gamification of surveys leverage the increase in perceived fun which results in increased willingness of respondents to complete and recommend the survey. With gamification, the questions of the survey are not modified or eliminated, and the gamification focuses on the peripheral aspects of presentation and flow.

This paper contributes a novel approach to addressing the challenges of survey fatigue through the lens of serious games. A key difference between our work and prior survey gamification work is that our serious game creates proxy activities which are equivalent to the target survey questions but do not use the actual question wording anywhere within the game. We discuss the key elements of the design process necessary to convert a questionnaire into a game and then demonstrate the equivalence of the game results to questionnaire responses.

2 RELATED WORK

2.1 Gamification of Surveys

Likert-style surveys are a common class of instrument used in acquisition of self-reported data from participants. Standardized Likert-style survey instruments, which are internally validated through a design and evaluation process, are regularly used in research. HCI researchers have explored mechanisms to investigate and mitigate the disruptive effects of administration of a questionnaire [2, 29] in VR. The recommendation is to include the survey within the flow of the task and deliver it through the same medium (e.g., VR) as the task being evaluated. Similarly, other researchers also underline the need to make the questionnaires more pleasant for respondents and achieve a highly reliable response rate. For example, Harms et al. describes a design process to gamify a survey related to sports [15]. Using the MDA framework [19], they set the questions in a sports competition-like environment. They identify that the time taken to complete a gamified survey increases but this is balanced by the fun experience generated by the gamification. Other researchers have looked simplifying the input process by removing the need for precision pointing for responses [22]. Dorcec et al. [11] and Frommel et al. [14] demonstrate that including the survey questions into mini-games that are part of the flow is also feasible.

The reviewed literature exploring gamification of surveys tackles different aspects of the data-collection process aimed at making the survey instrument less disruptive and less onerous [21]. The contrast lies in the interpretation of the time taken to complete the survey. Harms et al. and Frommel et al. observe that it takes longer to complete the gamified survey, while Alexandrovsky et al. and Putze et al. focus on a comparable time of completion. Alexandrovsky et al. observed that *"participants perceived filling out the questionnaires in VR as faster"* in contrast to statistically insignificant but overall increased time for survey completion. In the context of gamification, and games in general, the argument favors fun and pleasurable experience over raw time-based completion performance.

The other common thread within this literature is that the text of the survey instrument is delivered as-is. Frommel et al. match the lane count in their car-driving game to the levels of the Likert-scale in use, but the questions (e.g., *"Currently I feel joy"*) are presented unchanged. Harms et al. and Drocec et al. use sports and game settings but once again the question text is never altered. In each case, the gamification is used to set the context and create engagement around the survey filling experience. We thus identify a gap in the literature exploring the use of a proxy activity in lieu of the actual survey instrument set in the context of a serious game. This approach of using a proxy serious game instead of the actual survey instrument seems unexplored and non-trivial.

2.2 Gamification versus Serious Games

The first challenge for serious game proxies is to distinguish clearly between gamification and serious games. Gamification is defined as the application of game design elements in non-game contexts [10]. Gamification uses game design elements like achievements, badges, score-boards and challenges to enhance experience instead of creating full games [9]. This is distinct from the concept of serious games which were first described by Abt [1]. The serious game is a completely functional game but designed with a learning-based *"serious"* objective in mind. A user can engage with the game-play without explicit knowledge about the serious objective or even awareness that the game-play is achieving the learning objective. Also, the term *"serious"* does not imply the game lacks fun, but rather that fun is not the only objective. Serious games have been designed to cover various aspects of learning-based objectives like educating the player [12, 26], collecting behavioural information [18] and even evaluate learning [25]. Thus the motivation of using a serious game proxy that replaces a survey instrument is to provide a better pathway for researchers to acquire self-reported data from participants. While Bailey et al. [3] and consequently Harteveld et al. [17] refer to this as '*hard gamification*' in the context of survey gamification, we align to Deterding et al.'s [9, 10] approach of differentiating gamification from serious games and exploring the design of a serious game proxy.

2.3 Food Choice Questionnaire

To explore the design process of creating a serious game as a proxy for a questionnaire, we chose Steptoe et. al.'s Food Choice Questionnaire [31]. It is a popular[1] means for studying people's dietary choices. The FCQ is a 36-question Likert-style survey that measures several factors or categories that influence food choice. These categories are health (He), sensory appeal (SA), price (Pr), convenience (Co), mood, natural content, weight control, familiarity, and ethical concerns [8]. The responses are on a 4-point scale going from "Not important at all" to "Very important". It is known for its wide adoption, being used for research into food selection choices around the globe [13] and being heavily tested and validated [27, 31]. It has been revised many times and tailored to different research scenarios that involved exploring food habits and choices. The FCQ has been applied to identify consumer segments [13] and even investigate sustainability within food choices [32]. Despite its popularity, we didn't find any research about making the questionnaire more appealing to the participants and keeping them engaged while answering the questions. Thus, the FCQ is a suitable candidate for the exploration of using a serious game proxy instead of the questionnaire being administered in a digital format. We intentionally did not rely on a bespoke questionnaire or survey. This would introduce an additional validity confound about the actual questionnaire itself and make the process of investigating equivalence unreliable.

3 DESIGN

Our aim is to explore the design space of creating a serious game proxy for an existing survey instrument. We developed FoodChoices(Q) as a serious game proxy for the FCQ. The design process is briefly described to cover the essential steps.

