JAVOK: Jurnal Akademisi Vokasi
ISSN: 2986-1578
Juni, 2023
vol.2 no.1

Exploring Two Methods of Usability Testing: System Usability Scale And Retrospective Think-Aloud

I Wayan Pio Pratama[1*], Ni Komang Rai Mirayanti[2], I Putu Eka Sudarsana[3], Angling Galih Cahaya. Widiyanto.[4]

[1]D-III Information Technology, Politeknik eLBajo Commodus, Labuan Bajo, Indonesia, pio.pratama@poltekelbajo.ac.id
[2]Computer Science, Ganesha University of Eduation, Singaraja, Indonesia, komangdata1@gmail.com
[3]D-III Information Technology, Politeknik eLBajo Commodus, Labuan Bajo, Indonesia, ekasudarsana@poltekelbajo.ac.id
[4]D-III Information Technology, Politeknik eLBajo Commodus, Labuan Bajo, Indonesia, anglingwidyanto@poltekelbajo.ac.id

Abstract

This research was conducted to measure the usability of a system using two methods, RTA (Retrospective Think Aloud) and SUS (System Usability Scale). An object used in this research was shazam. The subject were 5 people, who are smartphone users in the age range of 10 to 50 years old. Finally, in this research RTA method showed that the shazam application is effective with a completion rate of 94.3% and efficient with a time-based efficiency of 0.108 task/second. While SUS showed that Shazam is still categorized as a class B application.

Keywords: RTA, SUS, Completion rate, Time-based efficiency.

INTRODUCTION

International standard ISO 9241 pt. 11 (ISO, 1998) defines usability as the degree to which a specific user base can achieve certain goals with a product, focusing on effectiveness, efficiency, and satisfaction in a particular context of use. In this case, effectiveness is the user's ability to complete their tasks, efficiency refers to the speed at which tasks can be accomplished, and satisfaction relates to the user's enjoyment during their interaction with the product (Sauro & Lewis, 2016).

Good usability can generate recommendations for a product, while poor usability can deter potential customers. Unlike other market research data, usability examines user attitudes and actions. Any element in the user interface, such as buttons, labels, designs, or layout, impedes users from accomplishing their tasks, makes them challenging, or increases task times, which constitutes a usability issue (Sauro, 2015). While there are no rigid guidelines for gauging effectiveness, efficiency, and satisfaction, a large survey of nearly 100 summative usability tests (Lewis & Sauro, 2009) provides insights into common practices. The majority of tests typically include a blend of metrics such as completion rates, errors, task times, task-level satisfaction, test-level satisfaction, help access, and a catalog of usability problems (with frequency and severity often included). In the context of the RTA setup, the Recording Tool generated two separate records: one documenting the user's engagement with tasks in the application, and another capturing the participants' introspective observations as they assessed their task completion (Lewis & Sauro, 2009).

One of the most effective methods to uncover usability issues is to perform a usability test, which involves observing users as they engage with software, hardware, or a website to identify any difficulties they encounter. Although the temptation might be to

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

assist participants or inquire about their preferences, the objective of a usability test is to monitor participants without influencing their experience. There are established and trustworthy approaches for determining how design impacts the usability and the necessary alterations to enhance a product's usability for market success. In this study, two usability testing methods were employed, namely the System Usability Scale (SUS) and Retrospective Think-Aloud (RTA).

LITERATURE REVIEW

Retrospective Think-Aloud (RTA)

An application is considered usable if its functions can be utilized effectively, efficiently, and satisfactorily (Agustini et al., 2019). Usability evaluation is a process involving users to understand and use a product, aiming to optimize user comfort aspects such as effectiveness, efficiency, and overall system satisfaction (U.S. Department of Health & Human & Services, 2014). According to Zaphiris & Kurniawan, usability evaluation methods can be categorized into model/metrics-based, inspection, inquiry, and testing. Among these categories, usability testing has become the most commonly employed due to its accuracy (Sadewa et al., 2021).

