proceedings

Proceedings

User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana [†]

Ana Berdasco *, Gustavo López, Ignacio Diaz, Luis Quesada and Luis A. Guerrero

University of Costa Rica, 11501-2060 San José, Costa Rica; gustavo.lopez_h@ucr.ac.cr (G.L.); ignacio.diaz@ucr.ac.cr (I.D.); luis.quesada@ecci.ucr.ac.cr (L.Q.); luis.guerrero@ecci.ucr.ac.cr (L.A.G.)

* Correspondence: ana.berdasco@ucr.ac.cr; Tel.: +506-8334-3683

† Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.

Published: 20 November 2019

Abstract: Natural user interfaces are becoming popular. One of the most common natural user interfaces nowadays are voice activated interfaces, particularly smart personal assistants such as Google Assistant, Alexa, Cortana, and Siri. This paper presents the results of an evaluation of these four smart personal assistants in two dimensions: the correctness of their answers and how natural the responses feel to users. Ninety-two participants conducted the evaluation. Results show that Alexa and Google Assistant are significantly better than Siri and Cortana. However, there is no statistically significant difference between Alexa and Google Assistant.

Keywords: intelligent personal assistant; natural user interfaces; user experience; Google Home; Amazon Alexa; Apple Siri; Microsoft Cortana

1. Introduction

A natural user interface (NUI) is a system for human–computer interaction that the user operates through intuitive “invisible” actions. The goal of these interfaces is to hide the complexity of the system even if the user is experienced or the interactions are complex. Examples of the actions commonly utilized by NUI include touch and gestures. In more recent years, a new generation of voice-powered personal assistants has become common and widespread. These assistants were pioneered and commoditized by Apple when they introduced Siri in the iPhone in 2011 [1].

Even though intelligent personal assistants are now mainstream, evaluating these assistants represent a challenge due to the large variety and number of tasks they support. For example, the assistants found on the average smartphone supports a wide range of tasks, such as voice commands, web search, chat, and several others [2]. Due to the number of tasks that use voice commands, studies that attempt to measure the effectiveness of these assistants or compare them tend to focus on a small number of assistants and are targeted to a narrow field of usage scenarios in which authors perform measurements by themselves (for example, assistance during their day-to-day e-mail writing) [3].

This paper makes a comparison of four intelligent personal assistants (i.e., Google Assistant, Amazon Alexa, Apple Siri, and Microsoft Cortana) that have been developed to aid people in managing time commitments and performing tasks [4]. All assistants are compared based on the same aspects and services. This paper focuses on voice-activated intelligent personal assistants deployed in smartphones, smart speakers, or personal computers. All these assistants can be found on widespread devices such as Android or Apple phones as well as in Microsoft Windows [5–8].

The evaluation was conducted by 92 university undergraduate students of several different majors. Each participant evaluated all four personal assistants in two dimensions: how good were the answers, where good means how natural the responses feel to users, and how correct were the answers, where correct means free from error; in accordance with fact or truth.

The motivation of this study is to evaluate these assistants with many users, not just the personal experience of a single person. Another motivation for this study is to conduct an unbiased analysis. This is especially important because most comparisons or evaluations of personal assistants are conducted by the same companies that developed the assistants.

The rest of the work is structured as follows. Section 2 summarizes relevant previous works in the area. Section 3 describes the methodology and instruments used in this research. Section 4 presents the results and discussion of the research. Finally, Section 5 presents the conclusions and outlines future work.

2. Related Work

There are different ways in which personal assistants can be evaluated by voice; in some cases, the creators of the assistants offer an evaluation mechanism. However, rather than measuring how satisfied the users are with the assistants, they measure the capacity they have to perform specific tasks. For example, Amazon offers an evaluation guide for Alexa, where one of the tasks is to create a notification [8]. This allows evaluating the ability of Alexa to execute the task, but not the satisfaction of the user.

