UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

EVALUACIÓN DE ASISTENTES INTELIGENTES POR VOZ CON BASE EN LA CALIDAD DE LAS RESPUESTAS

Trabajo Final de Investigación Aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Computación e Informática para optar al grado de Maestría Profesional en Computación e Informática

ANA LAURA BERDASCO ROMERO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2020

Dedicatoria

A mis padres que siempre han sido mi apoyo incondicional y mi mayor tesoro.

A Gastoncito, mi inspiración para tomar riesgos y no mirar atrás.

A Rafa por apoyarme en cada decisión y motivarme a superarme cada día más.

Agradecimientos

Quiero agradecer a mi profesor guía el Dr. Gustavo López Herrera por todo el apoyo y confianza brindada durante este trabajo de investigación, además del soporte y la guía académica. De igual manera al M.Sc. Ignacio Díaz Oreiro y al Dr. Luis Quesada Quirós por toda su dedicación y críticas útiles.

Un especial agradecimiento a todos los participantes del estudio, por su colaboración y el tiempo invertido. También agradezco a los miembros del Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC) de la Universidad de Costa Rica, por el apoyo y la retroalimentación recibida durante todo el proceso de la investigación.

Agradezco a todos los profesores de los cursos de maestría, pero en especial a la profesora Marta Calderón por su dedicación en cada uno de los cursos que tome con ella.

"Este Trabajo Final de Investigación Aplicada fue aceptado por la comisión del programa de Estudios de Posgrado en Computación e Informática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Computación e Informática".

Dr. Luis Guerrero Blanco
Representante del Decano Sistema de Estudios de Posgrado

Dr. Gustavo López Herrera
Profesor Guía

Dr. Luis Quesada Quirós
Lector

M.Sc. Ignacio Díaz Oreiro
Lector

Dra. Gabriela Marín Raventós
Directora del Programa de Posgrado en Computación e Informática

Ana Laura Berdasco Romero
Sustentante

Tabla de contenidos

Índice de figuras

Índice de cuadros

Resumen

En los últimos años, los asistentes inteligentes por voz han tomado gran importancia y popularidad basados en la capacidad que tienen de ayudar a los usuarios con tareas cotidianas como crear alarmas, enviar correo, entre muchas funcionalidades. Esta investigación realiza una comparación de los asistentes inteligentes con base en la calidad y correctitud de las respuestas proporcionadas al momento de ejecutar diferentes tareas.

Noventa y dos estudiantes de diferentes carreras de la Universidad de Costa Rica participaron en la evaluación que determinó cual asistente ofrecía mayor satisfacción al usuario. Los resultados revelaron que Google Assistant y Alexa tienen el mejor rendimiento, seguidos de Cortana y Siri. Esta investigación fue publicada en la Conferencia Internacional en Computación Ubicua e Inteligencia Ambiental (UCAmI 2019). El artículo se puede ver en el anexo 1.

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Ana Laura Berdasco Romero, con cédula de identidad 800830202, en mi condición de autor del TFG titulado Evaluación de asistentes inteligentes por voz con base en la calidad de las respuestas.

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [x] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Ana Laura Berdasco Romero

Número de Carné: B10942 Número de cédula: 800830202

Correo Electrónico: ana.berdasco@ucr.ac.cr

Fecha: 17 / 6 / 2020 Número de teléfono: 83343683

Nombre del Director (a) de Tesis o Tutor (a): Gustavo López Herrera

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

Capítulo 1

Introducción

Con el avance de la tecnología, el uso de asistentes por voz se ha visto en aumento [1], particularmente en los últimos años con la aparición de una nueva generación de asistentes inteligentes potenciados por voz, liderados por Siri de Apple, Cortana de Microsoft, Alexa de Amazon y Google Assistant. Una de las razones para el mayor uso de los asistentes por voz, son las mejoras en la precisión del reconocimiento automático de voz [2].

A pesar de la importancia que han tomado los asistentes, evaluar estos dispositivos es un reto debido a la variedad y gran cantidad de tareas admitidas, por ejemplo, comando de voz, búsqueda web, chat, entre otras [3]. Por este motivo, es posible encontrar estudios en los que solamente se evalúa un dispositivo con base en las tareas soportadas [4], en otros casos, los estudios se basan en la comparación de estas tareas entre los mismos dispositivos [5,6].

En los últimos años, los estudios han mostrado la importancia de empezar a evaluar los asistentes inteligentes por voz tomando en consideración la satisfacción de los usuarios con las respuestas que reciben y la importancia de obtener una respuesta natural de parte de los asistentes inteligentes por voz [2]. Es decir, no solo evaluar la capacidad de los asistentes para ejecutar una tarea, sino la forma en la cual ejecuta la misma, entiéndase como la capacidad de comunicar apropiadamente una respuesta por parte de los asistentes [7].

Dada la popularidad que tienen los asistentes inteligentes por voz, existen estudios que evalúan y comparan los asistentes [2,7,8], pero no hay estudios que tomen en consideración la satisfacción del usuario con base en la calidad de la respuesta que está recibiendo de los asistentes, sino que se limitan a evaluar los asistentes desde las funcionalidades que estos ofrecen.

Otro detalle que es importante mencionar, es que la mayoría de las evaluaciones para conocer la satisfacción de los usuarios con los asistentes es realizada por las mismas compañías, por lo tanto, no son objetivas al momento de evaluar los asistentes, además que no hacen comparaciones con otros asistentes inteligentes por voz que existen en el mercado.

Con respecto a lo anterior, podemos agregar que las empresas que fabrican los asistentes inteligentes por voz habilitan plataformas de desarrollo para incluir nuevas funcionalidades a los asistentes, por lo que es importante que los desarrolladores de software cuenten con estudios imparciales que evalúen la experiencia de usuario, con base en la calidad de la respuesta de los dispositivos, lo cual les permita priorizar la mejora y desarrollo de nuevas funcionalidades.

1.1 Objetivos

A continuación, se detallan los objetivos de la presente investigación.

1.1.2 Objetivo General

El objetivo general de esta investigación es evaluar la satisfacción de los usuarios con base en las respuestas de los asistentes inteligentes por voz.

1.1.3 Objetivos específicos

Los objetivos específicos de la investigación son los siguientes:

1. Identificar los principales asistentes inteligentes por voz.
2. Crear un instrumento para evaluar las respuestas de los asistentes con el fin de medir la satisfacción de los usuarios.
3. Evaluar los asistentes inteligentes por voz.

Seguidamente, el capítulo 2 presenta un estado del arte que da evidencia de las revisiones de literatura efectuadas, en el capítulo 3 se describe un marco conceptual con los términos relevantes para este trabajo. Por su parte, en el capítulo 4 se explica la metodología que se utilizó para llevar a cabo la investigación, mientras que en el capítulo 5 se presentan los resultados de la evaluación. Por último, el capítulo 6 presenta las conclusiones y posibles trabajos futuros.

