

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

DISEÑO DE UN CUESTIONARIO DE EVALUACIÓN DE EXPERIENCIA DE
USUARIO POR MEDIO DE UN ASISTENTE INTELIGENTE POR VOZ

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

MARCO ANDRÉS CHACÓN CHAVES

Ciudad Universitaria Rodrigo Facio, Costa Rica

2021

Dedicatoria

A Dios.

A abuela, por siempre creer en mí.

A mis padres y hermano, por el apoyo incondicional.

Agradecimientos

Mi profunda gratitud a mi profesor guía el M.Sc. Ignacio Díaz Oreiro por la guía, el apoyo y los consejos dados durante este trabajo de investigación. De igual manera, quiero agradecer al Dr. Gustavo López Herrera y al Dr. Luis Guerrero Blanco por los consejos y críticas útiles brindadas durante las revisiones de mi proyecto. Además, quiero agradecer a los integrantes del grupo USING de la Universidad de Costa Rica por toda la retroalimentación brindada.

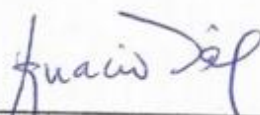
Agradezco enormemente a los participantes del estudio, por el tiempo invertido y los comentarios brindados que fueron de gran utilidad en la realización de este trabajo de investigación. Un especial agradecimiento a Alexandra Aguirre y Mariana Córdoba, por siempre apoyarme e inspirarme a crecer cada día como profesional y persona. De igual forma, quiero agradecer a Edgar Esquivel y Franklin Murillo por ayudarme a mantener los pies sobre la tierra sin dejar de soñar con un mejor futuro. A Catalina Paniagua, por darme el impulso cuando más lo necesitaba.

Un especial agradecimiento a la profesora Elzbieta Malinowski por las guías y consejos dados que me ayudaron y motivaron a empezar mi carrera profesional.

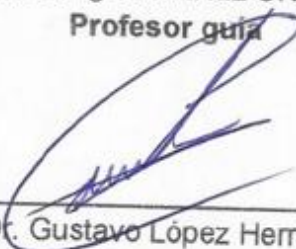
"Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Computación e Informática de la Universidad de Costa Rica, como requisito para optar por el grado y título de Maestría Profesional en Computación e Informática."



Dr. Luis José Quesada Quirós
**Representante de la Decana
Sistema de Estudios de Posgrado**



M.Sc. Ignacio Díaz Oreiro
Profesor guía



Dr. Gustavo López Herrera
Asesor



Dr. Luis Guerrero Blanco
Asesor



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



Marco Andrés Chacón Chaves
Sustentante

Tabla de contenidos

Dedicatoria.....	ii
Agradecimientos	iii
Hoja de aprobación.....	iv
Tabla de contenidos.....	v
Resumen	vi
Índice de Figuras	vii
Índice de Tablas.....	viii
Introducción	1
1.1 Objetivos	2
1.1.2 Objetivo general.....	2
1.1.3 Objetivos específicos	3
Trabajo relacionado	4
Marco conceptual.....	7
Metodología	11
Implementación del cuestionario por voz	19
Caso de Estudio: Mural.....	31
6.1 Evaluación de la plataforma MURAL mediante el cuestionario UEQ implementado con una interfaz conversacional.....	31
6.2 Evaluación del asistente por voz implementado mediante UEQ+	37
Conclusiones y trabajo futuro.....	51
Referencias.....	53
Anexo 1.....	58

Resumen

La presencia de los asistentes inteligentes por voz ha tenido un crecimiento considerable en los últimos años, ya que se pueden encontrar desde dispositivos especializados como los parlantes inteligentes, teléfonos inteligentes, hasta computadoras y aparatos del hogar. Dicha ubicuidad hace que su uso o interacción con las personas vaya más allá de tareas simples como crear recordatorios en el calendario, búsquedas rápidas en internet o creación de alarmas, sino que se expanda a tareas más complejas u otras áreas de aplicación.

Este trabajo aprovecha esa expansión en las áreas de aplicación de los asistentes inteligentes, y en particular de las interfaces conversacionales por voz, para proponer un instrumento de evaluación de experiencia de usuario que los utilice como medio de interacción con los participantes de la evaluación. Para ello, se presenta la implementación de un cuestionario estandarizado de experiencia de usuario, que utiliza una interfaz conversacional por voz como mecanismo de recolección de información de los usuarios, buscando una interacción natural, según los principios de diseño de interfaces conversacionales. En este trabajo también se expone un caso de estudio en el que se evalúa la experiencia de usuario al utilizar el instrumento propuesto.