Many of the works that stand out in the literature are focused on the evaluation of a single assistant and the tasks that it can perform from searches and configuration notifications, among other tasks. At the same time, they point out the challenges that users may face with attendees, for example, that sometimes the user must repeat the command that was used or that integration problems with other devices may arise, among other challenges [9].

A group of researchers of the Department of Future Technologies, University of Turku, Finland, investigated the usability, user experiences, and usefulness of the Google Home smart speaker. The findings showed that Google Home is usable and user-friendly for the user [9], but the study did not include other assistants like Alexa or Cortana.

The paper "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants" is an example, which not only makes an evaluation of the tasks that the assistants offer, like sending emails and messages, among others, but also includes topics such as privacy and the problems of security that the assistants face to handle the information of the users [10].

Another study was carried out by a group of researchers from Microsoft [2] that tried to automate the evaluation of the attendees and predict the quality of voice recognition. Most of the work is in creating a model that allows evaluating the tasks supported without needing a physical person to do it, and the satisfaction is evaluated in terms of the capacity of the assistant to understand the assigned task.

On the other hand, there are also studies that not only focus on evaluating the skills of the assistants but have begun to take into account as part of the evaluation the affective experiences of the users with the assistants [11]. Yang found that the affective responses differed depending on the scenario; for example, some factors that underlie the quality are the comfort in the conversation between the machine and the man, the pride of using cutting-edge technology, the fun during use, the perception of having a human person, privacy, and the fear of distraction

One approach worth mentioning is that of the authors Lopez, Quesada and Guerrero [12]. They proposed a study in which they evaluated the answers of the assistants based on the accuracy and naturalness of the answers of the devices. This maintains the focus of evaluating the tasks that the assistants perform but also consider the quality of the user–assistant interaction. Our work is partially based on this paper, which served as a reference for the evaluation of intelligent personal assistants.

3. Methodology

3.1. Evaluation Design

The first part of the study was the identification of the voice assistants that would be evaluated by the participants, which was achieved through a literature review. The selected assistants were Siri, Alexa, Cortana, and Google Assistant [10].

After the assistants were identified, the next step was to select the scenarios that would be evaluated. A scenario in this context is defined as a task in which a person would want the assistant's help. This definition is intentionally loose to accommodate a wide range of tasks. Examples include a person requesting assistance on how to navigate from their current location to another, simple mathematical questions, and "general knowledge" questions.

The scenarios were selected with the collaboration of a group of four HCI (human computer interaction)experts, all professors at the University of Costa Rica (UCR), and it was based on previous research [12]. The evaluation was performed in two dimensions: an objective one that measures the correctness of the answer (i.e., whether it is factually accurate) and a subjective one that measures its quality (as perceived by the person interacting with the device).

The next stage was to perform an unscripted pilot, which was performed by a group of 10 participants with varied backgrounds, such as economics, computer engineering, biology, and others. The goal was to understand how they naturally interacted with the personal assistants on each scenario, with minimal guidance. They were provided only with a vague scenario, and they were asked to request the assistant to help them solve it. Interactions enabled the gathering of questions naturally asked by people to the assistants when attempting to solve the scenarios. An example of the guidance provided to the members of the pilot is: "Imagine that you want to make a sum". Each participant asked questions to the assistant in slightly different ways, such as one of them asking "How much is the sum of three plus four" while others asked "three plus four".

As part of the results of this pilot, it was identified that depending on how the question is posed, it may or may not be understood by the assistants. Therefore, a question that was understood by all the assistants had to be selected for each scenario that was going to be evaluated. This was done to guarantee that the performance of all the assistants was measured under fair and equal circumstances, in which they all understood the question being asked.

After the pilot, a video was recorded with one person asking each assistant a set of requests. Only one person participated in this recording to assure that each assistant answered the same question with the same tone and accent. Each answer was recorded, and these recordings were presented to the participants during the evaluation.