Capítulo 2

Estado del Arte

Se realizó una revisión de literatura acerca de las técnicas existentes para evaluar los asistentes inteligentes por voz. La evaluación estuvo enfocada en investigaciones que evaluaran aspectos como la calidad de las respuestas y la satisfacción de los usuarios con las mismas.

2.1 Trabajo Relacionado

Existen diferentes formas en las cuales se pueden evaluar los asistentes por voz, inclusive los desarrolladores de estos ofrecen un mecanismo de evaluación que consiste en un *checklist* para evaluar la funcionalidad es de los mismos. Pero estos mecanismos, más que medir qué tan satisfechos están los usuarios con los asistentes, miden la capacidad de los asistentes para ejecutar ciertas tareas. Por ejemplo: Amazon ofrece una guía para evaluar a Alexa en donde una de las tareas es crear una notificación [9]. Claramente esto permite evaluar la capacidad de Alexa para ejecutar la tarea, pero no la satisfacción del usuario.

Muchos de los trabajos que resaltan en la literatura están enfocados en la evaluación de un solo asistente y las tareas que éste puede realizar, desde búsquedas, configuración de notificaciones, entre otras tareas [1,10]. Al mismo tiempo, Por otro lado, en [4] se señalan los desafíos que pueden enfrentar los usuarios con los asistentes, por ejemplo, que en ocasiones el usuario debe de repetir el comando que se usó o que se pueden presentar problemas de integración con otros dispositivos, entre otros desafíos.

Otro ejemplo de artículos que evalúan los asistentes es “Alexa, Siri, Cortana, and More: *An Introduction to Voice Assistants”* [5], en el cual no solo se hace una evaluación de las tareas que estos ofrecen, sino que se desarrollan temas como la

privacidad y los problemas de seguridad que los asistentes enfrentan con la información de los usuarios.

En el tema de evaluación de los asistentes por voz, podemos mencionar el trabajo realizado por un grupo de investigadores de Microsoft [2] que trató de automatizar la evaluación de los asistentes y también predecir la calidad del reconocimiento por voz. La mayor parte del trabajo se centra en crear un modelo que permita evaluar las tareas soportadas sin necesidad de que una persona física lo realice, y la satisfacción es evaluada en términos de la capacidad del asistente para entender la tarea asignada.

También existen estudios que no solo se enfocan en evaluar las capacidades de los asistentes y han empezado a tomar en cuenta, como parte de la evaluación, las experiencias afectivas de los usuarios con los asistentes [11], es decir, las emociones humanas como parte de la evaluación. Algunos de estos estudios señalan las emociones como parte importante de la experiencia de usuario [12].

Uno de los enfoques que es importante mencionar, es el de los autores Gustavo López, Luis Guerrero y Luis Quesada [13]. Ellos plantearon un estudio en el cual evaluaron las respuestas de los asistentes con base en la exactitud y naturalidad de las respuestas de los dispositivos, lo cual no solo se preocupa de evaluar las tareas que los asistentes realizan, sino que toma en consideración experiencia del usuario.

Nuestro estudio a diferencia de los anteriores, los participantes evalúan las respuestas de los asistentes en base a la calidad y correctitud y sin tomar en consideración la interacción de los usuarios con los dispositivos. Ya que lo que se busca es identificar cual asistente inteligente por voz ofrece más satisfacción a los usuarios al ejecutar una tarea.

Capítulo 3

Marco Conceptual

Para comprender en su totalidad la idea principal de este trabajo final de investigación aplicada (TFIA), es necesario explicar los conceptos más importantes.

Una interfaz natural (*Natural User Interface (NUI)*, su nombre en inglés) una forma más natural para que las personas interactúen con la tecnología. NUI se refiere tanto a las entradas sensoriales como el tacto, el habla y los gestos [19].

Un asistente inteligente por voz es un servicio de software que está junto a un dispositivo de hardware especializado, como un altavoz inteligente o simplemente una función que se ofrece en un dispositivo informático de uso general, como una computadora personal, tableta, teléfono inteligente o computadora portátil (como un reloj de pulsera digital), el cual ofrece un conjunto de habilidades de un asistente humano tradicional, que responde preguntas y realiza tareas utilizando el procesamiento de voz y lenguaje natural respaldado por inteligencia artificial [9].

Los agentes inteligentes por voz tienen como propósito realizar tareas o servicios por medio de la interacción con el usuario, gracias a la capacidad de acceder a información de una variedad de fuentes en línea.

En la Tabla 1 se describen los asistentes inteligentes utilizados en esta evaluación:

Tabla 1. Descripciones asistentes inteligentes por voz

Asistente	Descripción
Alexa	Alexa es el servicio por voz ubicado en la nube de Amazon, disponible en los dispositivos de Amazon y dispositivos terciarios con Alexa integrada. Alexa puede controlar varios dispositivos inteligentes que sean compatibles con su sistema operativo como altavoces, televisores, electrodomésticos entre otros dispositivos.
Google Assistant	Es un asistente virtual desarrollado por Google que se puede encontrar en diferentes dispositivos y teléfonos Android y iOS. Al igual que Alexa se puede programar para controlar dispositivos inteligentes o conectarse con dispositivos externos para aumentar la capacidad de los asistentes.
Cortana	Cortana es un asistente personal inteligente desarrollado por Microsoft que puede ser usado en diversos dispositivos compatibles con el sistema operativo Windows 10. Tiene la capacidad de controlar otros dispositivos inteligentes, aunque es un espectro más limitado que Alexa y Google Assistant.
Siri	Siri es un asistente inteligente personalizado para para iOS, macOS, tvOS y watchOS. Posee la capacidad de conectarse con otros dispositivos que sean compatible con el sistema operativo de Apple.

En esta investigación se evalúan los asistentes inteligentes por voz en 2 dimensiones: una objetiva que mide la correctitud de la respuesta (es decir, si es

objetivamente precisa) y una subjetiva que mide su calidad (tal como la percibe la persona que recibe la respuesta del dispositivo).

Por ejemplo, para la pregunta: ¿Quién es el presidente de Canadá? Se obtuvieron las siguientes respuestas de Alexa y Google Assistant que se pueden observar en la Tabla 2

Tabla 2. Respuesta Asistentes a la pregunta: ¿Quién es el presidente de Canadá?

Alexa	Google Assistant
Canadá no tiene presidente, pero el primer ministro es Justin Trudeau.	Justin Trudeau. Aquí hay un resumen del sitio web: Wikipedia.org. El actual ministro de Canadá es el líder del partido liberal…

Al observar las respuestas de la Tabla 2 se puede notar que la respuesta correcta es proporcionada por Alexa, ya que la pregunta es ¿Quién es el presidente de Canadá? y Canadá no tiene presidente sino Primer Ministro, por el otro lado la respuesta de Google no es correcta a la pregunta, pero si nos da una respuesta que puede ser percibida con mayor calidad, ya que me da más detalles y no simplemente un nombre.

Capítulo 4

Metodología

Esta sección describe la metodología utilizada para alcanzar cada uno de los objetivos planteados. La Figura 1 muestra los objetivos, método y actividades realizadas.

