Evaluating the Interface Using Expert-heuristic Method

Ulyana Khaleeva [1]

[1] *Nizhny Novgorod State Technical University n.a. R.E. Alekseev, 24 Minin str., Nizhny Novgorod, 603950, Russia*

Abstract

The research aims to form a new method for evaluating interfaces, ensuring its multi-criteria nature and eliminating the shortcomings of previous methods. A combination of expert and heuristic approach is proposed, to detect a wide range of UI/UX problems, to ensure assessment competence and to reduce the level of distrust of the expert. In the first experiment, two groups of interfaces with different characteristics were evaluated, with two interfaces in each group. Fifteen heuristics were evaluated: ten general purpose criteria and five specialized criteria. Thirteen experts were involved, for whom weighting coefficients were previously calculated, taking into account their professional competencies and personal qualities influencing the reasonableness of the evaluation. After analyzing the results of the first experiment, it was decided to investigate the influence of the number of experts in the sample on the overall UI score. Therefore, for the second experiment, the optimal number of experts in the group was calculated to ensure the lowest score variance. Applications were evaluated in five groups (the number of heuristics did not change). Also, in each experiment, the outlier weights of the experts were calculated to ensure consistency of the opinions of the sample group members. In the conclusion, an analysis of the feasibility of applying the new method to mobile interfaces was performed. Conclusions on the suitability of the chosen mathematical apparatus and further ways of development of the method have been made.

Keywords

expert evaluation, heuristic evaluation, evaluation methods, user interface, expert weighting, UI, UX

1. Introduction

In a highly competitive environment, companies are forced to invest huge sums in the development of advertising and information support for business - sites and applications are becoming a necessary component to ensure the success of the enterprise, and thus make a profit.

According to the statistics [1] (Table 1), the cost of website development, taking into account analytical activities ranging from 29 000 rubles. - landing page, up to 400 000 rubles - portal.

Table 1

The cost of the various stages of website development in 2021

Development phase	Time spent	Minimum price rub./hour	Maximum price rub./hour
Analytics and strategy	80-360 hours	1500	3400
UI / UX design	80-400 hours	1200	3200
Front-End development	120-600 hours	1800	3800
Back-End development	120-600 hours	4000	6000
Total	175-760 hours	8500	16400

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27-30, 2021, Nizhny Novgorod, Russia
EMAIL: u.gulyaeva@nntu.ru (U. Khaleeva)
ORCID: 0000-0002-3527-4752 (U. Khaleeva)

Note that a significant portion of the cost is spent on design and user interaction strategy. This stage also involves evaluating the interface, which can significantly reduce costs by reducing the number of edits and, as a consequence, iterations of redesign.

Based on the foregoing, the goal of the study was determined: the development and testing of a new method for assessing the interface, combining a qualitative and quantitative component.

To do this, it is necessary to perform the following tasks:

- Analysis of existing methods for assessing interfaces;
- Development of an evaluation algorithm with the following properties: flexibility based on the functional features, complexity and / or scope of the interface; speed and ease of use; potential for formalization; the possibility of reducing or completely eliminating subjective perception;
- Selection of the mathematical apparatus;
- Approbation of the method on various interfaces;
- An overview of potential opportunities for formalization;
- Development of recommendations for improving the method.

It is assumed that as a result of using the new method, the customer will be able to obtain both an overall assessment of the interface and individual criteria, which helps to determine the elements that need to be modified in the first place. Additionally, it is possible to develop recommendations based on expert opinion to improve the project.

In a previous study [2] was considered the method of expert-heuristic evaluation of interfaces, which allows with sufficiently high accuracy to evaluate user interfaces also due to the elaborate system of heuristics, taking into account both general and specific features of those or other groups of interfaces. Also note that this algorithm significantly reduces the subjective component of the evaluation and allows to eliminate the disadvantages of using the GOST system.

2. Calculation of interfaces estimation using expert-heuristic method. Experiment 1

At the first stage described in the previous part of the experiment [2] according to the method of calculating weighting coefficients based on a questionnaire survey to determine the level of competence of an expert, the following data was obtained (Table 2). In the first experiment of applying the method, a group of 13 experts was formed.