In the video, the questions were presented sequentially. Each question was presented followed by the answer provided by each one of the assistants. To guarantee the comprehension of the viewers, both the questions and the answers included the audio in English as well as a transcript (English and Spanish). Figure 1 shows an example of the presentation format. The following questions were used:

1. How does a dog sound?
2. Thirteen plus seventeen.
3. What is the speed of the light?
4. Where does Keylor Navas play?
5. Which team won the soccer world cup of Italy 90?
6. I want to play a game.
7. How many US dollars are 10,000 Costa Rican colons?
8. Who is Canada's president?
9. What is the chemical formula for water?
10. Set the alarm to six o'clock AM.

3.2. Evaluation Execution

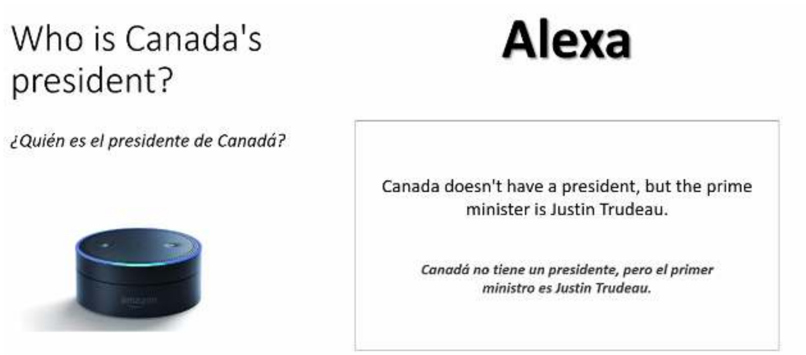
The video was presented to 92 university students, divided into five groups. These were active students from the University of Costa Rica and were aged between 18 and 26 years old at the time of the study.

All participants evaluated the quality and the correctness of the answers provided by each one of the intelligent personal assistants by responding the following two questions: “How good were the answers?” and “How correct were the answers?”. Before the video was presented, a brief explanation of the goal of the study, the video that they would see, and what was expected from them was explained. On average, each group evaluation lasted 20 min. Table 1 shows an example of the questions used by the participants to evaluate the assistants.

All participants responded using a 5-point Likert scale for goodness and correctness of the response. The scale was: (1) very poor, (2) poor, (3) average, (4) above average, and (5) excellent.

Table 1. Example of the questions for evaluating the assistants.

Question	Google Assistant	Alexa	Siri	Cortana
How good were the answers?				
How correct were the answers?				



Who is Canada's president?

¿Quién es el presidente de Canadá?

Alexa

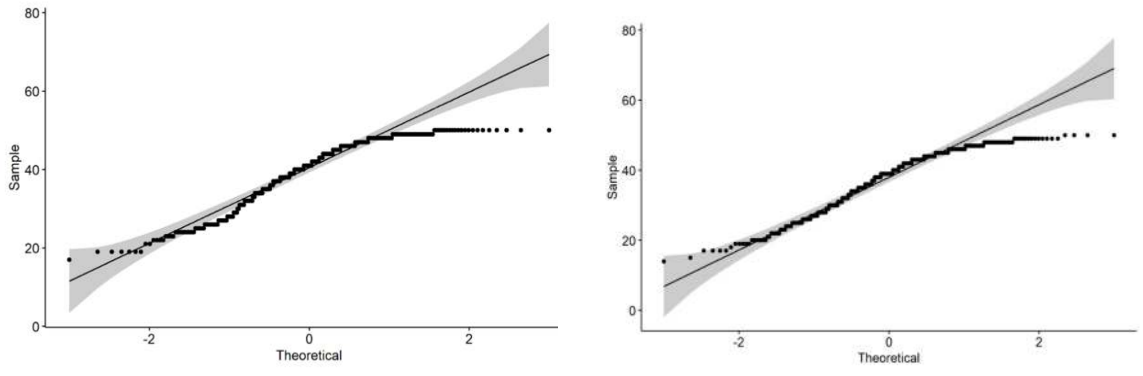
Canada doesn't have a president, but the prime minister is Justin Trudeau.

Canadá no tiene un presidente, pero el primer ministro es Justin Trudeau.

Figure 1. Examples of the video showed to participants.

