

# Game for Heuristic Evaluation (G4H): A Serious Game for Collaborative Evaluation of Systems

Paulyne Matthews Jucá, Ingrid Teixeira Monteiro, and José Cezar de Souza Filho

Universidade Federal do Ceará, Quixadá- CE, Brazil  
paulynejuca@gmail.com@gmail.com, ingrid@ufc.br,  
cezarbx@gmail.com

**Abstract.** Many initiatives have promoted Collaborative Heuristic Evaluation in order to avoid discrepancies between evaluator's ratings. This paper presents a gamification called G4H (Game for Heuristic Evaluation), a card game proposed to increase the engagement of different evaluators in an evaluation process based on Heuristic Evaluation. This paper presents all the rules, cards, game loop and the results of a preliminary study made to validate the game. The G4H can be used as complementary material in HCI courses. The preliminary study demonstrated an increase in satisfaction in participating of system evaluation using the G4H.

**Keywords:** Heuristic evaluation·Gamification·Card game.

## 1 Introduction

Heuristic Evaluation (HE) is a usability inspection method, in which some evaluators inspect a system interface searching for violations of one or more usability heuristics [1], which are general principles for interaction design. One of the steps of HE is rating the violations in order to prioritize the fixing efforts. Despite being a very popular evaluation method, HE faces serious criticisms of its validity and reliability [2]. One of these criticisms are the large discrepancies between the individual severity ratings of evaluators, indicating challenges with the rating process [3]. Some initiatives have been proposed to increase the number of evaluators in order to promote collaboration and enable sharing the “challenge and frustration of the evaluation process” [2]. Other authors have been investing in turn HE easier, for instance, [4] that proposes a guide for expert evaluation and [5] that proposes UX Check, a tool to support HE. In addition, Collaborative Heuristic Evaluation (CHE) [2], which is not new, represents a possible improvement in the evaluation results. This paper presents a gamification called G4H (Game for Heuristic Evaluation) based on competitive mechanics in order to increase the engagement of different evaluators in an evaluation process based on HE.

## 2 Related Work

There are two main types of related work to this paper: collaborative evaluation and gamification used for HCI. The first group contains papers like [1-5] that debate the challenges and advantages on using collaborative methods to achieve better evaluations. They are related to this paper because they have the goal to promote collaboration, but the main difference with this paper is that they don't use game mechanics to promote engaging. This paper intends to use game elements to increase, promote and engage user (especially non-specialists and end-users in heuristic evaluation).

The second group is composed by works that already promoted the use of gamification with HCI. In [4], the authors propose a gamification that should be used for specialists. That itself represent a significant difference for the proposition presented by this paper. In [7], the author intends to indicate how the gamification is being used in HCI area. It comments that the majority of the work that combines gamification and HCI intends to understand how different players and the system personalization influences on interaction. This paper intends to present a gamification that uses Nielsen heuristics to promote collaborative evaluation that can be used by non-specialists and end-users. G4H was initially tested using undergraduate students. It is expected that an improved version of this gamification can be used by HCI teachers during their lectures as a side class exercise.

## 3 Gamification

Gamification [6] is the use of game elements and mechanics on non-game problems in order to provide fun to increase engagement. "In HCI, the study of gamification has been often part of the sub-domains of Player-Computer Interaction (PCI) and Player Experience (PX), which study the experience of players interacting with games" [7]. This paper presents a gamification focused on Heuristic Evaluation.

The gamification framework proposed by Kevin Werbach and Dan Hunter [8] was used to create the gamification in this paper. The main objective is to engage non-expert and/or end users in collaborative heuristic evaluation. To achieve this goal, the gamification focus on the extrinsic motivation of been recognized by the pair as a good evaluator and the intrinsic motivation of doing accurate evaluation and violation identification. The competitive character of the gamification is presented on the possibility of buy a new reevaluation phase if the player is not the winner of the level. The competition is designed to increase debate on the evaluation and it increases the quality of the final evaluation. The proposed gamification was based on the 10 heuristics of Nielsen, but it can be adapted for other heuristic frameworks. The Game for Heuristic Evaluation (G4H) will be presented in the next section.