Table 2

Weighting coefficients of experts

Expert №	w_j	% in the total estimate
Expert 1	0,25	6,693440428
Expert 2	0,28	7,49665328
Expert 3	0,385	10,30789826
Expert 4	0,15	4,016064257
Expert 5	0,49	13,11914324
Expert 6	0,315	8,43373494
Expert 7	0,24	6,425702811
Expert 8	0,315	8,43373494
Expert 9	0,28	7,49665328
Expert 10	0,2	5,354752343
Expert 11	0,35	9,3708166
Expert 12	0,3	8,032128514
Expert 13	0,18	4,819277108

The second stage of the experiment included the direct evaluation of UI. As prototypes were used works of 4th year students of NSTU n.a. R.E. Alekseev, studying on 09.03.02 "Information systems

and technologies" major in "Information technologies in design" within the study of "Mobile application development" discipline.

In the first experiment, each expert was asked to evaluate 4 interfaces grouped in pairs: Group A - browsing and maintaining (creating) content, Group B - training applications and simulators (Figure **1**) - according to 15 heuristics [3].

A set of heuristics, among which there were 10 general and 5 highly specialized questions, provides a quick experiment and allows us to determine the applicability of the method to mobile interfaces.

The heuristics included the following general questions:

1. Level of interface compliance with HIG (Human Interface Guidelines - Apple's application and interface development guidelines);
2. The level to which the interface is easy to navigate;
3. The level of clarity, the obviousness of the icons and symbols;
4. The level of consistency of the interface color palette with the target audience (TA);
5. The level of readability of textual information and headings;
6. Level of compositional integrity;
7. The user friendliness [4] of the interface;
8. Convenience of the registration procedure;
9. Easy filtering and categorization;
10. The convenience of the search procedure.

The heuristics also included questions for a specific application category, such as Group A (viewing and maintaining content):

1. Easily save and view bookmarks/favorite entries;
2. Easy to add a new publication/record;
3. The convenience of chatting / correspondence;
4. Easy to set up a profile/account;
5. Level of personal satisfaction with the color palette of the interface.

For group B (training applications and simulators), the special questions were:

1. Ease of interaction with content/tasks/exercises;
2. Easy display of statistics/progress;
3. The convenience of adding a mark of completion of the task;
4. Easy to set up a profile/account;
5. Level of personal satisfaction with the color palette of the interface.

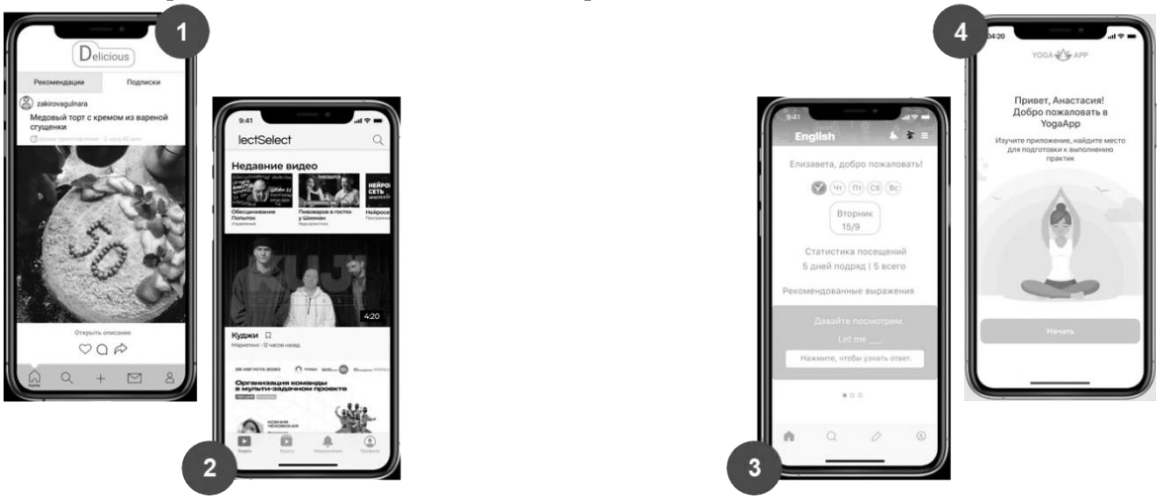


Figure 1: Interfaces for evaluation. 1,2 – group A, 2,3 – group B

Then we calculated the total score by assigning points to a single criterion r_i according to the formula [5]:

$$r_i = \frac{\sum_{j=1}^{n} r_{ji} \cdot w_j}{\sum w_j},\ i = \overline{1,m}, \quad (1)$$

where m is the number of heuristics, r_{ji} – normalized (by multiplying by 0.1 to bring the score value in the range from 0 to 1) score of interface compliance with the allocated criterion from 0 to 10,

w_j is the weight coefficient of the expert, calculated in the first phase of the experiment [2].

The resulting score $r_i \cdot 100\%$ characterizes the average value of user satisfaction with this criterion and its compliance with the principles of usability.

If we consider the results of the evaluations of each of the experts as realizations of some random variable, we can apply the methods of mathematical statistics to them. The average value of the estimate for the *i-th* criterion

$$\overline{r_i} = \frac{\sum_{j=1}^{L} r_{ji}}{n} = \frac{1}{n}\sum_{j=1}^{L} r_{ji} = \frac{r_i}{n}, \quad (2)$$

where *n* is the number of experts.

The average value $\overline{r_i}$ expresses the collective opinion of the group of experts. The degree of consistency of the experts' opinions is characterized by the value

$$\sigma_i^2 = \frac{1}{n}\sum_{j=1}^{n}(r_{ji} - \overline{r_i})^2, \quad (3)$$

called the variance of the estimates. The smaller the value of the variance, the more confident you can rely on the found values of the $\overline{r_i}$ estimate of the importance of a particular criterion. As a measure of reliability of the cited expertise, we take

$$\beta = \frac{\sigma_i}{r_i}, \quad (4)$$

called variation. The average value of the estimate is used to $\overline{r_i}$ determine the weighting coefficients

$$\lambda_i = \frac{\overline{r_i}}{\sum_{i=1}^{m}\overline{r_i}}, \ i = \overline{1, m}, \quad (5)$$

λ_i reflects the degree of influence of the evaluation *of the i-th* criterion on the overall assessment of the interface, calculated by the formula:

$$r = \sum_{i=1}^{n} r_i \cdot \lambda_i \quad (6)$$

Thus, the overall degree of satisfaction with the interface in percentage terms is defined as $r \cdot 100\%$

The screenshot of a fragment of the calculation and evaluation table in Excel is as follows (Figure 2).

	A	B	C	D	E	F
1		Evaluate the degree of co	Evaluate the ease of navig	Evaluate the degree of cla	Evaluate the degree of co	Evaluate the degree of yo
2	Expert 1	10	10	10	9	7
3	Expert 2	10	8	9	10	10
4	Expert 3	9	8	8	8	8
5	Expert 4	9	10	9	10	10
6	Expert 5	9	10	10	10	10
7	Expert 6	8	9	8	9	9
8	Expert 7	9	8	9	9	10
9	Expert 8	10	10	10	9	9
10	Expert 9	9	7	6	10	10
11	Expert 10	7	6	7	6	8
12	Expert 11	8	9	8	9	9
13	Expert 12	10	10	10	9	8
14	Expert 13	9	9	9	9	9
15	Expert 14	7	6	5	8	6
16						
17	Expert 1	0,24	0,609911055	0,609911055	0,609911055	0,548919949
18	Expert 2	0,315	0,800508259	0,640406607	0,720457433	0,800508259
19	Expert 3	0,15	0,343074968	0,304955527	0,304955527	0,304955527
20	Expert 4	0,315	0,720457433	0,800508259	0,720457433	0,800508259
21	Expert 5	0,25	0,571791614	0,635324015	0,635324015	0,635324015
22	Expert 6	0,49	0,996188056	1,120711563	0,996188056	1,120711563
23	Expert 7	0,28	0,640406607	0,569250318	0,640406607	0,640406607
24	Expert 8	0,2	0,508259212	0,508259212	0,508259212	0,457433291
25	Expert 9	0,2	0,457433291	0,355781449	0,304955527	0,508259212
26	Expert 10	0,35	0,622617535	0,533672173	0,622617535	0,533672173
27	Expert 11	0,385	0,782719187	0,880559085	0,782719187	0,880559085
28	Expert 12	0,28	0,711562897	0,711562897	0,711562897	0,640406607
29	Expert 13	0,3	0,686149936	0,686149936	0,686149936	0,686149936
30	Expert 14	0,18	0,320203304	0,274459975	0,228716645	0,365946633
31	r_i	3,935	8,771283355	8,631512071	8,472681067	8,923761118
32	$\overline{r_i}$	8,622980577	0,62652024	0,616536577	0,605191505	0,637411508
33	λ_i		0,072657039	0,071499242	0,070183563	0,073920091
34	r	0,58878724	0,045521106	0,044081898	0,042474496	0,047117517
35	Total	58,878724				