3.3. Data Analysis

Each answer was considered individually (“How good were the answers?” and “How correct were the answers?”) and then grouped. To group the results, the ten scores from a single participant were added. This provides an aggregated score with a minimum value of 10 and a maximum of 50 per participant. Normality tests were conducted, and the data did not show a normal distribution. Figure 2 shows the distributions of the tests.



(**a**) (**b**)

Figure 2. Results of the Shapiro–Wilk test for normality. (**a**) Results for the responses of “How correct was the answer?” and (**b**) results for the responses of “How good was the answer?”.

The non-normality of the data prevented the use of a parametric ANOVA test to compare the means. Therefore, the Kruskal–Wallis tests, which is a non-parametric equivalent of the ANOVA tests that do not require the data to be normally distributed, was used.

To qualitatively categorize the results, values were discretized into five categories: "excellent", "above average", "average", "below average", and "poor". Since the values can have a range from 10 to 50, this range was split into five equal segments. Therefore, each one of them spans eight units. For example, the "very poor" range includes all answers between 10 and 18, while the "excellent" one includes those between 42 and 50.

4. Results and Discussion

This section describes the results of the evaluation with 92 participants. It is interesting to mention that 99% of the participants were aware of the existence of the various assistants, but only 86% had used at least one of them. The results show no differences between the preferences of women and men.

Figure 3 shows for each of the assistants the result obtained to evaluate "How good were the answers?". The best two, by a wide margin, are Alexa and Google Assistant. The latter is the best one, beating Alexa by approximately 12% in the excellent category. Figure 4 shows a comparison of the sum of the responses of the participants separating each assistant to compare based on "How correct were the answers?". The superiority of both Google Assistant and Alexa is also apparent in this figure.

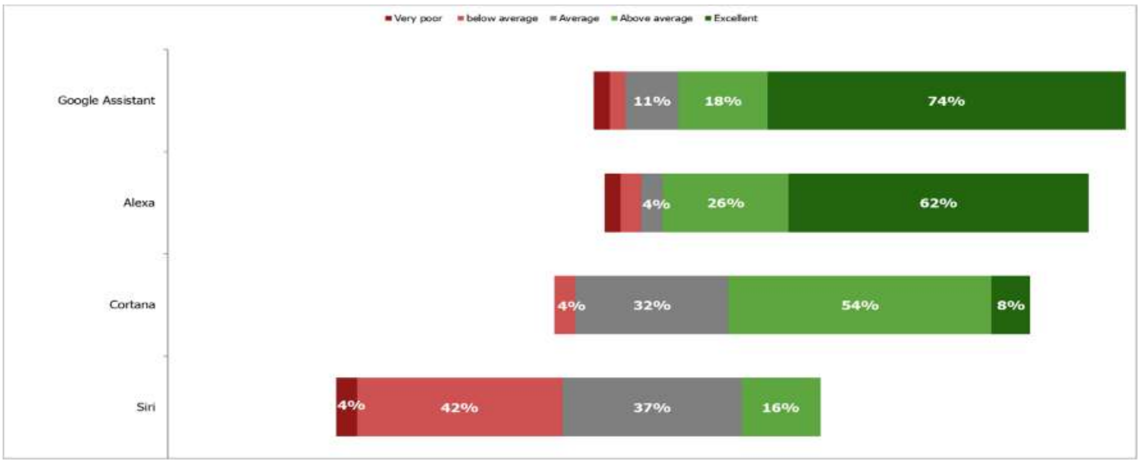


Figure 3. Results for the question "How good were the answers?".