Figura 1.Metodología

Para cumplir con el primer objetivo específico, se realizó una revisión de literatura que consistió en las siguientes actividades:

- **Diseño de la revisión**

1. Definición de las preguntas de investigación.

    - ¿Cuáles son los dispositivos más usados en el mercado?
    - ¿Cómo son evaluados los asistentes inteligentes por voz?

2. Definición del proceso de búsqueda: esto incluye la selección de los criterios de selección y calidad de los estudios, además de la definición de los datos a extraer.

    - Motores de búsqueda.
    - Año de publicación de los estudios.

- **Ejecución de la revisión y análisis de resultados**

En esta etapa se procedió a ejecutar la revisión de literatura con el fin de identificar cuáles son los dispositivos inteligentes más usados. La información se recolectó de diferentes bibliotecas digitales (Springer y ACM). Los artículos tomados en consideración están ubicados entre el año 2003 y el año 2019.

Para cumplir con el objetivo número dos: *Crear un instrumento para evaluar las respuestas de los asistentes con el fin de medir la satisfacción de los usuarios,* se realizaron las siguientes actividades:

- **Selección de escenarios**

Después de la selección de los asistentes, se procedió a la selección de los escenarios que los participantes deberán evaluar. La selección de los escenarios se hizo en conjunto con un grupo de expertos en HCI. Un escenario en este contexto se define como una tarea en la que una persona estaría interesada en que

el asistente lo ayude. Esta definición es intencionalmente flexible para adaptarse a una amplia gama de tareas.

- **Selección del método de evaluación**

Los dispositivos serán evaluados con base en la calidad y correctitud de las respuestas que ofrezcan a cada uno de los escenarios definidos previamente.

- **Evaluación exploratoria (Piloto)**

Se realizó una evaluación con 10 personas a las que se les proporcionó un escenario y solicitó interactuar con el asistente para que este les brindara una respuesta. Un ejemplo de la orientación proporcionada a los participantes del piloto fue: "Imagina que quieres hacer una suma". Cada participante hizo preguntas al asistente de maneras ligeramente diferentes, algunos preguntaron "¿Cuánto es la suma de tres más 4?", Mientras que otros preguntaron "3 más 4"

Estas interacciones permitieron la recopilación de las preguntas en inglés, ya que el escenario se proporcionó en español. El objetivo era comprender cómo interactuaban naturalmente los participantes con los asistentes inteligentes en cada escenario, con la mínima orientación.

- **Diseñar un modelo de evaluación**

Después de obtener los resultados del piloto, se creó un modelo de evaluación que funcionara para todos los dispositivos y que no permitiera que uno estuviera en ventaja sobre otro.

En esta etapa se definieron las preguntas en inglés para evaluar las respuestas de los asistentes inteligentes por voz. La Tabla 3 muestra las preguntas utilizadas en español e inglés.

Tabla 3. Preguntas de la evaluación en español e inglés

Pregunta en inglés	**Pregunta en español**
How does a dog sound?	¿Cómo suena un perro?
Thirteen plus seventeen	Trece más diecisiete
What is the speed of the light?	¿Cuál es la velocidad de la luz?
Where does Keylor Navas play?	¿Dónde juega Keylor Navas?
Which team won the soccer world cup of Italy 90?	¿Qué equipo ganó el mundial de fútbol de Italia 90?
I want to play a game	Quiero jugar un juego
How many US dollars are 10000 Costa Rican colons?	¿Cuántos dólares estadounidenses son 10000 colones costarricenses?
Who is Canada's president?	¿Quién es el presidente de Canadá?
What is the chemical formula for water?	¿Cuál es la fórmula química del agua?
Set the alarm to six o'clock AM	Configure la alarma a las seis en punto de la mañana

Las preguntas seleccionadas a pesar de ser sencillas permiten evaluar a los asistentes en diferentes escenarios como ciencias, matemáticas, actividades cotidianas y la interacción directa con el dispositivo.

En esta etapa se creó el instrumento de evaluación basado en dos preguntas para evaluar la calidad y correctitud de las respuestas:

- ¿Qué tan buenas fueron las respuestas? (*How good were the answers?)*
- ¿Qué tan correctas fueron las respuestas? (*How correct were the answers?*)

Los participantes evaluaron los asistentes con base en una escala Likert de 5 puntos, la cual se explica en la Tabla 4.

Un punto clave de esta investigación es que la evaluación de los asistentes se hizo en el idioma inglés, ya que muestran mejores resultados versus otros idiomas

como el español en el cual los asistentes inteligentes por voz no tienen el entrenamiento suficiente. La Tabla 5 muestra un ejemplo de la evaluación realizada.

Tabla 4.Escala Likert para evaluar los asistentes

Categoría	Valor
Excelente	5
Por encima del promedio	4
Promedio	3
Pobre	2
Muy pobre	1

Tabla 5. Ejemplo evaluación de los asistentes

Preguntas	Google Assistant	Alexa	Siri	Cortana
¿Qué tan buenas fueron las respuestas?				
¿Qué tan correctas fueron las respuestas?				

- **Implementación del instrumento**

El instrumento de evaluación es el mecanismo por el cual los participantes evalúan los asistentes. Dado que el objetivo de este trabajo es evaluar la satisfacción de los participantes con los asistentes, se creó un video para poder incluir un mayor

número de participantes, el video presentaba el escenario a evaluar y la respuesta de cada uno de los asistentes.

- **Ejecución de la evaluación**

La evaluación de los asistentes fue llevada a cabo con 92 estudiantes universitarios, divididos en 5 grupos. Estos eran estudiantes activos de la Universidad de Costa Rica y tenían entre 18 y 26 años al momento del estudio. En el Anexo 2 se encuentra la evaluación completa.

- **Análisis de Datos**

Cada respuesta se consideró individualmente (¿Qué tan buenas fueron las respuestas? Y ¿Cuán correctas fueron las respuestas?) Y luego se agruparon. Para agrupar los resultados, se agregaron los diez puntajes de un solo participante. Esto proporciona un puntaje agregado con un valor mínimo de 10 y un máximo de 50 por participante. Se realizaron pruebas de normalidad y los datos no mostraron una distribución normal.

Se utilizo la prueba del Shapiro Wilk para probar la hipótesis nula de si los datos son normales, utilizando un valor de significancia de "*0.05*". El valor *p* fue menor a "*0.05*" para todas las preguntas en ambos escenarios de correctitud y calidad. Por lo que se rechaza la hipótesis nula de que los datos son normales. Los gráficos cuantil cuantil están en concordancia con los resultados del Shapiro Wilk, pueden observarse en la Figura 2 y 3.

La no normalidad de los datos impide el uso de una prueba paramétrica ANOVA para comparar las medias. Por lo tanto, se utilizó la prueba de Kruskal Wallis, que es un equivalente no paramétrico de las pruebas de ANOVA que no requieren que los datos se distribuyan normalmente. El valor p de las pruebas Kruscal Wallis en todos los casos fue menor al valor de significancia de 0.05 por lo cual se rechaza la hipótesis nula de que los datos vienen de la misma distribución y se acepta la hipótesis alternativa que son diferentes.