3.1 Implicit goals

The goal of the FoodChoices(Q) game is to allow the user to play the game as a series of inter-connected actions fitting within the flow of the game. Internally, the game collects data that can be used to compute result(s) that are equivalent to the responses that would be given by the user if they answered the FCQ questions instead. In this regard, the game must satisfy one of the three goals of questionnaire design, viz.: questions used are accurate to the overall research that is

[1] estimated citation count exceeding 2200 at time of writing

intended [4, 20]. The use of an existing and established survey questionnaire (the FCQ) eliminates the risk of the source questions being relevant. However, we still need to validate the equivalence of the proxy game's results with the survey results. Hence, we need to design the evaluation such that it tests the equivalence of game results to the survey responses. As a game activity, the reduction in the task completion time is not an implicit target. The time required to complete the survey for the proxy is the game-play time. The main design goal is to replace the onerous activity of survey filling with one that is potentially more enjoyable and fun.

3.2 Design considerations

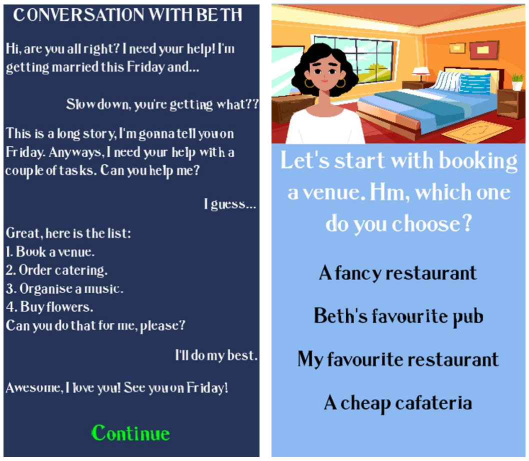
Likert-style survey instruments consist of questions which can be independent or grouped into categories. The designers of the instruments describe how to compute a result using the instrument and involves scoring the questions in a pre-defined and validated manner. The design challenge is to decide how the in-game activities and the subsequently computed scores map to elements of the survey instrument. The open question here is if the in-game activities map to a single question or to a category consisting of several questions. The game proxy also needs to satisfy the requirements of being a serious game. The design must address the implicit goal of the serious game by creating opportunities where the user's game-play results in generation of data that can be used instead of self-reported data in a survey instrument. It also must address the requirements of presenting a game experience with meaningful play [30].

3.2.1 FCQ Sub-set

FCQ is a 36-question questionnaire which are grouped into categories. Research has established that the most important criteria categories [31] for determining a person's food choices are health (He), price (Pr), sensory appeal (SA) and convenience (Co). These categories are covered through 18 questions. Since our aim is to explore the design process of creating the proxy, we focus on this subset of questions covering the four categories within the available time and practical constraints. To address the question of proxy-design for categories versus individual questions, we chose both alternatives and design game elements around them. The design considers three of the categories (SA, Pr, Co) directly. The aim is to generate a score that can be compared directly to the category response scores. For the fourth category, health (He), we chose two questions Q22 and Q27 to generate the in-game results for comparison with the question responses. We considered the rationale for choosing to map a category directly to the in-game activity and grounded it in the methodology used to design a survey instrument. During the design of a survey instrument, the designers identify which questions are able probe the category satisfactorily. In practical use, category questions are Likert-Scale questions which roll-up to the category. The standard practice for analysis of Likert-Scale questions is to average the question responses per category for each user and then use this average value for each user in further analysis of the categories. For the price, sensory appeal, and convenience categories, we explore the use of a proxy to extract the category score directly instead of relying on proxies for individual questions. An arguably flawed but accepted practice (at least in HCI research) is to compare the average (across user responses for the same question) of Likert-style questions. The health category questions allow us to explore the use of a proxy for a single question in this manner.

3.2.2 Game setting

Instead of designing a game from ground-up, we chose to use an existing game as a template. The challenge is to break the game into discrete stages or steps providing a clear demarcation of activities. The stages can then be used to compute results that map to the questionnaire responses. This requirement is met well by games that use embedded mini-games within a main-story arc. We designed the main game to be an interactive fiction game inspired by narrative expositionbased games like Choices[2]. The mini-games require choices to be made to move the game forward. To limit the design effort, the choices are considered as snap decisions which don't alter the story-telling but still relevant to the flow. The game storyline follows a wedding planning scenario, and the player is presented with the mini-games involving choices related to food as a part of the narrative. The story is presented in the format of conversations and choices to be made (see Figure 1). The mini-games designed as FCQ proxies did not show the actual question-text from the FCQ at any point in the game. The game also included additional choice-based mini-games intended to move the narrative forward but did not have relevance to the FCQ itself.



CONVERSATION WITH BETH

Hi, are you all right? I need your help! I'm getting married this Friday and...

Slow down, you're getting what??

This is a long story, I'm gonna tell you on Friday. Anyways, I need your help with a couple of tasks. Can you help me?

I guess...

Great, here is the list:
1. Book a venue.
2. Order catering.
3. Organise a music.
4. Buy flowers.
Can you do that for me, please?

I'll do my best.

Awesome, I love you! See you on Friday!

Let's start with booking a venue. Hm, which one do you choose?