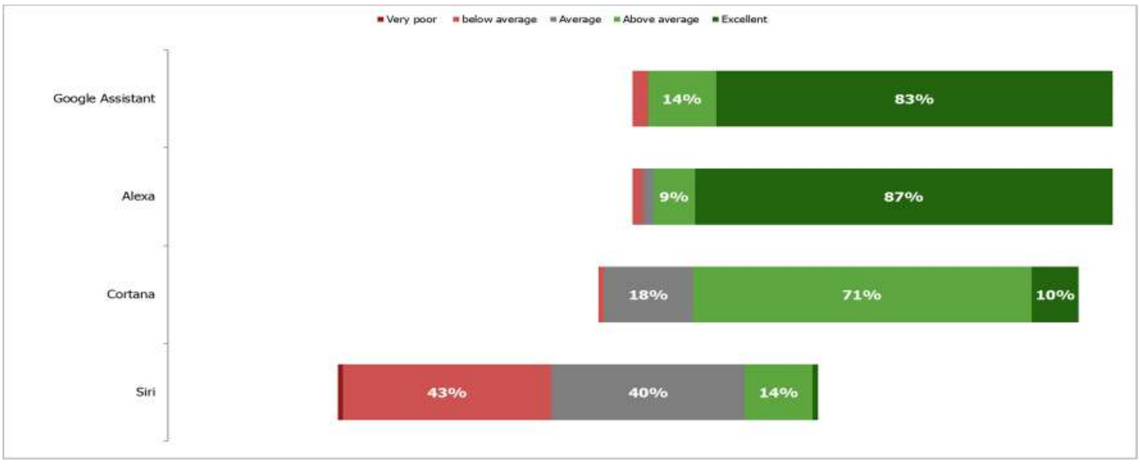


Figure 4. Results for the question "How correct were the answers?".

None of the participants considered that the answers of Siri were excellent. Only 16% considered them above average while 37% of them considered them average, 42% below average, and 4% very poor.

In the case of Cortana, only 8% of the evaluators consider that their answers were excellent, but 54% of them considered them above average and 18% average. Overall, the distribution of responses for Siri is more skewed towards negative results than that of Cortana. It can be concluded that the performance of Siri is the worst out of the four assistants, followed by Cortana and that both Google Assistant and Alexa are better than them.

Figure 5 shows for each of the assistants the result obtained to evaluate: "How good were the answers?". Google and Alexa have a similar performance in this question, with Alexa having a slight edge of 4% in the excellent category while Google Assistant has 5% more in the above average one. Given that the median and the IQR of Alexa and Google are quite close (45 and 4 for Google Assistant and 44 and 5.25 for Alexa) there is no statistical evidence that they are significantly different. Figure 6 shows a comparison of the sum of the responses of the participants by assistant based on "How correct were the answers?".

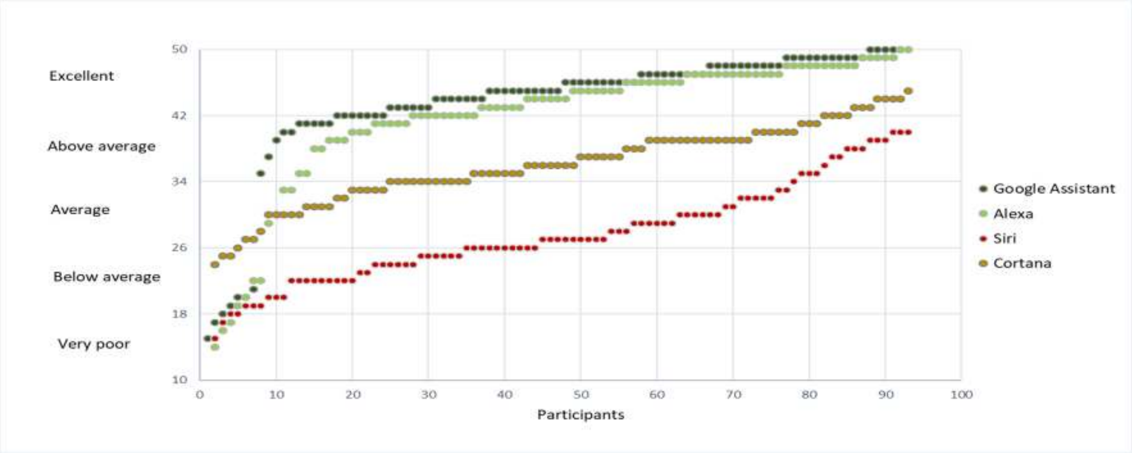


Figure 5. Individual responses to the question "How good was the answer?".