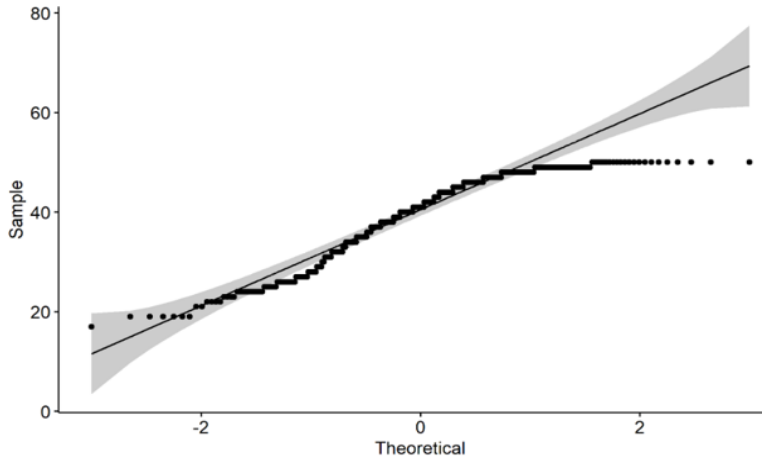


Figura 2. Resultados de la prueba de Shapiro Wilk para la normalidad: ¿Qué tan correctas fueron las respuestas?

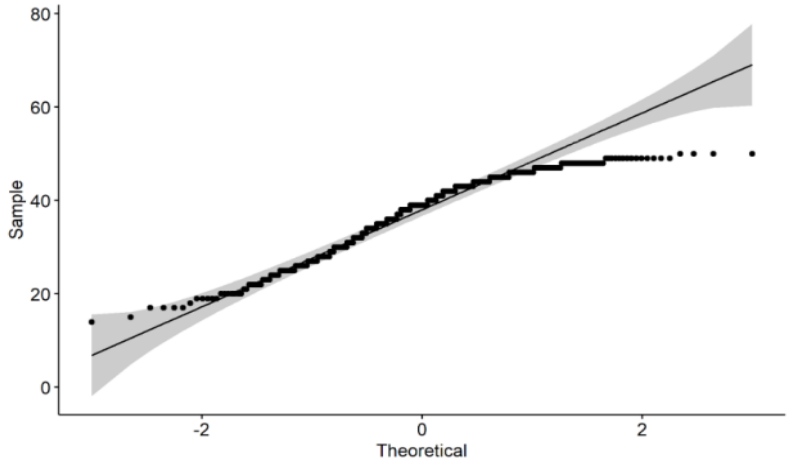


Figura 3. Resultados de la prueba de Shapiro Wilk para la normalidad: ¿Qué tan buenas fueron las respuestas?

Para categorizar cualitativamente, los valores de los resultados se dividieron en cinco categorías: "excelente", "superior al promedio", "promedio", "inferior al promedio" y "pobre". Dado que los valores pueden tener un rango de 10 a 50, este rango fue dividido en cinco segmentos iguales. Por lo tanto, cada uno de ellos abarca ocho unidades, por ejemplo: el rango "muy pobre" incluye todas las respuestas entre 10 y 18, mientras que el "excelente" incluye aquellas entre 42 y 50.

En el anexo 3 se muestran los valores p para la prueba Shapiro Wilk y Kruskal Wallis.

Capítulo 5

Resultados

Esta sección describe los resultados de la evaluación con 92 participantes. Es interesante mencionar que el 99% de los participantes eran conscientes de la existencia de varios asistentes, pero solo el 86% había utilizado al menos uno de ellos. La distribución de los participantes de la encuesta se explica en la Tabla 6.

Tabla 6. Distribución de los participantes en la evaluación

Tipo de participante	Distribución participantes	Edad promedio	Conoce los asistentes inteligentes por voz	Ha utilizado algún asistente inteligente por voz
Mujer	24%	23	100%	84%
Hombre	76%	22	98%	91%

La Figura 4 muestra para cada uno de los asistentes el resultado obtenido al evaluar: "¿Qué tan buenas fueron las respuestas?"; los dos mejores por un amplio margen, son Alexa y Google Assistant. Este último es el mejor, superando a Alexa en aproximadamente un 12% en la categoría excelente. La Figura 5 muestra una comparación de la suma de las respuestas de los participantes que separan a cada asistente para contrastar según "¿Qué tan correctas fueron las respuestas?". La superioridad de Google Assistant y Alexa también es evidente en esta figura.

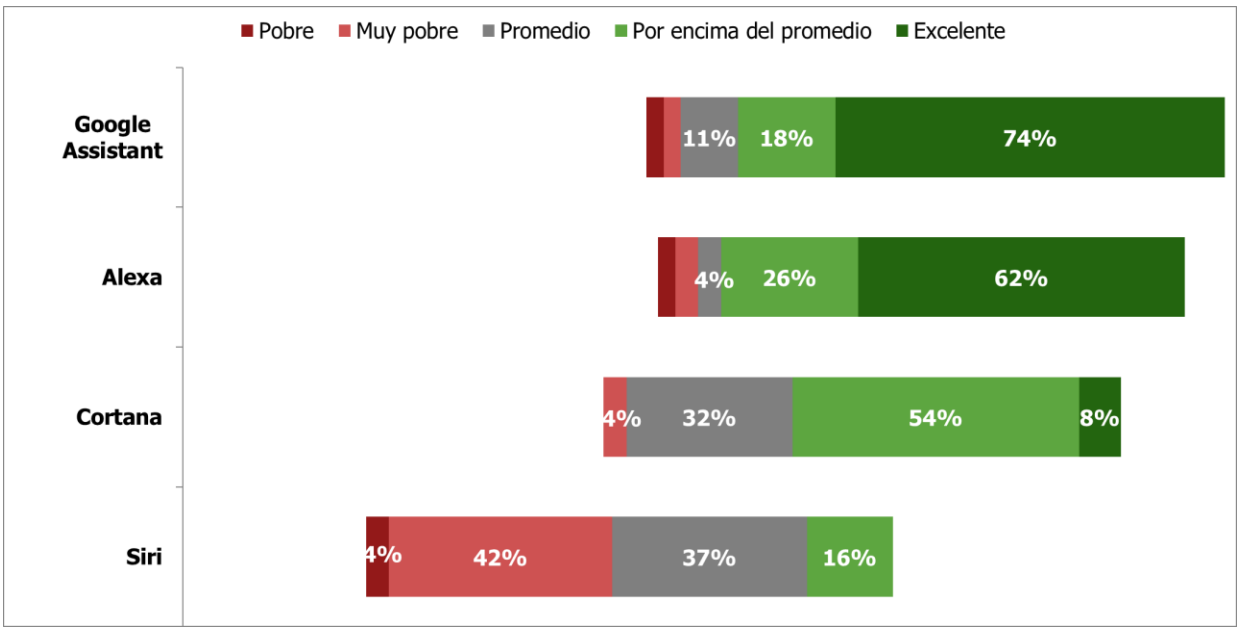


Figura 4. Resultados de la pregunta "¿Qué tan buenas fueron las respuestas?"