Figure 1: Screenshots of the main game plot

3.2.3 Mini-games

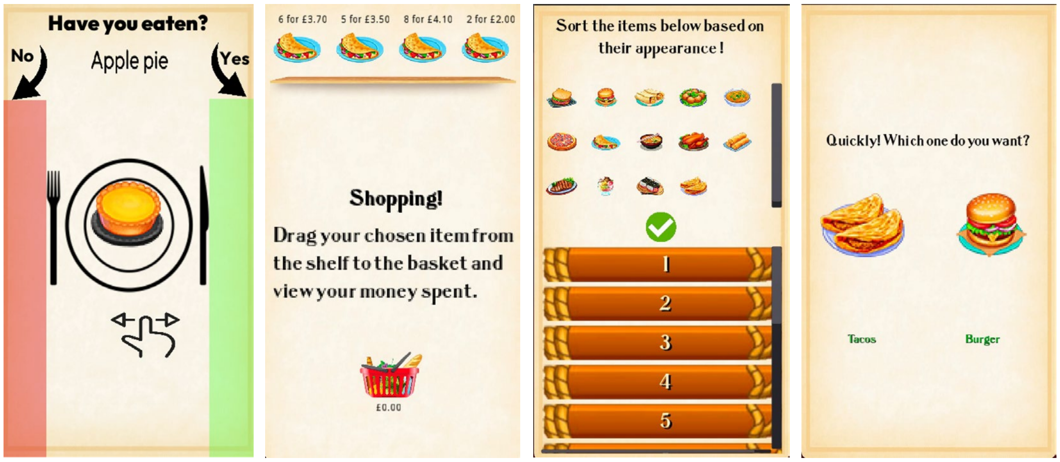
There are five mini-games within FoodChoices(Q) which are relevant to the FCQ. The first mini-game establishes a selection of food items (or dishes) the player will see in the subsequent mini-games. The mini-game is essential due to the diversity of dietary preferences and variation in food familiarity within the target population (UK-based). The aim of the mini-game is to identify a set of dishes the player has eaten or eats on a regular basis. The initial database is built using 35 of the most commonly eaten dishes in UK. The player selects as many of these as they want with a minimum requirement of choosing five. The choices are made using a swipe left or right mechanic to reject or select the dish (see Figure 2). Once the player completes their selection, the dishes shown in the remainder of the game are only chosen from the selected list.

For the Price category of the FCQ, the objective of the mini-game is to probe the user's preference towards value for money in the context of food. This objective is same as the Price category of FCQ. The mini-game places the player in a supermarket shopping setting. The user is presented with four choices of meal deals and asked to select one (see Figure 2). The player uses a drag and drop mechanic for selecting the preferred deal. This process of showing different deals is repeated thrice. Each deal has a score assigned to it based on actual value for money (from lowest unit price to highest unit price). The proxy-score for the Price game is thus based on the deals chosen by the player and helps identify the user's preference towards value for money based on the meal deal they choose in each iteration.

[2] Choices: Stories You Play https://play.google.com/store/apps/details?id=com.pixelberrystudios.choices

The mini-game associated with Sensory Appeal category uses a two-step process. To capture the subjectiveness of the sensory appeal of food items, the player is asked to rank food items based on their appearance or looks. In the next step, the player is presented with two options from the ranked list and asked to choose one (see Figure 2). The scoring is based on the relative appeal of the chosen items from what was displayed. The number of iterations depends on number of dishes selected in first mini-game and each dish is shown at least once.

The mini-game associated with Convenience category uses a setting that involves food preparation. We attempted to measure convenience as amount of effort the player will put into preparing a dish. Players are shown four types of toast, with increasing sophistication and thus increasing effort to make them. The player interactively "rubs" the pan to mimic the cooking effort and this fills up a progress-bar (see Figure 3). If the user stops, the progress is saved only up to the most recently attained level and to get to the next level, they need to start interacting with the pan again. When the player is satisfied with the dish, they can stop rubbing the pan and click finish. The completion state is used as the metric to measure the preference of convenience over effort.



Have you eaten?

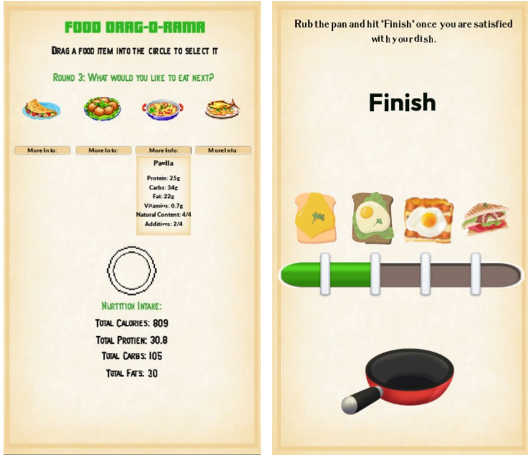
Shopping!

Drag your chosen item from the shelf to the basket and view your money spent.

Sort the items below based on their appearance!

Quickly! Which one do you want?

Figure 2: Mini-games: (L to R) first mini-game, price mini-game, sensory appeal mini-game in two steps.



FOOD DRAG-O-RAMA

DRAG A FOOD ITEM INTO THE CIRCLE TO SELECT IT

ROUND 3: WHAT WOULD YOU LIKE TO EAT NEXT?

Paella

Protein: 25g
Carbs: 34g
Fat: 22g
Vitamins: 0.7g
Natural Content: 4/4
Additives: 2/4

NURTITION INTAKE:

TOTAL CALORIES: 809
TOTAL PROTIEN: 30.8
TOTAL CARBS: 105
TOTAL FATS: 30

Rub the pan and hit 'Finish' once you are satisfied with your dish.