## 4 Game for Heuristic Evaluation (G4H)

The gamification has a simple game loop in four steps: initial heuristic violation classification, initial severity classification, negotiation phase and reevaluation. In G4H, each heuristic is a level in the game. In each level, the player has to decide if the system in evaluation violates that heuristic (or not) and gives a grade according to the gravity of the fault. It is expected that different evaluators have different grades, especially if the evaluator has different experience in HCI and heuristics usage.

The game enables negotiations where players have to convince each other about the found violations and about the given grades. After negotiation, all players vote for violations that must be kept, changed or removed and also for a new grade for each heuristic violation based on their arguments. The player that grade correctly (most voted) wins the level and earn a point. At any level, the player can trade two points for the possibility of a new negotiation phase (with new votes replacing the previous votes). It increases the player's chance of convincing others on his/her arguments and win the level. The player who has more points will win the game.

The G4H will be presented in details in next sections.

### 4.1 G4H Setup

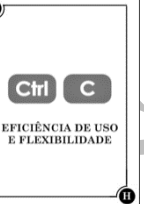
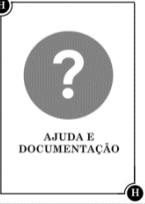
The setup represents all the necessary things that is supposed to be available before the game starts. All items that need to be available before the game are listed below:

- Heuristic violation cards (one set per player): each card represents one of the 10 heuristics.
- Severity cards (one set per player): each card represents one of the five severity levels of each violation found.
- Point cards: represent the rewards granted to player for evaluating the severity in the right way. The player with the greater amount of point at the end of the list of problems previously found wins the game.
- Tasks to be performed during the previous evaluation of the system. Each player individually performs the system inspection in order to find problems. The list of problems found will be used during the game, but the evaluation itself is out of the game context. Each problem identified will represent a round in the game.
- Access to the system under evaluation.
- A support card with a brief description of each heuristic. It can be used by the players to remember what each heuristic is about.
- List of problems found by each player.
- At least 3 players

### 4.2 G4H Cards

G4H has 3 types of cards: heuristics, severities and points. Table 1 shows the prototyped version of heuristics cards.

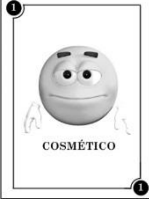
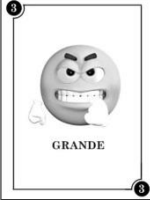
**Table 1.** Heuristics Cards

|                                                                                                                                                |                                                                                                                                    |                                                                                                                             |                                                                                                                          |
|------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
|  <p>VISIBILIDADE DO STATUS DO SISTEMA</p>                     |  <p>MAPEAMENTO ENTRE O SISTEMA E O MUNDO REAL</p> |  <p>CONTROLE E LIBERDADE DO USUÁRIO</p>   |  <p>USO DE PADRÕES E CONSISTÊNCIA</p> |
|  <p>PREVENÇÃO DE ERROS</p>                                    |  <p>RECONHECIMENTO AO INVÉS DE MEMORIZAÇÃO</p>    |  <p>EFICIÊNCIA DE USO E FLEXIBILIDADE</p> |  <p>ESTÉTICA E DESIGN MINIMALISTA</p> |
|  <p>RECONHECIMENTO, DIAGNÓSTICO, E RECUPERAÇÃO DE ERROS</p> |  <p>AJUDA E DOCUMENTAÇÃO</p>                    |  <p>NÃO É VIOLAÇÃO</p>                  |                                                                                                                          |

Each player receives a set of these 11 cards representing the 10 violation types and 1 card to represent that the player believes that the problem presented is not a violation. A problem found previously (see G4H setup) on the system can represent one or more violations.

Table 2 presents the severity cards used by the players to grade the violation's level of gravity.