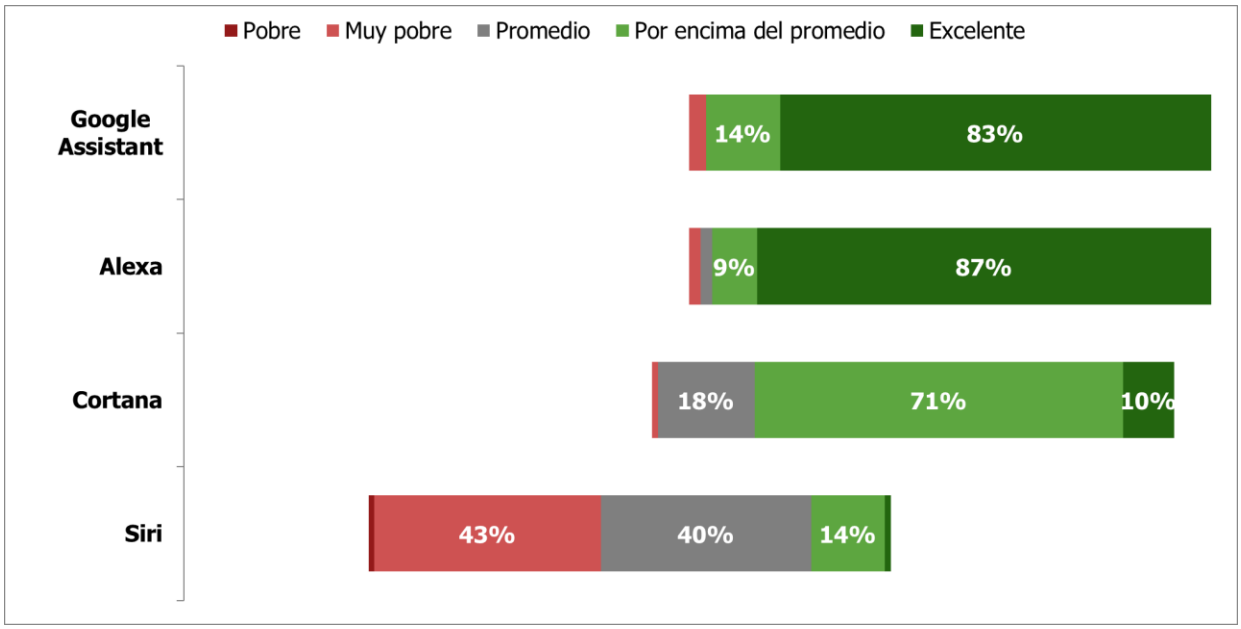


Figura 5. Resultados de la pregunta “¿Qué tan correctas fueron las respuestas?"

Ninguno de los participantes consideró que las respuestas de Siri fueron excelentes. Solo el 16% las consideró por encima del promedio, mientras que el 37% las consideró promedio, el 42% por debajo del promedio y el 4% muy pobres.

En el caso de Cortana, solo el 8% de los evaluadores considera que sus respuestas fueron excelentes, pero el 54% las consideró por encima del promedio y el 18% promedio. En general, la distribución de respuestas para Siri está más sesgada hacia resultados negativos que la de Cortana. Se puede concluir que el desempeño de Siri es el peor de los cuatro asistentes seguido de Cortana, mientras tanto Google Assistant y Alexa son mejores que ellos.

La Figura 7 muestra para cada uno de los asistentes el resultado obtenido al evaluar: "¿Qué tan buenas fueron las respuestas?", Google y Alexa tienen un desempeño similar en esta pregunta. Alexa cuenta con una ligera ventaja del 4% en la categoría excelente, mientras que Google Assistant tiene 5% más en la categoría: por encima del promedio. Dado que la mediana y el IQR de Alexa y Google están bastante cerca (45 y 4 para Google Assistant y 44 y 5.25 para Alexa) no hay evidencia estadística de que sean significativamente diferentes. La Figura 6 muestra una comparación de la suma de las respuestas de los participantes por asistente según: "¿Qué tan correctas fueron las respuestas?".

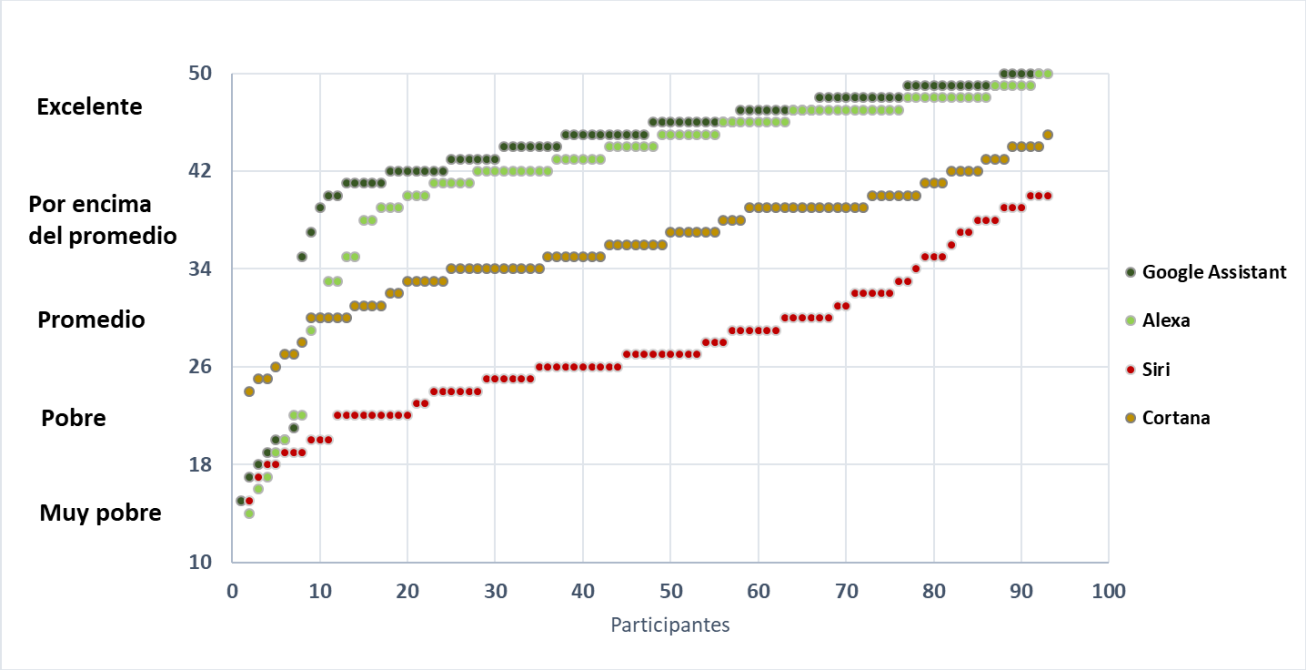


Figura 6. Respuestas individuales a la pregunta: ¿Qué tan buena fue la respuesta?