During usability testing, some users naturally express their thoughts and experiences. Requesting users to verbalize their thoughts in real time can offer valuable insights into their perception of the product or system and its alignment with the design's intent. However, users' accounts may not always be comprehensive, as they might consciously or unconsciously omit certain aspects. It's important to refrain from asking users to think aloud when timing tasks, as this can significantly slow down performance. An alternate approach is the RTA method. This method guides users to recollect their thoughts and actions post-task completion. During a stimulated RTA, users articulate their thought process while watching a screen recording of their interaction with the system. This approach ensures that the user's task performance is not directly impacted.

During a usability test session, it's common for some participants to naturally articulate their thought processes. Encouraging participants to vocalize their thoughts throughout the session can offer valuable insights into their understanding of the product or system and its alignment with the intended design. Nonetheless, participants often selectively share their experiences and thoughts, either consciously or unconsciously, leading to potential gaps in the collected data. It's crucial to refrain from requesting participants to 'think aloud' during time-sensitive tasks as verbalizing thoughts can substantially impede performance.

The Retrospective Think Aloud (RTA) method presents an effective alternative. This method involves guiding users to recollect and articulate their thoughts and actions post-completion of the predefined task(s). In a stimulated RTA process, participants verbalize their thoughts while viewing a screen recording of their interaction with the system under review. This method ensures that the participant's task performance remains free of direct interference.
An alternative to having participants RTA method. The RTA method instructs users to recall their thoughts and actions after they have finished hed predefined task(s). During stimulated RTA, subjects verbalize their thoughts while reviewing a screen recording of their performance while they interact with the system under study. In this way, no direct interference of a subject’s task performance occurs.

System Usability Scale (SUS)

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

The process of measuring usability involves both observing customers using a product and their attitude toward it. At times, a broad indication of a system's usability in comparison to its competitors or prior versions may be all that's necessary. When choosing metrics, measures that don't involve extensive effort and cost for data collection and analysis are often preferable. The selected measure needs to be easy and swift to apply, yet reliable enough to compare user performance shifts across different versions of a software product. Consider a scenario where customers are asked to participate in evaluation exercises lasting between 20 minutes and an hour, culminating in a subjective assessment of system usability. After sucperiodriod, particularly if they faced difficulties without assistance, users could be highly frustrated. If they were then asked to answer an exhaustive questionnaire exceeding 25 questions, it's likely they might not complete it, resulting in inadequate data to gauge subjective responses to system usability.

In response to these requirements, a simple usability scale was developed by John Brooke in 1986. It consists of ten items scale giving a global view of subjective assessments of usability. During the test, respondents are asked to give ratings on these items. Each item has a rate from Strongly Disagree (1) to Strongly Agree (5). SUS employs a Likert scale. There's a common belief that a Likert scale merely incorporates forced-choice queries, in which a statement is presented and the respondent marks their level of agreement or disagreement with the statement on a point scale. Nonetheless, the formulation of a Likert scale is more nuanced than this portrayal (Brooke, 2020).

Given such statements, there will generally be somewhere there are agreements between respondents. Moreover, some of these statements might evoke strong levels of agreement or disagreement from all respondents. The goal is to pinpoint such statements for incorporation into a Likert scale, as, if appropriate examples are chosen, there should be a consensus on extreme attitudes toward them.

METHOD

Test Object

This research focuses on the usability of Shazam, an application developed and maintained by Apple Inc. The primary function of this application is to identify a wide array of audio content, including music, movies, advertisements, and television shows, through brief audio samples captured via the device's microphone. Upon tagging an audio snippet for a duration of approximately 10 seconds, the application generates an "audio fingerprint." This process involves the analysis of the captured sound sample, which is then matched against an extensive database containing millions of songs. If the application successfully locates a corresponding match, it delivers essential information to the user, such as the song's artist, title, album, and lyrics.

Shazam's functionality extends beyond merely identifying songs. Certain iterations of the application incorporate hyperlinks to various relevant platforms, including iTunes, Spotify, YouTube, and Groove Music, thus providing users with additional context or access to the identified music. It's noteworthy that Shazam can identify music from any source, assuming that the level of background noise is not excessive to the point of inhibiting the creation of an acoustic fingerprint and that the song exists within the application's comprehensive music database.