**Table 2.** Severity Cards

|                                                                                   |                                                                                   |                                                                                   |                                                                                     |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|  |  |  |  |
| Cosmetic                                                                          | Minor                                                                             | Major                                                                             | Catastrophic                                                                        |

Regarding the point cards, the players receive one of the cards presented in Table 3 when they win a round. They can use these cards to buy reevaluation rounds. Players win one point when the result of the evaluation is unanimous. Otherwise, the players that win the round receive two points.

**Table 3.** Point cards

|                                                                                    |                                                                                     |
|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|  |  |
| One point                                                                          | Two points                                                                          |

### 4.3 G4H gameloop rules

The gamification has a simple game loop in four basic steps: initial evaluation of the heuristic violation, first evaluation of the severity, negotiation phase and reevaluation. The G4H full gameloop steps are:

1. The first player chooses one of the problems found previously by him/her. This will be the problem of the round. Repeated problems shouldn't be reevaluated unless that it represents a different violation. The player of the round may describe the problem found to the others. The players discuss the problem and make comments and questions to the player that found the problem at the system. The players can also access the system to demonstrate the problem by executing again the task that has the problem.
2. Heuristic Evaluation: the players choose one or more cards of heuristic (Table 1) to indicate the violation for the problem found. These cards should be hidden at first in order to not influence other player's evaluation. When all the players have ended their evaluations, they all turn their cards at the same time.
3. Debate about the heuristic selection. The players, one at a turn, debate how and why they have selected the violations cards.
4. Final selection of heuristic violation. The players vote again on which violations they consider right. The violated heuristic chosen by the greater number of players is the

one considered to the next steps. One exception is when the majority of players decide that it is not a violation and the game returns to step 1.

5. Severity classification. Since a heuristic violation is confirmed, the players need to classify the severity of this violation. Again, each player chooses a severity card (Table 2) and maintains it hidden until all the players finish selecting a card in order to avoid cross player influence. The players show their cards at the same time. If all the players agree with the same classification, each player receives 1 point (Table 3) and the round ends here. If there is some divergence on the severity classification, the game continues to renegotiation step. It is important to remember the selection of each player in order to define the winner of the round after the renegotiation.
6. Negotiation. In this step the players need to explain the reasons that lead them to choose their severity classification (step 5). G4H doesn't define a fixed amount of time for negotiation, but it is recommended that each player talks at least one time. It can be a conversation and the player may make questions and debate. When the player decides that they have enough information to make a new classification, the game proceeds to next step.
7. Reclassification. All players select a severity card (that can be the same of the step 3 or a new one representing a change on his/her opinion). Again the selected card is maintained hidden until all players select their cards. The classification that is most selected is the severity chosen. Next step consists in defining the winner(s) of the round
8. Distributing points. The game awards the players that better guessed the final severity level (defined at the end of round 7) at first chance (step 5). The distribution of points for the players is made as follows: the player that has more selections of severity in final selection similar of their severity selection in step 3 receives 2 points. For instance: imagine one scenario where there are 3 players A, B and C. If during step 3 player A chooses "major", and player B and C chooses "minor" and after negotiation (step 7) the majority of player chooses "major", only player A wins 2 points. But if after renegotiation the majority of players chooses "minor", players B and C each receives 2 points. The last possibility is for the majority select a different severity classification, for instance "catastrophic". In this case, the player that is closer to the final classification wins 1 point. In the example, player A (that chosen "major" in step 5) wins 1 point. This can sounds difficult to understand, but the players in experiment had no difficult to use these rules.
9. Buying new negotiations. After negotiation, players can buy the chance of a renegotiation. It is interesting when the decision is almost even and a player wants another chance to win the round. In order to do this, the player gives 3 points away. These 3 points are lost independently of the results of the new negotiation. The points of the last step are returned and steps 6, 7 and 8 are replayed. There is only one renegotiation for round.
10. Game continues with the player at left. If this player doesn't have new problems anymore in his/her list, the game continues with the player at left until all problems previously found be evaluated.