Índice de Figuras

Figura 1: Representación de un diferencial semántico.....	8
Figura 2: Metodología	11
Figura 3: Implementación de mensaje de bienvenida e instrucciones generales del Asistente	27
Figura 4: Implementación de las preguntas segmentadas.	28
Figura 5: Implementación de la conversión y escritura de respuestas.....	29
Figura 6: Implementación del mecanismo de salida de la interacción	30
Figura 7: Gráfico de resultados UEQ MURAL Grupo 1.	32
Figura 8: Gráfico de resultados UEQ MURAL grupo 2.	33
Figura 9: Gráfico de resultados conjuntos UEQ MURAL Grupo 1 y Grupo 2.....	34
Figura 10: “Benchmark” de evaluación de MURAL comparado a 452 estudios previos de UEQ.....	35
Figura 11: Escala de medias evaluación UEQ+ Grupo 1.....	38
Figura 12: Respuesta de escala Likert Grupo 1	40
Figura 13: Nube de palabras con las respuestas del Grupo 1	41
Figura 14: Escala de medias evaluación UEQ+ Grupo 2.....	42
Figura 15: Respuesta de escala Likert Grupo 2	44
Figura 16: Nube de palabras con las respuestas del Grupo 2.....	45
Figura 17: Comparación de las medias del Grupo 1 y Grupo 2, en las 5 subescalas evaluada en UEQ+	46
Figura 18: Comparación de las respuestas adicionales con medias superiores para el Grupo 1 sobre el Grupo 2	48
Figura 19: Comparación de las respuestas adicionales con medias superiores para el Grupo 2 sobre el Grupo 1	49

Índice de Tablas

Tabla 1: Escalas de evaluación propuestas por Klein	17
Tabla 2: Escalas de evaluación propuestas por el investigador	17
Tabla 3: Resultados de encuesta de primera iteración del Mago de Oz.....	20
Tabla 4: Resultados de encuesta de la segunda iteración del Mago de Oz.....	21
Tabla 5: Escala UEQ (media y varianza) evaluación de MURAL Grupo 1.....	33
Tabla 6: Escala UEQ (media y varianza) evaluación de MURAL Grupo 2.....	34
Tabla 7: Escala UEQ (media y varianza) evaluación conjunta de MURAL Grupo 1 y Grupo 2.	35
Tabla 8: Escala con respuesta inconsistentes.....	36
Tabla 9: Medias y Varianzas en las subescalas Grupo 1	39
Tabla 10: Medias y Varianzas en las subescalas Grupo 2	43
Tabla 11: Comparación de media de las subescalas de UEQ+ para el Grupo 1 y Grupo 2.	46
Tabla 12: Preguntas adicionales con medias superiores para el Grupo 1	47
Tabla 13: Preguntas adicionales con medias superiores en el Grupo 2	48
Tabla 14: Resultados Pruebas T en resultados de UEQ+	50
Tabla 15: Resultados Pruebas T en resultados de preguntas de usabilidad	50



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, Marco Andrés Chacón Chaves, con cédula de identidad 114860326, en mi condición de autor del TFG titulado Diseño de un cuestionario de evaluación de experiencia de usuario por medio de un asistente inteligente por voz

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 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: Marco Andrés Chacón Chaves

Número de Carné: B11750 Número de cédula: 114860326

Correo Electrónico: mandres0112@gmail.com

Fecha: 23-11-2021 Número de teléfono: 88601404

Nombre del Director (a) de Tesis o Tutor (a): M.Sc. Ignacio Díaz Oreiro

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

La experiencia del usuario (UX) es actualmente un factor clave para establecer la calidad de un producto o servicio [1–3], y para evaluar esta experiencia de usuario se utilizan métodos y herramientas cuyo objetivo es determinar la percepción sobre el uso de un sistema o producto. Existen diferentes métodos para evaluar la experiencia de usuario, tales como: evaluación de expertos, estudios etnográficos, cuestionarios diseñados específicamente para un caso de estudio, entrevistas y cuestionarios estandarizados, por nombrar algunos de los más utilizados.