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

Participant

A common misconception exists regarding the necessity of large sample sizes (typically over 30) for reliable statistical analysis and interpretation of quantitative data. In reality, even small samples can provide significant insights. A frequent concern related to small sample sizes is that they may not adequately represent the broader population. However, sample size and representativeness are distinct concepts. For instance, a sample of five individuals could represent the population accurately, while a sample of 1,000 may not. The issue lies not with the sample size but rather with its representativeness.

In user research, regardless of whether the data is qualitative or quantitative, the primary objective is to ensure that the participant sample accurately represents the population we wish to study. If this criterion is not met, we lack the logical foundation to generalize our results from the sample to the population. No amount of statistical adjustment can justify drawing inferences about one population based on observations from a different one. For example, to improve the design of snowshoes, a sample of five Arctic explorers provides more valuable insights than a sample of 1,000 surfers.
The confusion between sample size and representativeness often arises when our population comprises diverse groups, and our sample size isn't sufficient to represent each one. To tackle this, we should develop a sampling plan ensuring that a representative sample is drawn from every group under study. In essence, more accurate conclusions can be derived from a somewhat non-random sample from the right population than from a perfectly random sample from an incorrect population.

This study was carried out with a sample of five participants aged between 10 and 50. At the time of the study, all were smartphone users, indicating some familiarity with using such devices. However, none had previously used Shazam, making them novice users and therefore representative of the primary target group for the application. Participants were solicited to partake in the experiment, with the only participation criteria being that they were smartphone users within the specified age range. Gender and other factors were not considered as prerequisites for participation. In conclusion, the participants, consisting of both male and female individuals of varying ages, were evenly distributed in the experiment, with no notable differences in gender, age, or prior experience with Shazam.

Experimental Process

In this research, we collected usability data by observing and interacting with users as they attempted tasks.

TASK LIST

All task must be done using shazam app!

No	Task	Max Time (seconds)
1	Find the title of the music you are listening!	60
2	Get the lyric of the song in task 1!	60
3	Get youtube video of the song in task 1!	60
4	Find song that similar in task 1!	60
5	Sign in using facebook account!	60
6	Search song in discovery menu!	60
7	Find playlist Menu!	60

Figure 1. The task form is given in the research

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

In order to evaluate the shazam application by means of the retrospective think-aloud protocols, some tasks were formulated. Retrospective think-aloud protocols, some tasks were formulated. All tasks were designed to manifest all main features in the application.

Participants received usability testing tasks in paper forms as shown in Figure 1 and Figure 2. They were instructed to carry out the tasks in support of the tool without the assistance of the facilitator (silent condition).

REVIEW FORM

Name : Occupation : Age :

No	Task	Time (seconds)	Result	Comment
1	Find the title of the music you are listening!			
2	Get the lyric of the song in task 1!			
3	Get youtube video of the song in task 1!			
4	Find song that similar in task 1!			
5	Sign in using facebook account!			
6	Search song in discovery menu!			
7	Find playlist Menu!			

Note : ✓ : success, × : failed

Moderator

……,…/……/……

(……)

Participant

……,…/……/……

(……)