#### 4.4 Frequent Suggestions About the Rules

The game has been informally tested many times before the first experiment with users. The interesting part is that some beta testers and others HCI researchers that had access to the game have made questions in order to change the rules, specially about giving points to heuristic violation classification. The questions and suggestions were so common that they need to be also presented in here.

Balancing a game may be difficult and it is necessary to imagine how the solutions provided could be misused by the players. The gamification designers need to prevent misuse because it can lead into results much different from the previously desired. Some of the most common suggestions are listed below.

##### **Reward the player that found a problem in the system**

The number of problems found is not the same for each player. The main goal of the game is not award the better inspector or specialist but it is to promote the collaboration. This is the motive why heuristics violation identification in step 1 does not result in points. If it does, the game would need to make the concurrency just by limiting the number of rounds by the minimum number found by each player.

##### **Reward renegotiate heuristic violation classification**

Both the heuristic violation classification and severity classification are part of one round. The rewards are from the round. Putting more points on the table could make the renegotiation cost too cheap and not valuable to the players.

##### **Reward player according their guess after negotiation and not the severity level chosen before negotiation**

There is also a reason to not reward the person that changed the opinion. By doing so, some players could during the renegotiation chose a classification that is not correct just to win points. In this way, the players that understood that their first classification were incorrect tend to reevaluate the classification in better way since it doesn't alter their point rewards.

## 5 Using G4H

In order to investigate the reception of the initial version of the game, we performed a preliminary study with participants evaluating an academic system.

The study was made with 5 undergraduate students of the Federal University of Ceará. All the students have concluded at least one course on HCI and had previous basic knowledge on Nielsen Heuristics and System Evaluation. It was the main criteria for participating of the study since the player had to prepare a list of problems found in the system to start the game.

The system under test was a system to search and compare prices of different products and find products with better prices. This system is very popular in Brazil<sup>1</sup>.

All the students previously received the necessary information to evaluate the system (and a set of tasks to perform on the system) and were instructed on the G4H gamification rules.

The names of the students were changed in order to respect their privacy. A term of consent was also signed by all the participants.

To guarantee that all the players understood the rules of the game, a demonstration round was performed.

The players had 30 minutes to evaluate the system individually in order to create the list of problems found. During the study, each player used 1 problem (5 in total) because of time limitations (the game last around 50 minutes). The number of problems found (within the 30 minutes provided) by each player was:

- Natalia: 5 problems;
- Igor: 3 problems;
- Marcos: 4 problems;
- Davi: 4 problems;
- Alice: 2 problems.

### 5.1 Demonstration Round

In this demonstration round the problem found in the system was the difficulty on finding a product category because they are located at the bottom of the page and the user has to scroll the full page in order to see the information.

The players debated about the violations, and after the card selection the most chosen violated heuristic were: Consistency and standards and Flexibility and efficiency of use.

After the selection of the heuristic, the researcher instructed the players that they need to select the severity level of the violation and highlighted that this was an important evaluation since that it is the part of the game that will be considered in the distribution of points. At first, 2 players selected “major” for severity level and 3 selected “minor” for severity level.

The player of the round (the same that identified the problem) justified that he has chosen “major” imagining that his mother wouldn’t use this system since she couldn’t find the tags at the bottom of the page. This was one of the arguments in this discussion.

After negotiation, all players selected the severity level “minor” and 3 players won 2 points each.

During this round, some players made some interesting comments about the game and the cards. They are listed as follows:

- Alice commented that the cards were difficult to use. It happened because it was a prototype of the game cards, the cards were small (around 5cm x 3cm) and impressed in standard A4 paper.

---

<sup>1</sup> The system was Mercado Livre (“free market”): <http://www.mercadolivre.com.br/>



- Alice had doubts if she could use the guide with the heuristic descriptions during the game and commented that the usage could made the game last too long.
- Some players asked if they could select more than one heuristic violation.
- One player asked if the severity level was applied to each heuristic or to the problem.
- Some players have shown surprise (“Eita!” and “Vixe!” expressions that are similar to “Wow!” in English language) with the heuristic selected by other players.