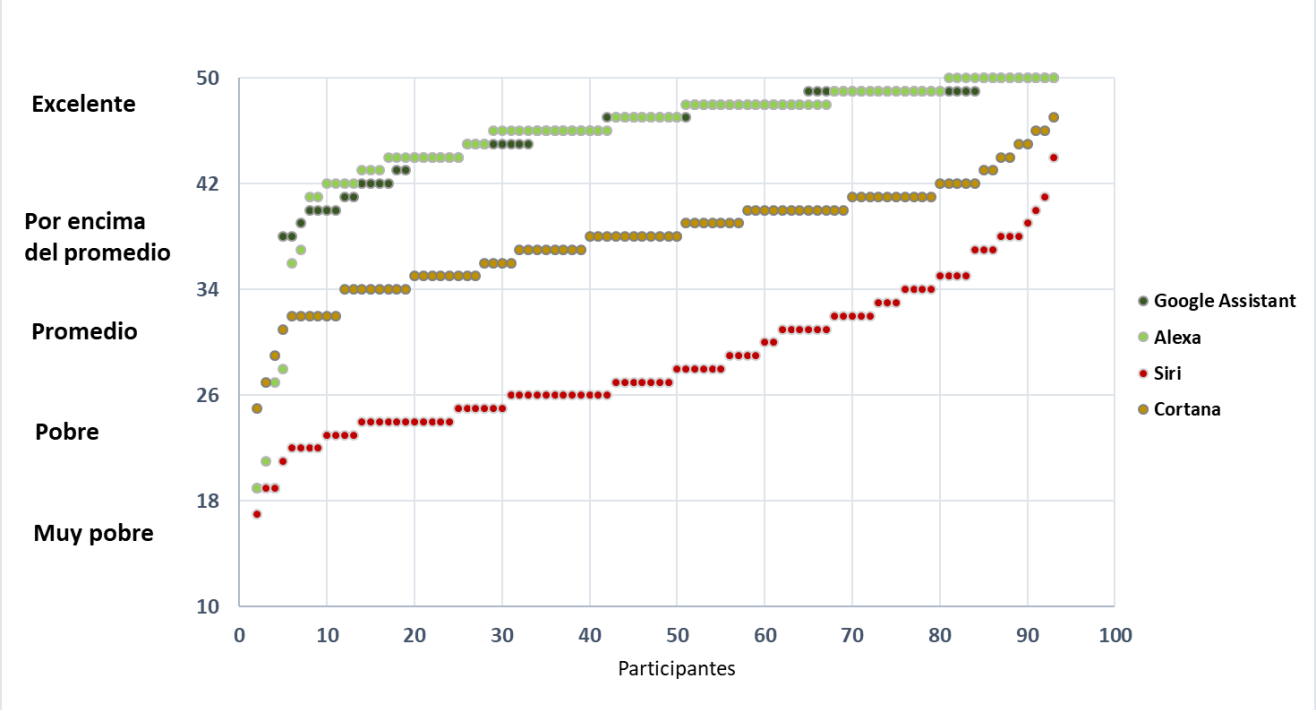


Figura 7. Respuestas individuales a la pregunta: ¿Qué tan correcta fue la respuesta?

En el caso de Siri, al evaluar si las respuestas son correctas, su desempeño es muy pobre ya que el 43% de los participantes considera que las respuestas están por debajo del promedio (respuestas incorrectas). Este es el peor desempeño entre los cuatro asistentes. En el caso de Cortana, el 71% considera que las respuestas estaban por encima del promedio y el 18% las considera como promedio.

En el caso de Siri y Cortana, los números son considerablemente bajos teniendo en cuenta que estos asistentes se utilizan para ayudar a las personas en sus actividades diarias o para resolver problemas cotidianos, lo más importante es garantizar que proporcionen una buena comunicación y respuestas correctas.

La Tabla 7 resume los resultados. Para cada asistente se calculó la mediana de cada pregunta y el valor resultante se discretizó con la misma lógica que los valores individuales. Google Assistant y Alexa son los mejores, tanto en calidad como en corrección. Cortana se ubica por debajo de ambos y Siri tiene el peor desempeño de los cuatro asistentes. Siri y Cortana en algunos casos no brindan una respuesta a las preguntas, y cuando lo hacen no siempre es correcta o de calidad.

Tabla 7. Resumen de los resultados para cada asistente

Asistente inteligente	**Calidad**	**Correctitud**
Google Assistant	Excelente	Excelente
Alexa	Excelente	Excelente
Cortana	Por encima del promedio	Por encima del promedio
Siri	Promedio	Promedio

Aunque no hay evidencia estadística para confirmar que Google es mejor que Alexa, en los resultados se puede observar que en la pregunta "¿Qué tan buenas fueron las respuestas?", los resultados de Google son ligeramente mejores. Esto

puede estar relacionado con los resultados obtenidos de varios estudios: la voz femenina del Asistente de Google tiende a ser más natural y expresar más emociones que los otros asistentes [13, 1].

Capítulo 6

Conclusiones y Trabajo futuro

En esta investigación, se realizó la evaluación de cuatro asistentes personales inteligentes para identificar al mejor asistente en función de qué tan buenas y correctas fueron sus respuestas. El estudio incluyó a los asistentes personales más populares del mercado: Siri, Cortana, Alexa y Google Assistant; además, 92 participantes realizaron el estudio.

Los resultados muestran que Alexa y Google son significativamente mejores que Siri y Cortana. No existe una diferencia significativa para confirmar que Alexa es mejor que el Asistente de Google o viceversa. Es interesante observar que, para ambos asistentes, las evaluaciones proporcionadas son muy positivas o muy negativas, con muy pocos evaluadores que les otorgan una calificación regular.

Por otro lado, Cortana y Siri muestran el peor desempeño, siendo el último el que produce los resultados más bajos. Es interesante que Siri siendo uno de los asistentes de voz más populares en el mercado, ya que está en el iPhone [1], tenga un rendimiento tan bajo en comparación con los otros tres asistentes. La mayoría de los evaluadores clasificaron las respuestas de Cortana como "superiores al promedio", lo que resulta interesante porque para Alexa y Google los evaluadores tendieron a calificarlas como "excelentes".

Un aspecto clave para este proyecto fue aplicar los conocimientos sobre revisiones de literatura aprendidos en diferentes cursos de la maestría, de esa manera se pudo conocer el estado del arte y tener una visión clara de lo que existía y lo que debíamos de aportar para que este estudio fuera relevante

De igual manera el desarrollo de este proyecto me permitió ampliar mi conocimiento sobre instrumentos de evaluación y cómo utilizarlos de manera idónea para conseguir los resultados deseados.