Figure 3: Mini-games: Health mini-game (left), Convenience mini-game (right)

The final mini-game doesn't focus on a category of the FCQ, but rather attempts to acquire results equivalent to individual questions. We chose two questions from the Health category viz., Q22 *("...contains lots of vitamins and minerals")* and Q27 (*"...is high in protein"*). The mini-game uses a sit-down meal setting situation and the player chooses the dishes to eat. Before they choose the dish, they can examine the content (protein, fat, carbohydrates, vitamins, minerals, and additives). This is repeated four times. The cumulative nutritional intake is updated and shown each time a dish is chosen (see Figure 3). The relative nutritional value of the choice from within the options offered is used to compute the metric.

These mini-games were arrived at using brainstorming sessions. Our game design choices are neither definitive nor restrictive. They are meant to showcase the process of designing a proxy element. Astute game designers could replace these with entirely different alternatives if the aim of creating a probe that matches the survey question (or category) objective is satisfied.

4 IMPLEMENTATION

FoodChoices(Q) was implemented using the GODOT game engine. GODOT is a free-to-use and open-source game development engine [24]. It was chosen mostly due to its ease of use and potential for wider platform compatibility. For the study, the game was installed on two Android devices. The players' choices and additional game logs were stored and later retrieved from the devices. The FCQ, which was limited to 18 questions, was deployed using Qualtrics and once again accessed through the browser on the devices. The aesthetics of the game assets were deliberately kept minimal. The dishes were shown at a higher fidelity of representation. To overcome the lack of artistic skills, we used images of foods available in the game *"Cooking Express: Cooking Chef"*[3]. This provided the game with a consistent quality of presentation of the dishes.

5 EXPERIMENT

The challenge facing a designer of the serious game proxy is how to validate the proxy against the original survey instrument. The aim of the evaluation is to establish equivalence between the game results of each mini-game versus the questionnaire responses from the same player. The methodology chosen involves within-subjects design and the mode of delivery (game or questionnaire) is the independent factor. Many HCI studies rely on null-hypothesis significance testing (NHST) to demonstrate differences between two systems. In the case of the proxy, a significant result of such a test is not a desirable outcome as it indicates the proxy diverges from the survey instrument results (assuming all other considerations e.g., normality, sphericity etc. are satisfied). At the same time, a null or non-significant result is not adequate to prove equivalence since NHST can be used to either reject null-hypothesis or fail to reject null-hypothesis [34]. Thus, an additional statistical tool is needed to prove equivalence. We make use of TOST in our current analysis. However, as with NHST, if the data violates normality, alternative tools like Bayesian ANOVA should be considered.

5.1 Participants

Due to the ongoing COVID-19 pandemic restrictions, we recruited participants within our personal bubbles to ensure that we followed the current restrictions. We recruited 24 participants, where 15 female, 8 male (1 preferred to self-describe as 'kind'). Participants fell within the age ranges 18-24 (17), 25-34 (1), 35-44 (3), 45-54 (1), and 65+ (2). Majority of participants had some form of gaming experience, and all of them were familiar with using mobile devices, as well as with

[3] https://play.google.com/store/apps/details?id=com.gameicreate.mycafeexpressrestaurantchefcookinggame

text-based questionnaires. Each participant took part in the experiment individually and was supervised by a member of the research team throughout the whole process. The study was carried out after following standard procedure for ethical approval within the university.

5.2 Procedure

The study was carried out on an Android device. Each participants filled the FCQ hosted as a Qualtrics survey and played the game. The order of survey and game were counterbalanced between participants. Before the start of the study, the participant was presented with the research information sheet, a consent form, and instructions to familiarize themselves with the setup. The participants completed the first task (game or FCQ) and continued with the second task (FCQ or game) after a short break in between. To maintain the game-experience, the participants were allowed to complete the whole game to arrive at a proper in-game narrative ending resulting from the choices made by the players during the game. After both the tasks, a semi-structured interview was conducted to gather general feedback about the game. Participants could leave the study at any time. Each session lasted less than half an hour, and participants were not compensated for the participation as the study was voluntary.

5.3 Measures

5.3.1 Questionnaire Data

The participants filled in a digital version of the FCQ consisting of 18 questions associated with the 4 categories viz. Health (He), Convenience (Co), Sensory Appeal (SA) and Price (Pr). The responses to the questions are on a 4-point scale ranging from "Not important at all" to "Very important" which were coded to scale responses: [1,4].

5.3.2 Game Logs

The game was instrumented to log data through game logs for the mini-games and main-game actions. The result to be compared with the FCQ responses was computed for each mini-game as follows:

1. Price mini-game: There were three iterations of deal selection. Each deal was assigned with a rank from 1 to 4, where 1 meant the best value for money. Choosing the best deals implies caring about the value for money, we reversed the scale and then averaged the results of the iterations to arrive at the game result.
2. Sensory Appeal mini-game: The scoring activity was carried out in the second step of the mini-game (see 3.2.3: Mini-games). For each iteration, the participant's choice was scored based on the relative position of the two options within the ranked list produced in the first step. Higher ranked choices were scored 1 and lower ranked choices were scored 0. The game result used the average of all iterations to compute the quartile (e.g., if 80% choices were higher ranked, result would be top quartile i.e., 4) in which the average fell into.
3. Convenience mini-game: The final completion status of the progress bar was used to compute the score for this mini-game. This produced a ranking of 1 (least effort) to 4 (most effort). This ranking was reversed to produce the game result to match the convenience category of the FCQ as the more effort the players put, the less they cared about ease of preparation and thus convenience.
4. Health minigame: The food choices presented during each of the four iterations also provided reference values for protein and vitamins and minerals contained in the dishes. For each iteration, the four displayed options were ranked 1 (low amount) to 4 (high amount). The ranks of the chosen items from all iterations were averaged to arrive at the game results, one for protein and one for vitamins and minerals.