Los cuestionarios estandarizados, uno de los instrumentos más utilizados en la evaluación UX, contienen un conjunto invariable de preguntas, expuestas siempre en el mismo orden, que son respondidas por los propios participantes para expresar sus sentimientos y experiencias con respecto a los diferentes aspectos de un producto [4]. Estos cuestionarios utilizan escalas Likert [5] o diferenciales semánticos [6] para recopilar la opinión de los usuarios en relación con las características pragmáticas o hedonistas de los productos. Las características pragmáticas se refieren a esos rasgos de un producto como si es predecible, confuso, simple, complicado, entre otros. Como complemento, las características hedonistas son aquellas que representan sentimientos, es decir, si un producto nos resulta aburrido, interesante, novedoso, decepcionante, así como otros atributos relacionados con estimulación, identificación y evocación [7]. Cabe mencionar que los cuestionarios estandarizados son económicos y fáciles de usar, ya que el usuario los aplica por sí mismo y, además, se consideran confiables y válidos para medir UX [8].

Los cuestionarios estandarizados más reconocidos son AttrakDiff, UEQ y meCUE, formados por 28, 26 y 34 preguntas respectivamente. El hecho de que todas las preguntas estén estructuradas de la misma manera y con el mismo formato es positivo, dado que los usuarios que los llenan se acostumbran rápidamente a este formato [9]. Sin embargo, este hecho también podría influir negativamente debido al cansancio o aburrimiento que puede experimentar quien esté llenando el formulario, luego de haber introducido un conjunto importante de respuestas. Adicionalmente,

dado que estos cuestionarios se aplican por medios escritos (en papel o a través de una aplicación de software), no son totalmente accesibles para personas con dificultades visuales o motoras.

Por otro lado, en las últimas décadas se han desarrollado diversas interfaces no tradicionales que buscan explotar diferentes maneras de interactuar de las personas con los sistemas informáticos. Entre estas interfaces se encuentran las interfaces por voz, implementadas por medio de asistentes inteligentes, las cuales se han visto impulsadas por su potencial económico [10], y también su uso se ha diversificado, con tareas que van más allá de crear recordatorios o hacer búsquedas en internet [11]. Parte del éxito que están experimentando los asistentes por voz, se debe no solo a lo accesible que pueden llegar a ser en términos de uso [12], sino también a las mejoras sustanciales continuas que experimentan en reconocimiento de patrones de voz, expansión a otros idiomas además del inglés, entre otros. Debido a lo anterior, los asistentes inteligentes por voz se han vuelto muy populares, al punto de vender millones de unidades de estos dispositivos [13][14], y esta popularidad y la aceptación los convierte en un excelente medio para obtener información directamente del usuario.

Ahora bien, dado que la interacción a través de una interfaz por voz, como la que se encuentran en asistentes inteligentes, se asemeja a una conversación con otra persona [15], podrían utilizarse una interfaz conversacional por voz, la cual es una interfaz donde la interacción es conducida principalmente por palabra dicha y no escrita [16], como medio para implementar el mecanismo de recolección de información de un cuestionario estandarizado de experiencia de usuario.

1.1 Objetivos

A continuación, se detallan el objetivo general y los objetivos específicos de la presente investigación.

1.1.2 Objetivo general

El objetivo general de esta investigación es proponer una interfaz conversacional por voz para la evaluación de experiencia de usuario.

1.1.3 Objetivos específicos

Los objetivos específicos de la presente investigación son los siguientes:

1. Determinar el nivel de naturalidad necesaria para recolectar información de experiencia de usuario a través una interfaz conversacional por voz.
2. Especificar una interfaz conversacional por voz de evaluación de experiencia de usuario.
3. Analizar la experiencia de usuario de la interfaz conversacional por voz por medio de un caso de estudio.

El capítulo 2 presenta el trabajo relacionado referente a cuestionarios estandarizados, en particular de cuestionarios de experiencia de usuario y sobre los asistentes inteligentes por voz, enfocado principalmente en su capacidad como interfaces conversacionales por voz. El capítulo 3 describe el marco conceptual con los términos más importantes relacionados con este trabajo. El capítulo 4 detalla la metodología empleada en la realización de este trabajo, mientras que el capítulo 5 describe la implementación de la interfaz conversacional por voz. El capítulo 6 muestra los resultados del caso de estudio. Por último, el capítulo 7 presenta las conclusiones y posibles trabajos futuros derivados de este trabajo.