Aunque nuestros resultados son prometedores, se deben realizar estudios o réplicas similares en diferentes contextos para reunir más evidencia empírica sobre el uso de los asistentes personales inteligentes, sería interesante expandir esta investigación explorando otro tipo de asistentes personales inteligentes, por ejemplo, las pulseras inteligentes. Otra área de trabajo futuro a explorar sería cómo mejorar la calidad de las respuestas que proporcionaron los asistentes. Existe también oportunidad de realizar nuevos estudios que evalúen también que le falta a Cortana para que sus respuestas sean excelentes, tal como las del Asistente de Google y Alexa, porque a pesar de tener un buen rendimiento, sus respuestas no son consideradas "excelentes" por los participantes. Se necesitan más estudios para evaluar la interacción del usuario con los asistentes personales inteligentes y comprender mejor cómo la interacción puede afectar los resultados obtenidos. Además, podríamos incluir una población diversa que pueda fortalecer los resultados, por ejemplo, adultos mayores o niños que puedan ayudar a entender la interacción de estos grupos con los asistentes.

Referencias

[1] Aron, J. (2011). How innovative is Apple's new voice assistant, Siri?.

[2] Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic online evaluation of intelligent assistants. In Proceedings of the 24th International Conference on World Wide Web (pp. 506-516).

[3] van Beurden, M. H., Ijsselsteijn, W. A., & de Kort, Y. A. (2011). User experience of gesture-based interfaces: a comparison with traditional interaction methods on pragmatic and hedonic qualities. In International Gesture Workshop (pp. 36-47). Springer, Berlin, Heidelberg.

[4] Myers, K., Berry, P., Blythe, J., Conley, K., Gervasio, M., McGuinness, D. L., & Tambe, M. (2007). An intelligent personal assistant for task and time management. AI Magazine, 28(2), 47-47.

[5] Hoy, M. B. (2018). Alexa, siri, cortana, and more: An introduction to voice assistants. Medical reference services quarterly, 37(1), 81-88.

[6] López, G., Quesada, L., & Guerrero, L. A. (2017). Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces.
In International Conference on Applied Human Factors and Ergonomics (pp. 241-250). Springer, Cham.

[7] Derrick, B., & White, P. (2017). Comparing two samples from an individual Likert question. International Journal of Mathematics and Statistics, 18(3).

[8] Microsoft Corporation. [Online]. Available: https://www.microsoft.com/en-us/mobile/experiences/cortana/. [Accessed: 10- Jan-2019].

[9] Siri Support [Online]. Available: http://www.apple.com/ios/siri/. [Accessed: 10- Jan-2018].

[10] Google Assistant Support.” [Online]. Available: https://assistant.google.com/. [Accessed: 10- Jan-2019].

[11] Kepuska, V., & Bohouta, G. (2018). Next generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 99-103). IEEE.

[12] Canbek, N. G., & Mutlu, M. E. (2016). On the track of artificial intelligence: Learning with intelligent personal assistants. Journal of Human Sciences, 13(1), 592-601.

[13] Alexa skills kit.” [Online]. Available: https://developer.amazon.com/public/solutions/alexa/alexa-skills-kit. [Accessed: 10- Jan-2019].

[14] Yang, X., Aurisicchio, M., & Baxter, W. (2019). Understanding Affective Experiences with Conversational Agents. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

[15] Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. In Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques (pp. 261-268).

[16] Pyae, A., & Joelsson, T. N. (2018). 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 (pp. 127-131).

[17] Cohen, P., Cheyer, A., Horvitz, E., El Kaliouby, R., & Whittaker, S. (2016). On the future of personal assistants. In Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems (pp. 1032-1037).

[18] “Amazon has finally revealed how many Alexa devices have been sold” [Online]. Available: https://www.businessinsider.com/amazon-reveals-alexa-sales-2019-1. [Accessed: 10- Jan-2019].

[19] Kaushik, D., & Jain, R. (2014). Natural user interfaces: Trend in virtual interaction. arXiv preprint arXiv:1405.0101.

Anexo 1.
Artículo Publicado

Este anexo incluye el texto completo del artículo presentado en ‘13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.”

proceedings

MDPI

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; ana.berdasco@ucr.ac.cr, 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.

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

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.

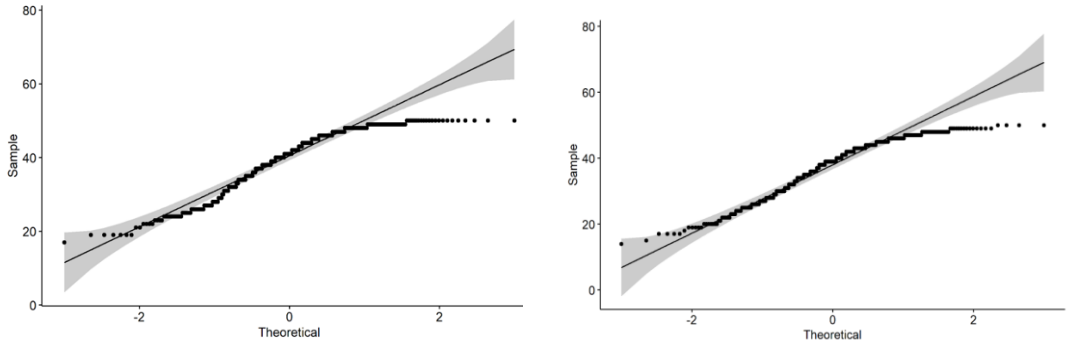
Table 1. Example of the questions for evaluating the assistants.

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

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

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.

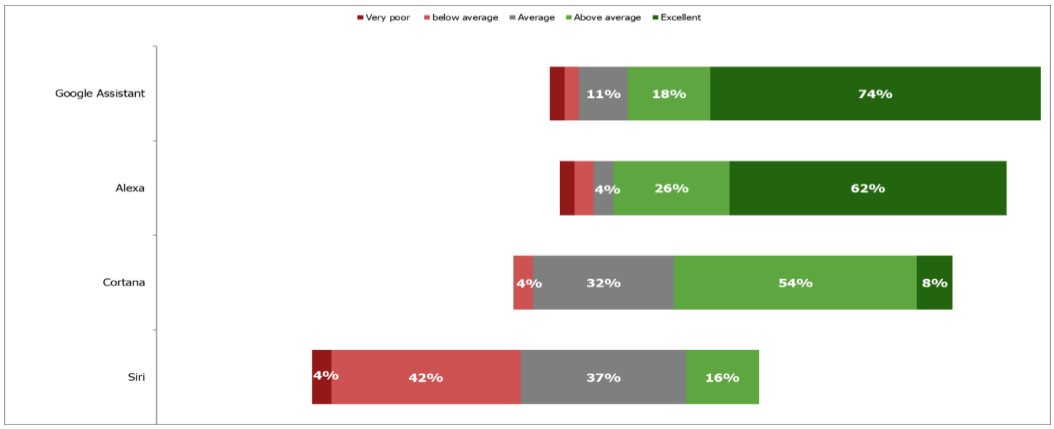


Figure 3. Results for the question "How good were the answers?".