Additional quantitative metrics like task completion times and dishes selected were also recorded. We also noted the observations and comments made by the participants during the post-activity semi-structured interview.

6 RESULTS

The analysis is devised to investigate the equivalence between the FCQ responses and the results from the mini-games. To compute the category scores from the FCQ, the responses were coded as 1 to 4 (with 4 representing Most Important) and then averaged across the questions belonging to the same category. The mini-games similarly produced scores between 1 to 4 for the associated measures. We used SPSS v28 to carry out preliminary tests (ANOVA). We also used Jamovi v1.6 and TOSTER 0.3.3 for equivalence testing (TOST). The preliminary step towards equivalence testing was to reject alternative hypothesis (or fail to reject null hypothesis) of significant difference between the category responses and game results. We present the statistical results here but defer the implications to the Discussion section.

6.1 Preliminary Tests

We used repeated measures ANOVA with single factor (mode as questionnaire vs game) for the three categories and two individual questions as dependent measures. The results are tabulated in Table 1. The Convenience mini-game showed strong significant difference to the Convenience category responses of the FCQ ($F_{(23)}=46.4$, $p<0.001$). The Convenience category could thus not be evaluated further for equivalence. Visual analysis of the histogram showed presence of an independent factor affecting the game results (See Figure 4).

Table 1: Preliminary tests results (ANOVA)

Measure	μQ (Questionnaire mean)	μG (Game mean)	F-value	p-value
Convenience	3.125	1.833	46.423	<0.001**
Sensory Appeal	3.167	3.083	0.216	>0.05
Price	3.014	2.834	0.706	>0.05
Health – Protein	3.000	2.666	1.670	>0.05
Health – Vitamins and Minerals	2.92	2.489	4.538	<0.05*

[a] df=23, * indicates statistically significant differences.

Both Sensory Appeal and Price categories showed no significant differences to their equivalent mini-game responses. Thus, they were analysed further using equivalence tests. The individual questions Health Q27 for Protein and Health Q22 for Vitamins and minerals were also analysed. The game results were significantly different than questionnaire responses for Health Q22 (Vitamins and minerals) and thus not tested for equivalence. The data for Health Q27 (Protein) was retained for further analysis of equivalence due to a NS result for difference.

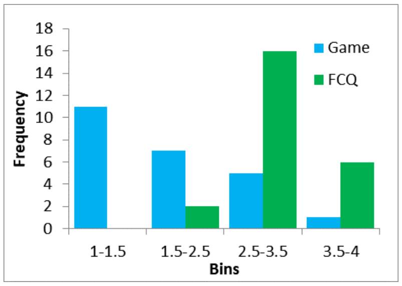


Figure 4: Histograms for Convenience category scores from Game (blue, series on left) and FCQ (green, series on right)

6.2 Equivalence Tests

The TOST procedure for testing equivalence among means [23] indicates that the responses to FCQ category Sensory Appeal were equivalent to the results of the Sensory Appeal mini-game: $t_{(24)} = -1.98$, $p<0.05$. The test used Cohen's d = 0.5, and computed raw bounds [-0.439, 0.439] compared against the 90% CI of the sample [-0.224, 0.39]. With Cohen's d set to 0.53, the TOST procedure for testing equivalences among means indicated that the responses to FCQ category Price were equivalent to the results of the Price mini-game: $t_{(24)} = -1.76$, $p<0.05$. The computed raw bounds were [-0.556, 0.556] and the 90% CI was [-0.187, 0.547]. For the individual question Health Q27, the TOST procedure only produced a significant result for Cohen's d >0.8. Normally, statistical analysis relies on pre-defined and fixed values of Cohen's d for effect sizes. Since we are demonstrating the process of establishing equivalence, we are not restricting the analysis to fixed values and rather show at what value of Cohen's d, significance could be considered to have emerged. We discuss the rationale for choice of a different Cohen's d-value for the Price mini-game in the Discussion section.

6.3 Other metrics

6.3.1 TCT

The task completion time (TCT) metric was collected with a prior understanding that the game TCT would be trivially higher than the FCQ TCT. This is in line with the expectations as established by previous research on gamification [11, 15]. Out of 24 records, two FCQ TCTs were discarded due to human error and 1 game TCT was lost due to program malfunction. The average FCQ TCT was 2:48 minutes while the average game TCT was 11:13 minutes. No other patterns of interest were observed with the TCT.

6.3.2 Items chosen in first mini-game

The number of items chosen in the first minigame ranged between 12 and 33 with an average of 24.17 items. The number of items increased the variety of choices available in subsequent mini-games but also increase the effort of ordering in the Sensory Appeal mini-game.