Figure 2. The review form is given in the research

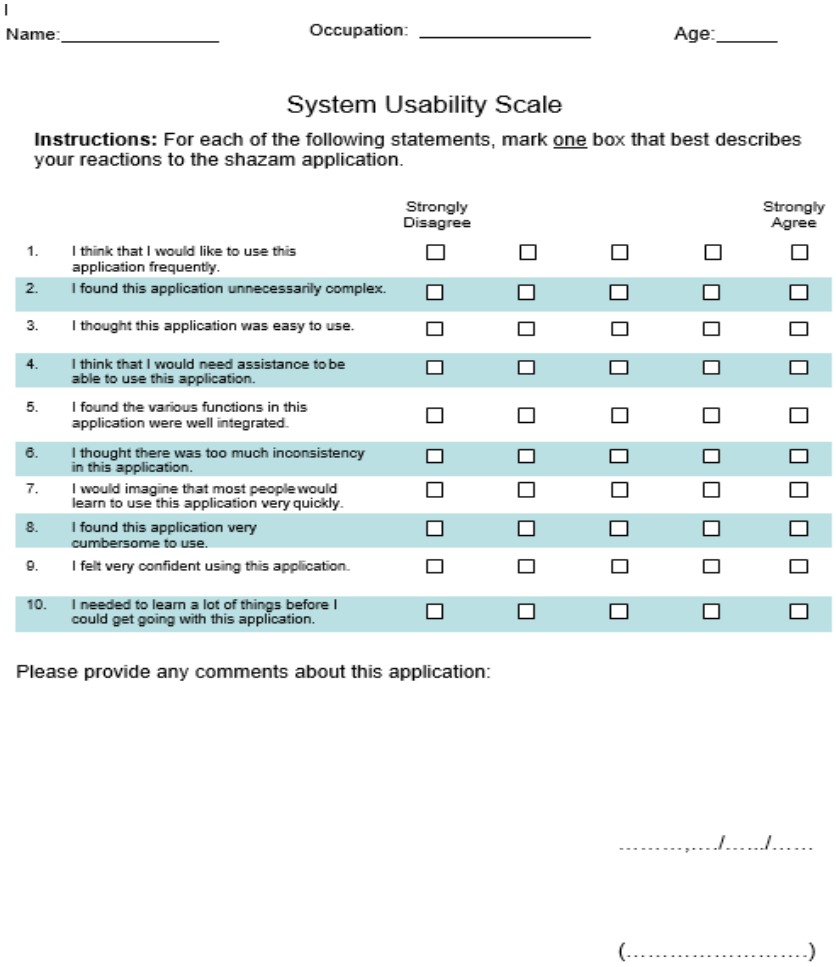
Directly after the subjects were finished, the silent test condition was stopped. Participants were initially given a task to practice articulating their thoughts aloud. Following this, they were required to express their thoughts retrospectively while watching a video recording of their interaction with the system, thus creating retrospective verbal reports. As per the standard RTA procedure, participants could halt the recording whenever they wished to elaborate on the observed actions, and the pause duration was included in the task measurement time. Participants were not permitted to rewind or review the video. Prior to viewing the task performance, participants were presented with a written form of the task to assist with recalling the task, aiming to mitigate any impact of memory loss on the RTA protocols.

Under the RTA scenario, the Recording Tool created two distinct documents: one chronicling the participant's task execution within the software, and the other capturing the participants' reflective commentary while they reviewed their performance on the tasks. In order to measure the efficiency and effectiveness of the application, the time and the success of each participant in completing the given tasks will be measured. Finally, the RTA method provided quantitative data. At the end of the RTA test, participants will be asked to explain while they take a look at the recording of what they

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

have done during the task session. This session provided qualitative data in terms of comments.

After the think-aloud reviewing section, ten SUS questionnaires as shown in Figure 3, will be given by the moderator. Each of the given questions will relate to the participant's satisfaction.



Name:_____ Occupation: _____ Age:_____

System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the shazam application.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this application frequently.	☐	☐	☐	☐	☐
2.	I found this application unnecessarily complex.	☐	☐	☐	☐	☐
3.	I thought this application was easy to use.	☐	☐	☐	☐	☐
4.	I think that I would need assistance to be able to use this application.	☐	☐	☐	☐	☐
5.	I found the various functions in this application were well integrated.	☐	☐	☐	☐	☐
6.	I thought there was too much inconsistency in this application.	☐	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this application very quickly.	☐	☐	☐	☐	☐
8.	I found this application very cumbersome to use.	☐	☐	☐	☐	☐
9.	I felt very confident using this application.	☐	☐	☐	☐	☐
10.	I needed to learn a lot of things before I could get going with this application.	☐	☐	☐	☐	☐

Please provide any comments about this application:

……,…./……/……

(……)

Figure 3. SUS questionnaires given in the research

It consists of ten items scale giving a global view of subjective assessments of usability. Each item has a rate from Strongly Disagree (1) to Strongly Agree (5). During the test, respondents are asked to give a rating on these items. If the participant strongly disagrees with the statement written in an item, they should fill in the strongly disagree column with a check mark (√). The purpose of giving these questionnaires is to measure the satisfaction of each participant. In other words, SUS acts to evaluate the results obtained in the RTA process.