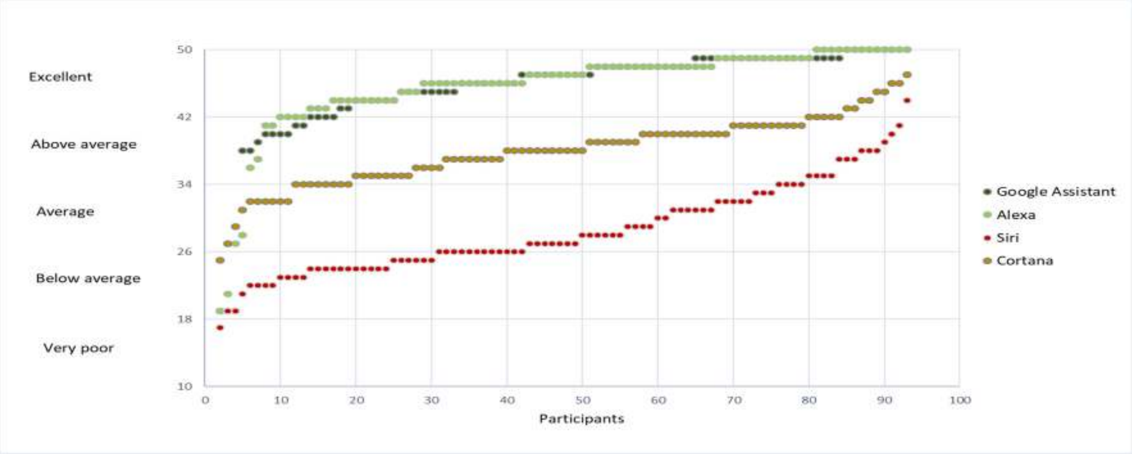


Figure 6. Individual responses to the question "How correct was the answer?".

In the case of Siri, when evaluating if the answers were correct, its performance was poor since 43% of the participants consider that the answers are below average (incorrect answers). This is the worst performance among the four assistants. In the case of Cortana, 71% consider that the answers were above average, and 18% regard them as average.

In the case of Siri and Cortana the numbers are considerably low, considering that these assistants are used to help people in their daily activities or to solve everyday problems, and the most important thing is to ensure that they provide good communication and correct answers.

Table 2 summarizes the results. For each assistant, the median of each question was calculated, and the resulting value was discretized with the same logic as the individual values. Google Assistant and Alexa are the best in both quality and correctness. Cortana ranks below both and Siri has the worst performance of all four assistants. Siri and Cortana in some cases do not provide an answer to the questions, and when they do provide it is not always correct or of quality.

In the case of Siri and Cortana, the numbers are considerably low, considering that these assistants are used to help people in their daily activities or to solve everyday problems, and the most important thing is to ensure that they provide good communication and correct answers.

Although there is no statistical evidence to confirm that Google is better than Alexa, in the results it can be noted that for the question "How good were the answers?" Google results are slightly better. This may be related to the fact obtained by many results of several studies: The female voice of Google Assistant tends to be more natural and express more emotions than the other assistants [1,13].

Table 2. Summary of the results for each assistant.

Personal Assistants	Quality	Correctness
Google Assistant	Excellent	Excellent
Alexa	Excellent	Excellent
Cortana	Above average	Above average
Siri	Average	Average

5. Conclusions and Future Work

This paper described the results of an evaluation of four intelligent personal assistants, to identify the best assistant based on how good and correct their answers were. The study included the most popular personal assistants on the market: Siri, Cortana, Alexa, and Google Assistant. A total of 92 participants conducted the study.

Results show that Alexa and Google are significantly better than Siri and Cortana. There is no statistically significant difference to confirm that Alexa is better than Google Assistant or vice versa. It is interesting to note that for both assistants, the evaluations provided are either very positive or very negative, with very few evaluators giving them a regular score.