Figure 2: The screenshot of a fragment of the calculation and evaluation table

For clarity, the normalized average score for each criterion is formatted using color scales. This allows you to see the most (bright green) and the least (red) developed aspect of the interface.

For example, the following results were obtained for the examined interfaces (Figure 3, Figure 4):

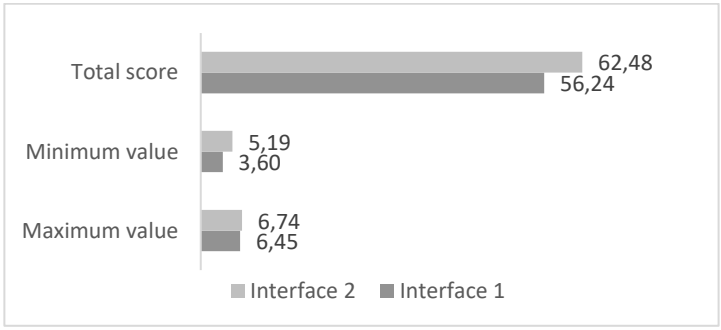


Figure 3: The result of the evaluation of Group A interfaces

The worst worked out:

- Interface 1 - "Mark of completion"
- interface 2 - "Search"

The best worked out:

- Interface 1 - "Search"
- Interface 2 - "Color Palette"

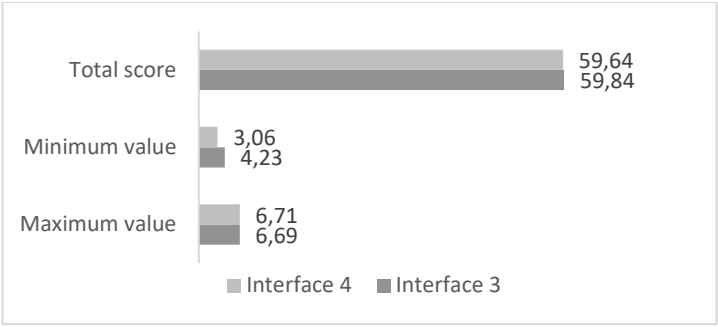


Figure 4: The result of the evaluation of Group B interfaces

The worst worked out:

- Interface 3 - "Mark of completion"
- interface 4 - "Search"

The best worked out:

- Interface 3 - "Search"
- Interface 4 - "Color Palette"

As a result of the experiment, the following regularities were confirmed:

- The position that the assessment is most dependent on the scores given by the expert with the highest coefficient of significance was confirmed;
- The degree of influence of the outlier grades given by "amateurs" is offset by their low ranking;
- Overestimates of experts with a high coefficient are averaged using the scores of the average expert category.

It was also decided to remove the question about individual color preferences from the list of heuristics, since this question concerns the subjective preferences of the expert. It is proposed to replace it with "Compliance with coloristic principles of interface construction".

3. Determining the number of experts in the sample group

For the second experiment, it was decided to change the number of experts in the sample.