6.3.3 Participants Feedback

Participant feedback was gathered through a semi-structured interview. The participants found the idea of creating a game based on the questionnaire really appealing. Most participants found the story engaging and stated that it made the game interesting. They also found the mini-games were fun. Some participants noted that the five mini-games were based on snap-decisions and indicated preference of these having more impact on the overall narrative. Another observation was the impact of number of items chosen in first mini-game for the sensory appeal mini-game's first step. It was also observed that for the Convenience mini-game, participants often ignored the fact that they can finish the game earlier and were focused on filling the progress bar or aiming for the highest reward (most appealing toast recipe) than the difficulty of preparation. Overall, the participants enjoyed the game and were curious about the findings of the study.

7 DISCUSSION

The results produce a set of interesting outcomes and observations in the context of designing a serious game proxy for questionnaires.

7.1 Convenience category results

The histogram for the game results showed strong skew towards convenience being least important. This was obviously contrary to common understanding and not matching the responses to the FCQ. Post-hoc game analysis suggested a surprising confound. Since the overall effort in the mini-game was less, participants chose to engage with the game with a strong sense of competition (agon against the artificial player) and attempted to reach the highest possible result. The choice of the game mechanic needs to balance the implicit objective versus the explicit effort required to attain it. This is comparable to Harms et al. [15] where one mini-game mimics a javelin throw activity as a response to the answer. We identify a risk that the user response is strongly skewed by the player mindset of winning and reward attainment rather than providing an usable response. At the same time, if the effort is too high, the results may be skewed to the lower end of the response scale due to player abandonment of a difficult task. It also ascertains our expectation that the design of a serious game proxy (or even a gamified survey) is a non-trivial activity.

7.2 Equivalence tests

The result of the TOST procedure allowed us to demonstrate equivalence of the results from the game to that of the FCQ results for the Sensory Appeal and Price category. The equivalence should be considered in context of the questionnaire being proxied. The choice of Cohen's d value at 0.5 is for a medium effect size and generally accepted within psychology studies [7]. If the chosen questionnaire was a behavioural one rooted in psychology, this would be adequate. However, the acceptable value of Cohen's d for medium effect size is different for other studies (e.g. educational research [7] and even HCI [34]). Our aim of varying the medium effect size parameter is to demonstrate that the choice is subjective to the survey instrument being proxied [34]. In the analysis for the Price game, we chose a Cohen's d of 0.52 and arrived at an equivalence result. The game designer will have to correctly identify the effect size based on: a. accepted norms in the field of study or b. accepted standard values which are then used to derive Cohen's d (e.g., IQ or BMI). The Price game required a certain level of numerical skills to make the choice. Our convenience sampling derived from a university-connected population avoids pitfalls due to lower numerical literacy. However, this may not be the case in real-world scenarios and should thus be factored in within design and statistical analysis.

7.3 Individual questions for health category

The mini-game proxy for Health Q27 did not produce an equivalent result despite not being significantly different. The value of Cohen’s d for equivalence far exceeded the acceptable ranges unlike the Price game. For Health Q22, the game results were significantly different and thus equivalence also could not be established. There are two possible interpretations of the results for the individual questions. The first is that the game-design failed to produce adequate equivalence and the mini-game was not fit for purpose. The mini-game showed the player protein per dish as well as total consumed based on their choices. However, it is likely that the players did not have a clearly defined value for ‘high in protein’ or that it differed from person to person. They could have also chosen the dishes using a different criteria like sensory appeal. The results for the two questions did not correlate with each other or any other metric. The mini-game doesn’t show detailed description of vitamins and minerals, instead showing bulk values and uses these for calculating the result. In either case, the game-proxy design requires further refinement with the values playing a better-defined role in the choices. An alternative interpretation is that the proxy approach works better on categories rather than single directed questions. Since a questionnaire like the FCQ is designed with pre-defined categories, a single mini-game directed at a single question is neither efficient nor practical. FoodChoices(Q) suggests that if the game-play can include the relevant mechanics that mimic action or responses to multiple individual scale-items, a single stage can generate category-level data without having to rely on multiple separate mini-games.

7.4 Implications on Proxy Design

Our work shows the use of game-proxy to replace an established questionnaire as feasible but with caveats. The four mini-games show different perspectives to the design process and highlight the non-trivial effort required to create the game proxies.

7.4.1 Skewing of results

During design, care needs to be taken to prevent standard in-game behaviours (like competition or boredom) from skewing the results generated by the proxy. The choice of the game mechanic needs to balance the implicit objective versus the explicit effort required to attain it. Presence of rewards and attainment must be carefully considered to prevent them from appearing as confounds. The results also indicate that in the absence of a well understood baseline questionnaire for comparison, this process is unlikely to produce a proper tool for research. The baseline questionnaire is essential to identify the presence of experimental confounds that emerge from the game-play itself.

7.4.2 Validation of equivalence

We suggest that the proxy be validated using a realistic population sample to demonstrate equivalence. The validation includes an NS result for means of paired sampling of questionnaire responses versus in-game results which is then followed up by the TOST approach described by Lakens et. al [23] with a suitable choice of Cohen's d. In case the data violates normality, Bayesian methods are recommended for further analysis. After validation through a suitable sample size, the game can be considered as an equivalent instrument to the survey and deployed to a larger population. At the same time, based on our methodology and observations, we do not recommend using a serious game as the instrument of first implementation instead of proxy as in our case.