On the other hand, Cortana and Siri show the worst performance, the last being the one that produces the lowest results. It is interesting that Siri, being one of the most popular voice assistants in the market since it is in the iPhone [1], has such a low performance when compared with the other three assistants. Cortana's answers were ranked by most evaluators as "above average", which proves interesting in that for Alexa and Google the evaluators tended to score them as "excellent".

Although our results are promising, similar studies or replications should be conducted in different contexts, to gather more empirical evidence on the use of intelligent personal assistants. It would be interesting to expand this research in the future by exploring other types of intelligent personal assistants. Another interesting area of future work is how to improve the quality of the answers that the assistants provided.

There is an opportunity to conduct new studies on evaluating why Cortana's answers are not as excellent as those of Google Assistant and Alexa, because despite having a good performance, its answers were not considered "excellent" by the participants.

Further studies are needed to evaluate the interaction of the user with intelligent personal assistants and gain a better understanding of how the interaction can affect the obtained results. In addition, we could include diverse populations, which can strengthen the results.

Author Contributions: Conceptualization, G.L., L.Q. and A.B.; methodology, G.L., L.A.G., A.B.; data curation, A.B.; writing—original draft preparation, A.B.; writing—review and editing, G.L. I.D., L.Q.; supervision, L.A.G. and G.L.

Acknowledgments: This work was partially supported by the Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC) and Escuela de Ciencias de la Computación e Informática (ECCI), both at Universidad de Costa Rica (Grant No. 834-B6-178). An acknowledgment also to all the survey respondents.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aron, J. How innovative is Apple's new voice assistant, Siri? *NewScientist* **2011**, *212*, 24, doi:10.1016/S0262-4079(11)62647-X.
2. Jiang, J.; Hassan Awadallah, A.; Jones, R.; Ozertem, U.; Zitouni, I.; Gurunath Kulkarni, R.; Khan, O.Z. Automatic online evaluation of intelligent assistants. In Proceedings of the 24th International Conference on World Wide Web, 18–22 May 2015; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2015; pp. 506–516.
3. Van Beurden, M.H.; Ijsselsteijn, W.A.; de Kort, Y.A. User experience of gesture-based interfaces: A comparison with traditional interaction methods on pragmatic and hedonic qualities. In *International Gesture Workshop*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 36–47.
4. Myers, K.; Berry, P.; Blythe, J.; Conley, K.; Magazine, M.G.-A. *An Intelligent Personal Assistant for Task and Time Management*; López, G., Quesada, L., Guerrero, L.A., Eds.; Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces; Springer: Cham, Switzerland, 2018, pp. 241–250.
5. Microsoft. Cortana. 21 May 2019. Available online: https://www.microsoft.com/windows/cortana/ (accessed on 28 October 2019).
6. Apple Inc. Siri. 21 May 2019. Available online: http://www.apple.com/ios/siri/ (accessed on 28 October 2019).
7. Google Inc. 21 May 2019. https://google.com/landing/now/ (accessed on 28 October 2019).
8. Amazon Inc. 21 May 2019 from Alexa Skills Kit. Available online: https://developer.amazon.com/public/solutions/alexa/alexa-skills-kit (accessed on 28 October 2019).
9. Pyae, A.; Joelsson, T.N. Investigating the usability and user experiences of voice user interface: A case of Google home smart speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*; ACM: New York, NY, USA, 2018; pp. 127–131.
10. Hoy, M.B. Alexa, siri, cortana, and more: An introduction to voice assistants. *Med. Ref. Serv. Q.* **2018**, *37*, 81–88.
11. Yang, X.; Aurisicchio, M.; Baxter, W. Understanding Affective Experiences with Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2019.
12. López, G.; Quesada, L.; Guerrero, L.A. *Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces*; Springer: Cham, Switzerland, 2017.
13. Canbek, N.G.; Mutlu, M.E. On the track of artificial intelligence: Learning with intelligent personal assistants. *J. Hum. Sci.* **2016**, *13*, 592–601.