It is proved that the number of experts must be large enough [6], so that individual opinions do not have an inappropriately large value. However, a sharp increase in the number of experts in the group decreases the level of their competence, which significantly reduces the accuracy of expert evaluations.

To calculate the number of the group of experts, we used the ratio that is used in calculating the error of observations [7]

$$N = t_p^2/\varepsilon_l^2, \quad (7)$$

where N is the number of experts in the group,

$\varepsilon_l = \varepsilon / S$ – maximum permissible relative error of expert estimation,

S – is the standard deviation of the distribution of estimates of any value,

t_p – is the Student coefficient, which determines the width of the confidence interval and the dependence on the value of the probability estimate P (t_p is a tabulated value).

Depending on the given error of expert evaluation and the chosen probability value, the minimum possible number of experts in the group N can be determined (Table 3).

Table 3

Minimum allowed number of experts in the group

ε_l	Probability of estimation P							
	0,99	0,95	0,90	0,85	0,80	0,75	0,70	0,65
0,5	26	15	11	8	7	5	4	4
0,3	74	43	31	23	19	15	12	10

Empirically, it was found that experts of 13-15 people can be considered a sufficiently representative group to conduct the examination.

This is confirmed by the dependence of the accuracy and reliability of the results of the estimation of the date of occurrence of the event on the number of experts in the group *N* (Figure 5).

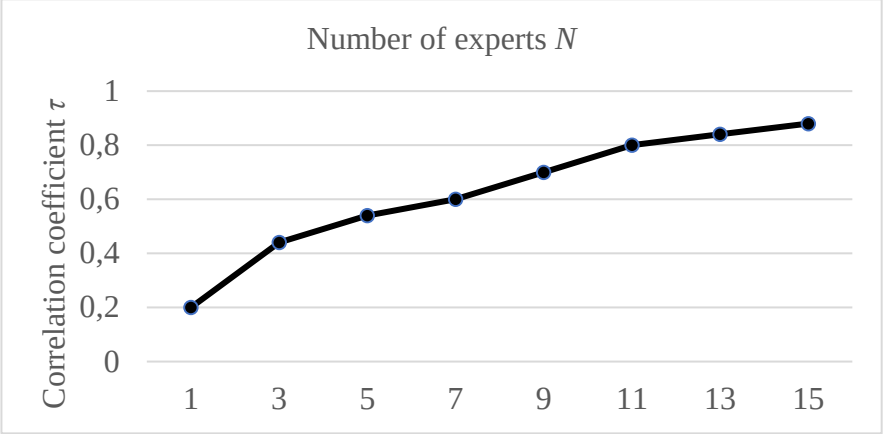


Figure 5: Relation of accuracy and reliability of the results of event timing estimation to the number of experts in the group *N*

Thus, it was concluded that the optimal solution would be to organize an expert group of 10-12 people.

4. Determination of expert weights that deviate from the main range of sample values

For example, the following values were obtained for the first experiment:
The median of the data set (Q2) is 0.28
The lower quartile (Q1) is 0.22
The upper quartile (Q3) is 0.3325
Interquartile range Q3 - Q1 = 0.1125
Determine internal limits $0.3325 + 0.1125 \times 1.5 = 0.50125$; $0.22 - 0.1125 \times 1.5 = 0.05125$

In our case, none of the calculated values of the weights exceeds the internal limits. In the case of such a situation, it is necessary to determine whether the number out of the range is a significant outlier.

To do this, determine the outer limits of the data set $0.3325 + 0.1125 \times 3 = 0.67$; $0.22 - 0.1125 \times 3 = -0.1175$

The determination of whether an outlier should be excluded from the data set must be based on a set of reasons. An outlier may not necessarily be a measurement error (and should be excluded), but may be related to new information or a trend and should be accounted for in the calculations.

It is also important to assess the degree of influence of the outliers on the median of the data set (its distortion), if the deviation of the median is not significant, then the outlier can be included in the data sample.

5. Calculation of interfaces estimation using expert-heuristic method. Experiment 2

To confirm the hypothesis that the evaluation will be performed with greater accuracy and a smaller number of outliers, it was decided to conduct a second experiment with a smaller (11 people) number of experts.