RESULT AND DISCUSSION

Efficiency

Efficiency is measured using time-based efficiency. One formula that we can use to calculate this is using this formula (Mifsud, 2015).

$$Time\ Based\ Efficiency = \frac{\sum_{j=1}^{R}\sum_{i=1}^{N}\frac{n_{ij}}{t_{ij}}}{NR}$$

Where:

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

N = The total number of tasks (goals)
R = The number of users
n_{ij} = The result of task i by user j; if the user successfully completes the task, then n_{ij} = 1, if not, then n_{ij} = 0
t_{ij} = The time spent by user j to complete task i.
This measurement is used to measure the speed of the user to find information that is needed from the application.

Time-Based Efficiency
Given data by participants like the below table.

Table 1. Task Taken Time Measurement Result

Task	t_{ij} (second)
1	6.09
2	5.47
3	5.5
4	10.29
5	-
6	4.63
7	3.40

Hence, we can calculate $\frac{n_{ij}}{t_{ij}}$

Table 2. Task Taken Time Measurement Ratio

Task	t_{ij} (second)	n_{ij}	$\frac{n_{ij}}{t_{ij}}$
1	6.09	1	0.164
2	5.47	1	0.183
3	5.5	1	0.182
4	10.29	1	0.097
5	-	0	0
6	4.63	1	0.216
7	3.40	1	0.294
Total			1.136

For all users, the calculation is displayed in the below table.

Table 3. Task Completion Effectivity

Participant	Task Solved	$he\sum_{i=1}^{N}\frac{n_{ij}}{t_{ij}}$
1	7	0.79
2	6	0.35
3	7	0.86
4	6	1.14
5	7	0.64
$\sum_{j=1}^{R}\sum_{i=1}^{N}\frac{n_{ij}}{t_{ij}}$		3.78
$\frac{\sum_{j=1}^{R}\sum_{i=1}^{N}\frac{n_{ij}}{t_{ij}}}{NR}$		$\frac{3.78}{7x5} = 0.108$

Time-based efficiency is to measure speed level of each user to solve the task using Shazam. This number is quite fast to solve 7 tasks even there is a gap between participants in order to solve the tasks. Shazam reaches 0.108 tasks per second, this

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

means Shazam is efficient for all age ranges since almost all users have no problem getting information from Shazam.

Effectiveness

Effectiveness can be determined by the proportion of participants who successfully finish all tasks. If a participant successfully completes a task, they are assigned a value of 1, while a value of 0 is given if the task is not completed (according to ISO/IEC 9126-4). Let's denote the percentage of participants who successfully finish all tasks as 'P'. This ratio can be computed using a specific formula.

$$P = \frac{Number\ of\ tasks\ successfully\ completed}{total\ number\ of\ tasks} \times 100\%$$

Completion Rate

Usability metrics fundamentally consist of measures like achievement rates, often known as completion rates. These are generally captured as a binary data point representing task accomplishment (marked as 1) or task non-completion (marked as 0). Achievement rates for a given task are calculated by determining the ratio of participants who effectively accomplished the task to the total number who attempted it. To illustrate, if a task is successfully executed by 8 out of 10 participants, the achievement rate stands at 0.8, typically conveyed as 80%. By deducting the achievement rate from 100%, we can express a non-completion rate, in this case, 20%. What's more, these binary rates are commonly used across scientific and statistical studies.

Table 4. Participant Successful Rate

Task	Task Solved	Number of Participants	Percentage of success
1	5	5	100%
2	5	5	100%
3	5	5	100%
4	5	5	100%
5	3	5	60%
6	5	5	100%
7	5	5	100%
Average			94.3%

The average of shazam is high, almost every task can be done by all users. Based on ISO/IEC 9126-4 said that a system is effective if it reached 78% when the user solves the tasks. Because shazam reach 94.3% that means shazam is effective, even 2/5 of users failed to solve task number 5.