## 5.2 Effective Rounds

The effective rounds were performed without the help of the researchers. The researchers acted only as observers of the study.

### Problem 1

Description: The system doesn’t allow the user to increase the number of products on the order and doesn’t inform max available number of the product in stock as an error message.

After debate the players selected Visibility of system status as the violated heuristic.

During the first severity level rating, 4 players selected “major” and 1 player selected “cosmetic”. After negotiation, all players agreed on “major” severity rating.

The researchers reminded the players of the possibility of buying renegotiations, but they decided that it was not worth it.

The most interesting comments of the round were:

- The player that has chosen “cosmetic” started the discussion reflecting on if the severity level should be selected considering only the user described on the task scenario or if they could evaluate based on the problem itself and considering other users. The research informed that the task is a guide, but the evaluation should be over the system and its usage.
- Alice asked “Nobody else has chosen more than one heuristic?” and “You have chosen only that?”. Davi answered “Accept it!”. And Alice responded “I do not.” and then laugh. Alice said “I don’t know. There isn’t any other violation?”. At the end she said “Ok, I accept it.”.
- During the heuristic violation debate, Marcos said “And what’s next?” and “Does anybody want to say anything more?”. Alice answered “No, all we need to do is to go”.
- During the negotiation of severity level, the players again have shown surprise about the cards selected. Davi said “Only me have chosen cosmetic?” and 2 other players asked “Why cosmetic?”
- Alice said to Davi “Now you are going to lose 2 points. That’s what is going to happen”. But Marcos asked Davi “What else you have to say about that?”
- After the new severity card selection Alice said “He has been convinced” and everybody laughed.

### Problem 2

Description: It isn't clear if the medal symbol that represents the best sellers refers to the seller of the product listed. To know for sure, it is necessary to click at the symbol and then the system selects a filter for best sellers. The page doesn't order the products for better qualified sellers.

At first, only one player selected 2 types of heuristic violation. But after the debate, the majority of players decided to select 2 violations: Visibility of system status and Consistency and standards.

During the selection of severity level, all the players agreed at first with "minor" and all won 1 point each.

The most interesting comments of the round were:

- Davi said "I think that is Consistency and standards" and Marcos answered "Ok, but only this?" and Igor said "No, I think that there is more than one in here."
- Natalia had trouble classifying the problem and joked "I think I will put all of them". And other player answered "I think that it will worth 4 points to whom get it right" making others players to laugh.
- When they all got the same severity level selection, Alice said "We only like when make fun of others, isn't it?"

### Problem 3

Description: The system doesn't allow selecting more than one brand on the search filters.

At first the majority of the players select the two heuristics that remain until the final selection: Consistency and standards and Flexibility and efficiency of use. Just on player had chosen just one violation at first.

The main discordance occurred during the severity rating when at first 3 players selected "minor", 1 player selected "major" and 1 player selected "cosmetic". The final rating was "minor".

The most interesting comments of the round were:

- Just after Davi presented the problem found, Alice said "I already know the answer" and laugh. And Natalia said "She knows because I said to her" and Davi said "What? That's persuasion" and they laugh again.
- But during the selection she already had second thoughts and said "Wait!", "It is not that", "Ah... it is also part of the answer" and "Done".
- After first selection of heuristic violation only Igor had selected just one violation and Davi said "We only have to convince Igor" and Alice agreed and Natalia said "We don't need to, because it is the majority that decides" and Alice said "Sorry, Igor"
- After the first selection of severity rating the players have shown surprise. They said things like "ops, ops" and "easy, easy"
- During the reclassification Marcos said "Hold on, I'm analyzing case by case" and all players laugh. Then Davi said "the grandma, the great grandmother..." and Marcos agreed saying "I'm also thinking since grandmother..."

- Alice said “With this error, I would lose my patience and quit buying anything”

#### **Problem 4**

Description: It is not possible to inform the Zip code to calculate the shipping tax and the page rolls to the end when user tries to write comments to the seller.