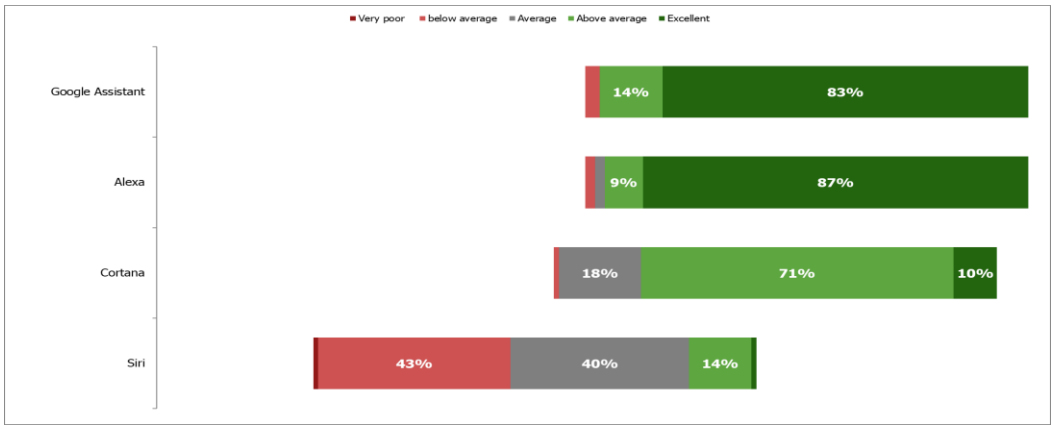


Figure 4. Results for the question "How correct were the answers?".

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

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?".

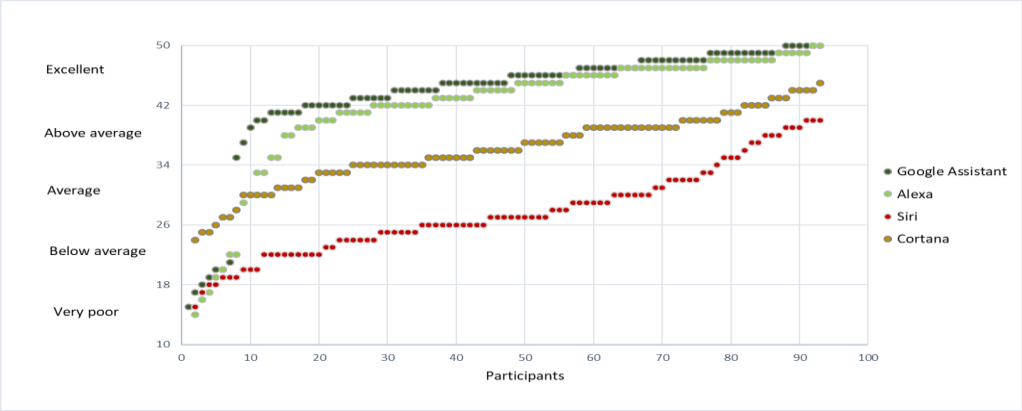


Figure 5. Individual responses to the question "How good was the answer?".

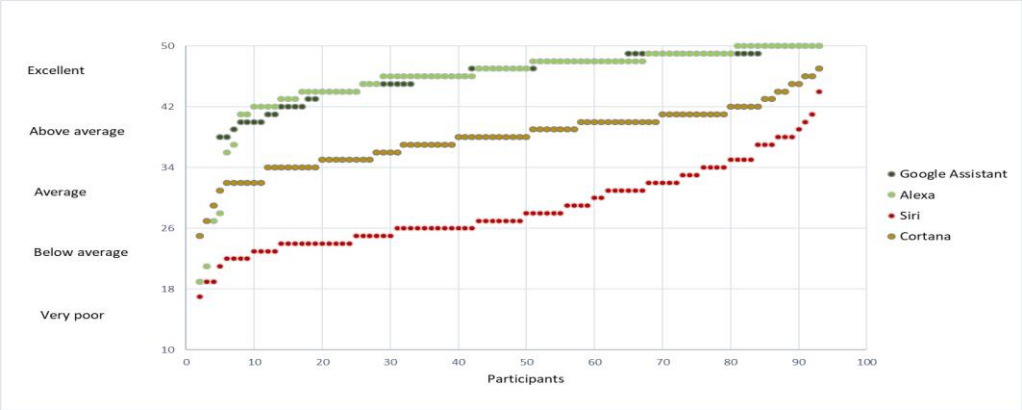


Figure 6. Individual responses to the question "How correct was the answer?".

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.

© 2019 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Anexo 2.

Encuesta utilizada para la evaluación de los asistentes inteligentes por voz.

	Preguntas / Dispositivos	Google Home	Alexa	Siri	Cortana
1	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
2	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
3	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
4	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
5	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
6	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**

7	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
8	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
9	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				
		Google Home	**Alexa**	**Siri**	**Cortana**
10	**P1**- ¿Qué tan ***buenas*** fueron las respuestas? How ***good*** were the answers?				
	P2 - ¿Qué tan ***correctas*** fueron las respuestas? How ***accurate*** were the answers?				

Encuesta Demográfica

*Obligatorio

1. Género *

- [ ] Mujer
- [ ] Hombre

2. Edad *: _____

3. Carrera (Profesión) *: _____

4. ¿Cómo define su nivel de inglés (conversacional)? *

- [ ] Principiante
- [ ] Intermedio básico
- [ ] Intermedio
- [ ] Intermedio Avanzado
- [ ] Avanzado

5. ¿Conoce algún asistente por voz (Siri, Cortana, Alexa, Google Home)? *

- [ ] Si
- [ ] No

6. ¿Ha usado alguno de los siguientes asistentes por voz? *

- [ ] Google Home
- [ ] Cortana
- [ ] Alexa
- [ ] Siri

Anexo 3.

Encuesta utilizada para la evaluación de los asistentes inteligentes por voz.

Pregunta	Tipo	shapiro.wilk.p.valor	kruskal.p.valor
Pregunta 1	Correcto	1.80E-27	1.99E-43
Pregunta 2	Correcto	1.23E-26	2.42E-36
Pregunta 3	Correcto	9.67E-33	0.103051
Pregunta 4	Correcto	3.96E-26	5.32E-39
Pregunta 5	Correcto	7.97E-26	3.70E-31
Pregunta 6	Correcto	1.00E-18	1.54E-25
Pregunta 7	Correcto	4.46E-21	1.16E-49
Pregunta 8	Correcto	4.86E-30	0.001731
Pregunta 9	Correcto	1.06E-26	6.09E-18
Pregunta 10	Correcto	4.49E-21	3.02E-32
Pregunta 1	Bueno	3.69E-25	3.45E-24
Pregunta 2	Bueno	4.73E-24	2.23E-26
Pregunta 3	Bueno	5.88E-26	1.02E-04
Pregunta 4	Bueno	2.42E-22	4.94E-28
Pregunta 5	Bueno	4.79E-20	3.19E-21
Pregunta 6	Bueno	4.70E-18	2.94E-23
Pregunta 7	Bueno	2.03E-19	7.24E-37
Pregunta 8	Bueno	1.18E-23	2.35E-10
Pregunta 9	Bueno	1.35E-22	6.89E-11
Pregunta 10	Bueno	2.50E-19	2.73E-23