Table 4
Obtained values of expert weights

Expert №	w_j	% in the total estimate
Expert 1	0,2925	8,087930319
Expert 2	0,24	6,636250518
Expert 3	0,28	7,742292272
Expert 4	0,35	9,677865339
Expert 5	0,28	7,742292272
Expert 6	0,54	14,93156367
Expert 7	0,35	9,677865339
Expert 8	0,385	10,64565187
Expert 9	0,25	6,912760957
Expert 10	0,22	6,083229642
Expert 11	0,429	11,8622978

The following values were obtained for the second experiment:

The median of the data set (Q2) is 0.2925

The lower quartile (Q1) is 0.25

The upper quartile (Q3) is 0.385

Interquartile range Q3 - Q1 = 0.135

Determine internal boundaries 0.385 + 0.135 × 1.5 = 0.5875; 0.25 - 0.135 × 1.5 = 0.0475 Thus, in our case, none of the calculated weights exceeds the internal limits.

Let's calculate the outer bounds of the data set to determine the weighting thresholds 0.385 + 0.135 × 3 = 0.79; 0.25 - 0.135 × 3 = -0.155

After forming a sample of experts and calculating weighting coefficients (Table 4), it was proposed to evaluate 5 groups of interfaces. The results of the evaluation are presented in Figure 6-Figure 10:

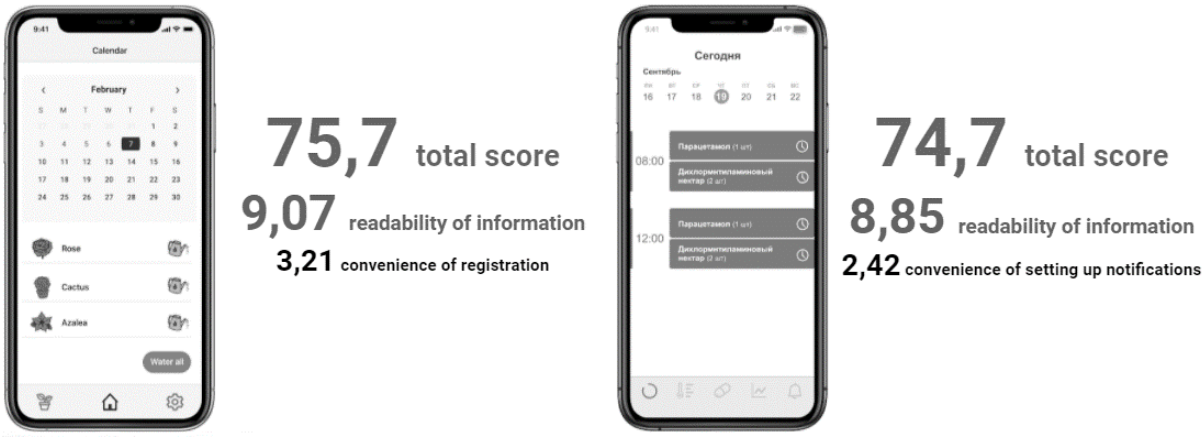


Figure 6: Group A interfaces - smart reminders (left - reminder to water, right - medication reminder)