When the problem is presented, Davi discuss that the problem doesn't occur with every product, just with some specific type of transactions. Because of this, Alice asked whether this is a case of non-violation, but after some debate they decide that it is a violation and identify not just one, but 3 of them: Help and documentation; Visibility of system status and Consistency and standards.

All players but Igor decided for “major” as severity level.

The most interesting comments of the round were:

- When classifying the heuristic for the first time, Marcos said “Eita (equivalent to Wow)” and Davi said “Lots of questions”
- Alice said “This was easier than the others”
- When she saw that her heuristic violation was similar to the majority, she said “Ehh” (winning sound) and “Everybody but Igor won 2 points, isn't it?”

After this round, the researchers decided to stop the study for 2 reasons: it already lasted 50 minutes (plus 30 for pre-evaluation of the system) and each player had run a round. But some players asked if this was the last round and have shown interest on continuing to play. After asked if they found interesting to apply this gamification during heuristic evaluation they all agreed that would be interesting and would like to participate. They also said that in this way, they could debate more about the classification than when they have done the evaluation alone.

The final points were:

- Natalia: 7 points;
- Igor: 7 points;
- Marcos: 7 points;
- Davi: 7 points;
- Alice: 4 points.

## **6 Risks**

If some player has significant superior level of convincement or even power or hierarchic position upon the others players it can unbalance the evaluation levels during negotiation phase.

Too competitive persons, like Alice can create a tension that can prevent other players to make comments. In this study, it wasn't a problem, because the students were colleagues and they understood her comments as jokes. But this problem would occur in every collaborative method and not only with this game.

## 7 Conclusions

This paper has proposed a gamification system to increase collaboration during heuristic evaluation. In order to investigate the reception of the initial version of the game, we performed a preliminary study with participants of different profiles evaluating a system for compare prices and sell products. During the application of the game, the volunteers demonstrated an increase in satisfaction in participating of system evaluation. The result is interesting especially when involving non-specialists and end-users working in the evaluation of the same system. Participants, in general, provide positive feedback and suggested some improvements that will be addressed in next versions of the game.

## References

1. Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90), Jane Carasco Chew and John Whiteside (Eds.). ACM, New York, NY, USA, 249-256. DOI=<http://dx.doi.org/10.1145/97243.97281>
2. Helen Petrie and Lucy Buykx. 2010. Collaborative Heuristic Evaluation: improving the effectiveness of heuristic. In Proceedings of the UPA Conference (Munich UPA2010).
3. Sascha Herr, Nina Baumgartner, and Tom Gross. 2016. Evaluating Severity Rating Scales for Heuristic Evaluation. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16). ACM, New York, NY, USA, 3069-3075. DOI: <http://dx.doi.org/10.1145/2851581.2892454>
4. Andrés Lucero, Jussi Holopainen, Elina Ollila, Riku Suomela, and Evangelos Karapanos. 2013. The playful experiences (PLEX) framework as a guide for expert evaluation. In Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces (DPPI '13). ACM, New York, NY, USA, 221-230. DOI=<http://dx.doi.org/10.1145/2513506.2513530>
5. Chris Gallelo. Low cost usability testing Or...why I built UX Check. <https://medium.com/@cgallelo/low-cost-usability-testing-61b5f8a2a1be#.xvgurmt7e>
6. Sebastian Deterding; Dan Dixon; Rilla Khaled and Lennart Nacke, 2011. From game design elements to gamefulness: defining "gamification". In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek'11). ACM, New York, NY, USA, 9-15. DOI=<http://dx.doi.org/10.1145/2181037.2181040aa>
7. Gustavo Tondello. An Introduction to Gamification in Human-Computer Interaction. Crossroads: The ACM Magazine for Students. <http://xrds.acm.org/blog/2016/04/introduction-gamification-human-computer-interaction/>
8. Kevin Werbach and Dan Hunter. For The Win: How Game Thinking Can Revolutionize Your Business. Wharton Digital Press, 2012.