SUS

SUS is calculated following the below algorithm.

For odd items, subtract 1 from the user response.

For even-numbered items, subtract the user responses from 5. This scales all values from 0 to 4 (with 4 being the most positive response).

Add up the converted responses for each user and multiply that total by 2.5. This converts the range of possible values from 0 to 100 instead of from 0 to 40.

Average together the scores for all participants.

Using that algorithm we calculate the data for participant number 1.

Table 5. SUS Raw Data

No	Raw	Scale
1	2	1
2	2	3

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni

3	3	2
4	2	3
5	4	3
6	3	2
7	3	2
8	4	1
9	3	2
10	2	3
	Total	22
	SUS Score	55

Doing the same calculation for other participants so we get below table.

Table 6. SUS Participant Score

Participant	SUS Score
1	65
2	77,5
3	90
4	85
5	47.5
Average	73

The average SUS score from over 500 studies is 68. That means a SUS score above 68 is considered above average and anything below 68 is below average.

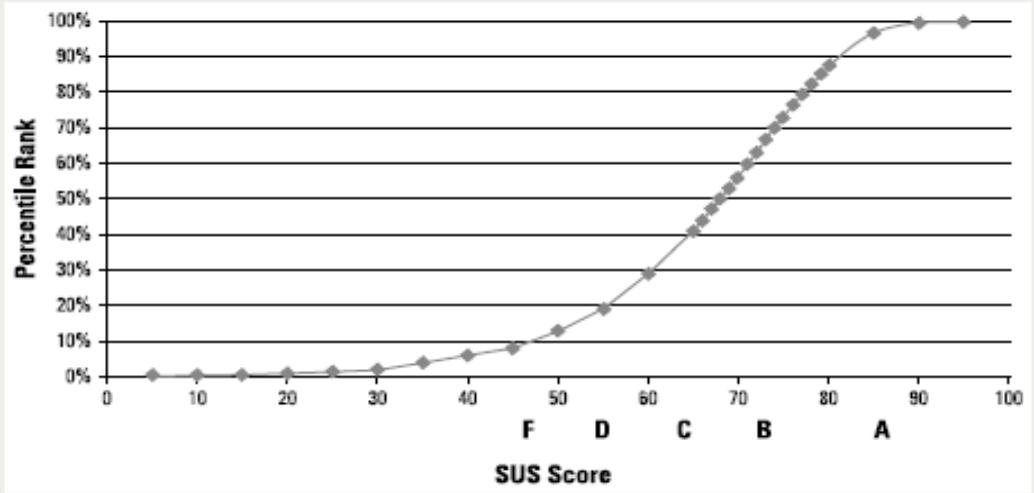


Figure 2. SUS Score Percentile Rank

From this curve shazam will get a B based on the SUS method.

DISCUSSION

This method showed that some participants found it difficult in completing task 1 and task 5. Task 1 is the main task that asked participants to find the title of the song that played during the test. It is because there is no intuitive information, especially for the main icon of this application. The main icon of this application is the big icon in the center of the system, as shown in Figure 4, it is used to find the title of the song that played during the test. So, some participants suggest that the icon in the middle of the application be changed to something more informative.

In addition, there is a difficulty in completing task 5. Task 5 asked participants to sign in using Facebook account. In this task, participants found it difficult to find the login menu. It is because the position of the login menu is hidden and the user needs to search until they find it.

This method showed that some participants completed their tasks quickly and precisely. Unfortunately, some participants found themselves not interested in the Shazam application. But overall indicates that RTA and SUS give positive correlation for shazam application. The majority of participants agree that Shazam is easy to use even doesn't guarantee all participants will often use this application.