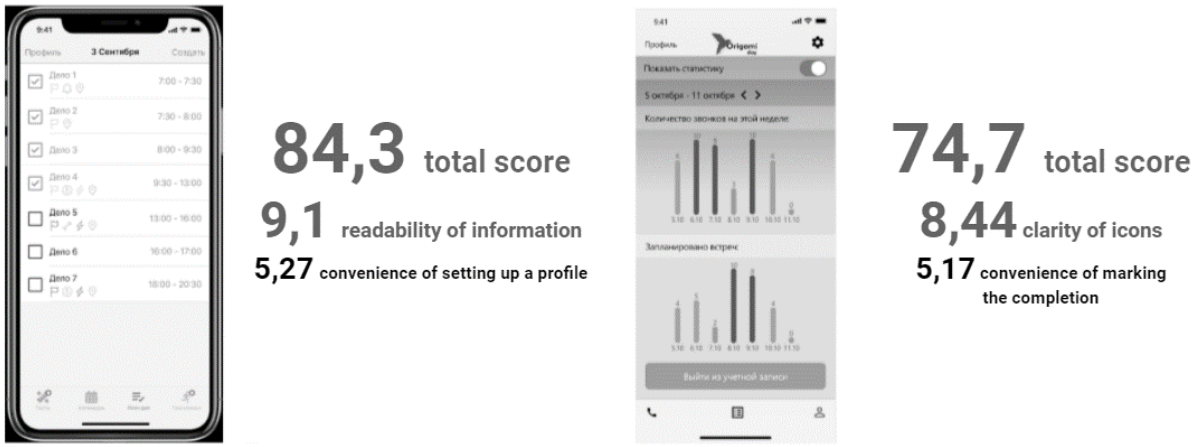


Figure 7: Group B interfaces - smart schedulers (left - task scheduler, right - meeting planner)

76,6 total score

8,62 color scheme for the target audience

4,78 convenience of registration

71,1 total score

8,06 viewing a list of attractions

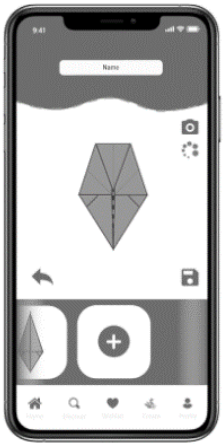
2,13 convenience of setting up a profile

Figure 8: Group C interfaces - tours and attractions (left - interesting places of the city, right - interesting city tours)

82,7 total score

9,11 integrity of the composition

6,67 convenience of setting up a profile



74,9 total score

8,85 color scheme for the target audience

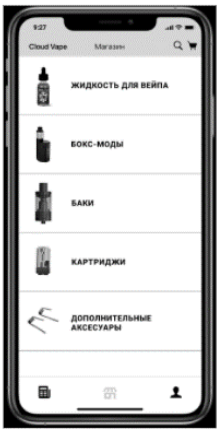
4,13 ease of level selection

Figure 9: Group D interfaces - games and simulators (left – game, right - origami simulator)

78,9 total score

9,12 color scheme for the target audience

2,77 convenience of setting up a profile



76,5 total score

8,59 viewing the shopping cart

5,07 convenience of saving in bookmarks

Figure 10: Group F interfaces - stores (left - bag store, right - vape shop)

6. Comparative analysis of the developed method with previously studied methods

Let's consider the most well-known methods for assessing interfaces and their applicability (Table 5)

Table 5
Comparative characteristics of methods for assessing interfaces

Comparison criterion	New method	Focus group method	Expert evaluation	GOMS	Game method
Number of features considered	More than 10 (depending on number of heuristics)	No well-defined criteria (what the group will notice)	More than 10 (depending on customer requirements)	1 (specific functionality)	No well-defined criteria
Ability to formalize	Yes	No	Partial	Yes	No
Difficulty of evaluation	Medium	Medium	High	Medium	High
Necessity of a ready-made interface	Not necessary (a prototype is possible)	Desirable (for final iterations)	Desirable (for final iterations)	Not necessary (a prototype is sufficient)	Desirable (for ease of experiment)
Degree of subjectivity of evaluation	Low	High	Medium	Low	High
Number of people to evaluate	11	7-9	From 1	1	2 (moderator and player)
Consideration of user experience	Partly (if the sample of experts includes ordinary users)	Partly (if the sample of experts includes ordinary users)	No	No	Yes

Thus, the developed method in the aggregate is more universal (in terms of the number of considered parameters), easy to implement and formalize (due to the simplicity and clarity of the mathematical apparatus).

Further development of the method presupposes its formalization on the basis of a web application and the creation of a system for developing recommendations for improving the analyzed interfaces. To date, a simulated layout of the service has been implemented using Google-services (https://sites.google.com/view/evalui).