7.4.3 Factor of development time

The time required to develop a suitable game and the time taken to complete the game as compared to equivalent steps for a survey will be a mitigating factor. FoodChoices(Q) took ~800 person hours (ph) to develop from initial conceptualization of the research question to completion of study and analysis. This includes time for novice game developers (no prior experience of Godot) and time taken to identify and understand the statistical procedures. Re-use of existing game artefacts cut down the asset development time by ~150 ph. This is comparable to the time taken to develop the original survey instrument itself. With experienced game designers and developers, the implementation time could be cut down to ~200 ph. With sufficient re-use of the instrument, this would be efficient for the targeted improvement in completion rates.

7.4.4 Factor of player time

The overall time taken to complete a single session of the game is considerably higher than filling in the textual survey. This must be contrasted with survey fatigue which is a persistent issue with text-based surveys. Games are a special category of applications where performance in terms of task efficiency is not the core criteria, rather experience is. In certain cases, the game design forces repetitive but meaningful actions which would be counterintuitive to HCI practitioners. Well-designed games provide a more pleasurable experience than applications designed for efficiency and time spent on the game becomes less important. For example, Candy Crush^{TM} players reportedly spend ~38 minutes of play time per day on average[4]. The survey instrument is a very good candidate to be presented either as a proxy game or as a gamified application since it cannot be auto-completed by a program and relies on the information sourced from the player. If the experience is pleasurable, the time taken to complete the survey becomes a lesser concern for the participant. As an alternative to stand-alone presentation of the proxy, it could be split into smaller pieces and embedded into other activities. For example, in-game advertising of other games is common in freemium games. The inspiration used for our proxy-design, the game called Choices, is often advertised within other games with pseudo-interactivity. The proxy mini-games could be deployed in place of the advertisements. This allows the respondent to engage with the proxy as part of their flow during a gaming session. However, we do not condone nor encourage the use of the proxy without explicit consent of the participant since consent is essential for research.

8 CONCLUSION

In the paper, we presented FoodChoices(Q), a serious game proxy for the Food Choice Questionnaire. We identify opportunities and challenges within the design and evaluation activity to validate the serious game as a usable proxy instead of an established questionnaire instrument. However, we also see advantages in using such a serious game proxy as it can overcome survey fatigue by obfuscating the perception of work associated with filling a questionnaire and replacing it with a fun activity within a game. We believe this is an exciting direction of research for serious games in general.

ACKNOWLEDGMENTS

Acknowledgments left blank for review.

REFERENCES

[1] Clark C Abt. 1987. *Serious games*. University press of America.

[2] Dmitry Alexandrovsky, Susanne Putze, Michael Bonfert, Sebastian Höffner, Pitt Michelmann, Dirk Wenig, Rainer Malaka and Jan David Smeddinck. 2020. Examining Design Choices of Questionnaires in VR User Studies. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 1–21.

[3] Pippa Bailey, Gareth Pritchard and Hollie Kernohan. 2015. Gamification in Market Research:Increasing Enjoyment, Participant Engagement and Richness of Data, but what of Data Validity? *International Journal of Market Research*, 57 (1). 17-28. 10.2501/ijmr-2015-003

[4] Paul P Biemer and Lars E Lyberg. 2003. *Introduction to survey quality*. John Wiley & Sons.

[5] Sheldon Cohen. 1994. Perceived stress scale.

[6] Scott D. Crawford, Mick P. Couper and Mark J. Lamias. 2001. Web Surveys:Perceptions of Burden. *Social Science Computer Review*, 19 (2). 146-162. 10.1177/089443930101900202

[7] Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.

[8] Luís Miguel Cunha, Diva Cabral, Ana Pinto Moura and Maria Daniel Vaz de Almeida. 2018. Application of the Food Choice Questionnaire across cultures: Systematic review of cross-cultural and single country studies. *Food Quality and Preference*, 64. 21-36. https://doi.org/10.1016/j.foodqual.2017.10.007

[9] Sebastian Deterding. 2012. Gamification: designing for motivation. *interactions*, 19 (4). 14–17. 10.1145/2212877.2212883

[4] https://investor.activision.com/news-releases/news-release-details/activision-blizzard-announces-first-quarter-2019-financial

[10] Sebastian Deterding, Dan Dixon, Rilla Khaled and Lennart Nacke. 2011. From game design elements to gamefulness: defining "gamification" *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, Association for Computing Machinery, Tampere, Finland, 9–15.

[11] Lara Dorcec, Dario Pevec, Hrvoje Vdovic, Jurica Babic and Vedran Podobnik. 2019. How do people value electric vehicle charging service? A gamified survey approach. *Journal of Cleaner Production*, 210. 887-897. https://doi.org/10.1016/j.jclepro.2018.11.032

[12] Martin Flintham, Richard Hyde, Paul Tennent, Jan-Hinrik Meyer-Sahling and Stuart Moran. 2020. Now Wash Your Hands: Understanding Food Legislation Compliance in a Virtual Reality Restaurant Kitchen. in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, Association for Computing Machinery, 169–180. https://doi.org/10.1145/3410404.3414237

[13] Christos Fotopoulos, Athanasios Krystallis, Marco Vassallo and Anastasios Pagiaslis. 2009. Food Choice Questionnaire (FCQ) revisited. Suggestions for the development of an enhanced general food motivation model. *Appetite*, 52 (1). 199-208. https://doi.org/10.1016/j.appet.2008.09.014

[14] Julian Frommel, Katja Rogers, Julia Brich, Daniel Besserer, Leonard Bradatsch, Isabel Ortinau, Ramona Schabenberger, Valentin Riemer, Claudia Schrader and Michael Weber. 2015. Integrated Questionnaires: Maintaining Presence in Game Environments for Self-Reported Data Acquisition *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, Association for Computing Machinery, London, United Kingdom, 359–368.