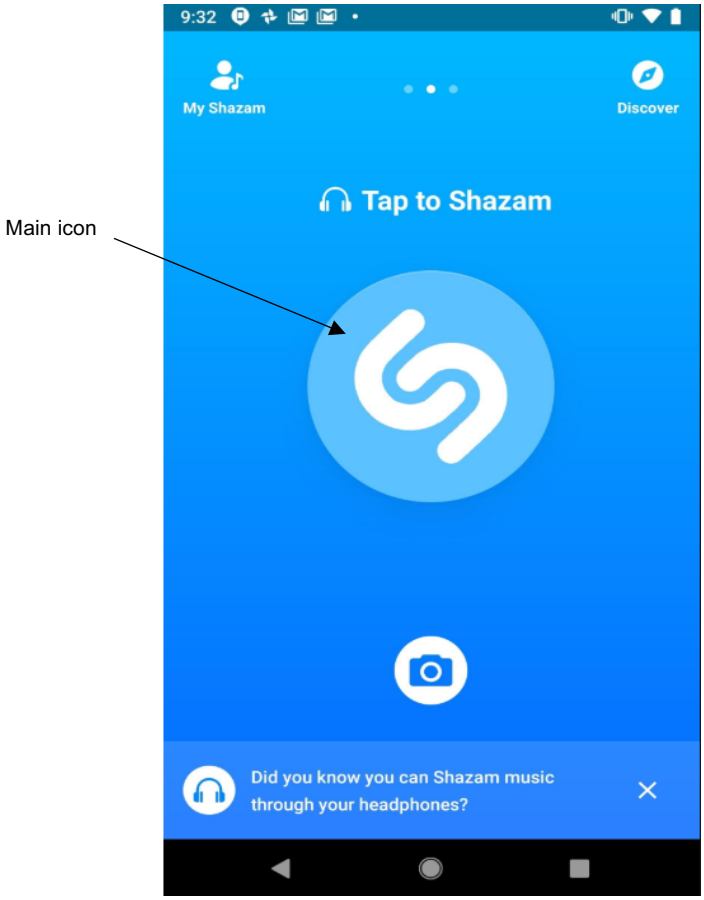


Figure 3. Shazam's appearance on Android

CONCLUSION

This study was conducted to evaluate the usability of Shazam, a widely used application, employing two distinct methods - RTA and SUS. A diverse group of five smartphone users, ranging in age from 10 to 50 years, were the subjects of this study. The results gathered through the RTA method highlighted Shazam's effectiveness, with a commendable completion rate of 94.3%, and efficiency, demonstrated by a time-based efficiency rating of 0.108 task/second. Meanwhile, the SUS evaluation classified Shazam as a class B application. Therefore, it can be concluded that Shazam demonstrates robust usability performance, offering both efficiency and effectiveness to its users. Nevertheless, there remains room for improvement, as indicated by its SUS classification, to elevate the user experience further.

REFERENCES

Agustini, K., Studi, P., Teknik, P., & Ganesha, U. (2019). *Usability testing*. *8*(1), 12–22.

Brooke, J. (2020). SUS: A "Quick and Dirty" Usability Scale. *Usability Evaluation In*

Industry, *November 1995*, 207–212. https://doi.org/10.1201/9781498710411-35

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5619 LNCS*(August), 94–103. https://doi.org/10.1007/978-3-642-02806-9_12

Sadewa, I. G. B. B., Divayana, D. G. H., & Pradnyana, I. M. A. (2021). Pengujian Usability Pada Aplikasi E-Sakip Kabupaten Buleleng Menggunakan Metode Usability Testing. *INSERT : Information System and Emerging Technology Journal*, *1*(2), 76. https://doi.org/10.23887/insert.v1i2.25975

Sauro, J. (2015). *Customer Analysis for Dummies*. John Wiley & Sons.

Sauro, J., & Lewis, J. R. (2016). Quantifying the User Experience: Practical Statistics for User Research, Second Edition. In *Quantifying the User Experience: Practical Statistics for User Research, Second Edition*.

U.S. Department of Health & Human, & Services. (2014). *Usability evaluation basics*. http://www.usability.gov/what-andwhy/%0Ausability-evaluation.html

 diterima: 19 bulan Juni direvisi: 22 bulan Juni diterbitkan: 30 bulan Juni