7. Conclusion

The following patterns were revealed as a result of the experiment:

- The overall score is higher when there is greater consistency among the experts, i.e., the lowest variance of the estimates
- The overall score is higher with a smaller degree of difference in the weight coefficients of the experts in the group
- When the number of experts decreased from 14 to 11, the quality of the expertise increased (the experts' evaluations differed less numerically)
- The overall heuristic score does not correlate with individual subjective preferences

Thus, this evaluation algorithm allows the maximum leveling of distrust of the expert due to the elaborate system of ranking of experts, and the formation of a general assessment of the interface is performed taking into account the degree of importance of this criterion in the overall grading system.

The results of the experiments allow us to draw conclusions about the applicability of the developed method for the evaluation of interfaces. The chosen mathematical apparatus is suitable for calculating the computational characteristics of the expert weights and the evaluation itself. In the future it is necessary to develop heuristics for different categories, also more detailed elaboration of the expert evaluation criteria for more accurate determination of the expert weights is possible.

8. References

[1] Wezom IT Company, How much does it cost to create a website - the price of website development 2021, 2021. URL: https://wezom.com.ua/blog/skolko-stoit-sozdat-sajt#cena-sajta-pod-klyuch-v-zavisimosti-ot-ehtapa.html. (in Russian).

[2] U. I. Gulyaeva, Formation of a group of experts with an expert-heuristic method for evaluating interfaces," in Proceedings of the XXVII International Scientific and Technical Conference Information Systems and Technologies IST-2021, NNSTU, Nizhny Novgorod, 2021. (in Russian).

[3] Academic, 2021, URL: https://academic.ru.html.

[4] Solutions Factory, User friendly, 2021. URL: https://www.glossary-internet.ru/terms/U/user_friendly.html. (in Russian).

[5] V.M. Gorbunov, The theory of decision-making: a textbook, Tomsk: National Research Tomsk Polytechnic University, 2010., pp. 37-43. (in Russian).

[6] A. Kryanev, S. Semenov, On the question of the quality and reliability of expert assessments in determining the technical level of complex systems, Functional reliability. Theory and Practice, volume 4, 2013, pp.90-109. (in Russian).

[7] G. Bobrovnikov, A. Klebanov, Complex forecasting of the creation of new technology, Moscow, 1989, p.205. (in Russian).

[8] V. Glushkov, On forecasting based on expert assessments, Science Studies. Forecasting. Informatics, 1970. (in Russian).

[9] V. Glushkov, Methods of program forecasting of the development of science and technology, Moscow: State University. USSR Soviet Ministry Committee on Science and Technology, 1971, p.270. (in Russian).

[10] G. M. Dobrov, Yu. V. Yershov, E.I. Levin L. P. Smirnov, V. S. Mikhalevich, (Ed), Expert assessments in scientific and technical forecasting, Kiev: Nauka. dumka, 1974, p.160. (in Russian).

[11] G. Shishkova, Management (Management decisions): Educational and methodological module, Moscow: Ippolitov Publishing House, 2002. (in Russian).

[12] R. Jeffries, J. R. Miller, K. Wharton, K. M. Ujeda, Evaluation of the interface in the real world: a comparison of four methods, Hewlett-Packard Laboratories, Chicago, 1991.

[13] C. Silva, V. Macedo, R. Lemos, M. Okimoto, Evaluating Quality and Usability of the User Interface: A Practical Study on Comparing Methods with and without Users, Design, User Experience and Usability. Theories, Methods and Tools for User Interface Design, volume. 8517, 2014, DOI:10.1007/978-3-319-07668-3_31.

[14] How to calculate outliers, URL: https://ru.wikihow.com/%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B8%D1%82%D1%8C-%D0%B2%D1%8B%D0%B1%D1%80%D0%BE%D1%81%D1%8B.html

[15] V. Zeng, Assessment of the quality of designing user interfaces of a new generation, News of TulSU. Technical Sciences, volume 12, 2019 pp.404-410. (in Russian).

[16] A.Kazaryan, How to conduct a heuristic assessment of usability, Designmodo Inc., New York, 2014.

[17] A. Ballav, Nielsen Heuristic assessment: Limitations in Principles and Practice, User Experience Magazine, volume 4, 2017.