[15] Johannes Harms, Stefan Biegler, Christoph Wimmer, Karin Kappel and Thomas Grechenig. 2015. Gamification of Online Surveys: Design Process, Case Study, and Evaluation. in, Cham, Springer International Publishing, 219-236.

[16] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. in Hancock, P.A. and Meshkati, N. eds. *Advances in Psychology*, North-Holland, 139-183.

[17] Casper Harteveld, Sam Snodgrass, Omid Mohaddesi, Jack Hart, Tyler Corwin and Guillermo Romera Rodriguez. 2018. The Development of a Methodology for Gamifying Surveys *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, Association for Computing Machinery, Melbourne, VIC, Australia, 461–467.

[18] Joshua Hill, Edward Corke, Mubarak Salawu, Ethan Cotterell, Matthew Russell, Joshua Gibbons, Tianxiang Mu and Abhijit Karnik. 2021. Point of Contact: Investigating Change in Perception through a Serious Game for COVID-19 Preventive Measures. *Proc. ACM Hum.-Comput. Interact.*, 5 (CHI PLAY). Article 274. 10.1145/3474701

[19] Robin Hunicke, Marc LeBlanc and Robert Zubek. 2004. MDA: A formal approach to game design and game research. in *Proceedings of the AAAI Workshop on Challenges in Game AI*, San Jose, CA, 1722.

[20] Laura Hyman, Julie Lamb and Martin Bulmer. 2006. The use of pre-existing survey questions: Implications for data quality. in *Proceedings of the European Conference on Quality in Survey Statistics*, Cardiff Wales, UK, 1-8.

[21] Florian Keusch and Chan Zhang. 2017. A Review of Issues in Gamified Surveys. *Social Science Computer Review*, 35 (2). 147-166. 10.1177/0894439315608451

[22] Tim Kuhlmann, Ulf-Dietrich Reips, Julian Wienert and Sonia Lippke. 2016. Using Visual Analogue Scales in eHealth: Non-Response Effects in a Lifestyle Intervention. *Journal of medical Internet research*, 18 (6). e126-e126. 10.2196/jmir.5271

[23] Daniël Lakens, Anne M. Scheel and Peder M. Isager. 2018. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1 (2). 259-269. 10.1177/2515245918770963

[24] Godot Engine. https://godotengine.org/. *Last Accessed: 28/07/2021*

[25] Rocio Lorenzo-Alvarez, Teodoro Rudolphi-Solero, Miguel J. Ruiz-Gomez and Francisco Sendra-Portero. 2020. Game-Based Learning in Virtual Worlds: A Multiuser Online Game for Medical Undergraduate Radiology Education within Second Life. *Anatomical Sciences Education*, 13 (5). 602-617. https://doi.org/10.1002/ase.1927

[26] Saturnino Luz, Masood Masoodian, Raquel Rangel Cesario and Manuel Cesario. 2016. Using a serious game to promote community-based awareness and prevention of neglected tropical diseases. *Entertainment Computing*, 15. 43-55. https://doi.org/10.1016/j.entcom.2015.11.001

[27] Jerko Markovina, Barbara J. Stewart-Knox, Audrey Rankin, Mike Gibney, Maria Daniel V. de Almeida, Arnout Fischer, Sharron A. Kuznesof, Rui Poínhos, Luca Panzone and Lynn J. Frewer. 2015. Food4Me study: Validity and reliability of Food Choice Questionnaire in 9 European countries. *Food Quality and Preference*, 45. 26-32. https://doi.org/10.1016/j.foodqual.2015.05.002

[28] Stephen R. Porter, Michael E. Whitcomb and William H. Weitzer. 2004. Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004 (121). 63-73. https://doi.org/10.1002/ir.101

[29] Susanne Putze, Dmitry Alexandrovsky, Felix Putze, Sebastian Höffner, Jan David Smeddinck and Rainer Malaka. 2020. Breaking The Experience: Effects of Questionnaires in VR User Studies. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 1–15.

[30] Katie Salen and Eric Zimmerman. 2004. *Rules of play: Game design fundamentals*. MIT press.

[31] Andrew Steptoe, Tessa M Pollard and Jane Wardle. 1995. Development of a measure of the motives underlying the selection of food: the food choice questionnaire. *Appetite*, 25 (3). 267-284.

[32] Muriel C. D. Verain, Harriette M. Snoek, Marleen C. Onwezen, Machiel J. Reinders and Emily P. Bouwman. 2021. Sustainable food choice motives: The development and cross-country validation of the Sustainable Food Choice Questionnaire (SUS-FCQ). *Food Quality and Preference*, 93. 104267. https://doi.org/10.1016/j.foodqual.2021.104267

[33] Jane Wardle, Carol Ann Guthrie, Saskia Sanderson and Lorna Rapoport. 2001. Development of the Children's Eating Behaviour Questionnaire. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42 (7). 963-970. 10.1017/S0021963001007727

[34] Koji Yatani. 2016. Effect Sizes and Power Analysis in HCI. in Robertson, J. and Kaptein, M. eds. *Modern Statistical Methods for HCI*, Springer International Publishing, Cham, 87-110.