Capítulo 2

Trabajo relacionado

En función de los objetivos de este trabajo, es importante definir conceptos relacionados con dos áreas del trabajo, la primera de ellas relacionada con las características de los principales cuestionarios para la evaluación de UX, y la segunda enfocada en los asistentes por voz más utilizados, sus funcionalidades y las características conversacionales que deben encontrarse en una interfaz por voz.

Los tres cuestionarios más reconocidos para la evaluación de UX son AttrakDiff, UEQ y meCUE [7][17-20]. El primero de los tres cuestionarios que apareció en la industria fue AttrakDiff en el 2003 [20]. Consiste en 28 ítems a ser marcados por el usuario, donde cada ítem está construido por un diferencial semántico de 7 puntos. Más tarde, en 2008 se presentó el "User Experience Questionnaire" (UEQ) [21]. Se compone de 26 elementos también contruidos por diferenciales semánticos de 7 puntos. Finalmente, en 2013, se propuso el cuestionario meCUE [22], construido con 34 ítems formados por escalas Likert de 7 puntos.

Estos tres cuestionarios han sido utilizados extensivamente en las evaluaciones de UX ya que constituyen un método económico y a la vez confiable de obtener retroalimentación de los usuarios respecto de un producto o sistema [4][5]. Este hecho se señala en [8], donde se afirma que el 84% de los estudios documentados en la literatura académica al 2016, habían utilizado cuestionarios en sus evaluaciones de UX.

En cuanto a modificaciones a los cuestionarios estandarizados, algunos investigadores han realizado propuestas al respecto. Por ejemplo, en [23] se propone una variación al UEQ adaptado a la cultura taiwanesa, mientras que en [24] se traduce al francés el cuestionario AttrakDiff, convirtiendo las palabras que forman los diferenciales semánticos de los rubros mediante un proceso formal de traducción y que permite evitar los sesgos asociados a las interpretación y subjetividad del investigador, frecuentemente presentes en casos de traducciones no metódicas o informales. Otros investigadores proponen cuestionarios específicos para ciertos

dominios, como por ejemplo medir la experiencia de usuario de la televisión interactiva [25] o medir la experiencia de interfaces auditivas [26].

Cabe señalar que la investigación de los cuestionarios estandarizados se basó en el trabajo de Díaz-Oreiro, López, Quesada y Guerrero [9], donde se presenta una revisión sistemática de literatura sobre cuestionarios estandarizados de experiencia de usuario. En su revisión, los autores recopilaron los artículos de AttrakDiff, UEQ y meCUE sin importar el uso que se les daba a estos cuestionarios, y posteriormente seleccionaron aquellos que utilizaban el cuestionario en estudios primarios. Entre los artículos descartados se encontraron investigaciones que proponían modificaciones a los cuestionarios. Estos artículos fueron facilitados por los autores de la revisión sistemática de literatura y se comprobó que ninguno de éstos realizaba algún tipo de implementación por medio de una interfaz por voz. Adicionalmente, dado que los datos recolectados en [9] correspondían a artículos entre 2006 y 2018, se realizó una búsqueda similar para artículos entre enero de 2019 y el inicio de este trabajo, en busca de artículos que introdujeran interfaces por voz en cuestionarios estandarizados de evaluación UX, pero nuevamente no se encontró ninguno.

En relación con los asistentes inteligentes por voz, existen estudios enfocados en comparar diferentes asistentes virtuales realizando las mismas tareas, así como revisar temas de seguridad y privacidad [27]. Otros estudios se enfocan en evaluar la experiencia de usuario desde un punto de vista emocional al utilizar estos dispositivos [28], mientras que otros estudios se enfocan en evaluar qué tan correcta o qué tan buena es la respuesta dada por un asistente inteligente por voz [29], además de la naturalidad con la que estos asistentes contestan [15]. Por su parte, en [12] se muestra cómo los asistentes inteligentes por voz ayudan a reducir la brecha tecnológica que puede existir entre usuarios con algún tipo de discapacidad o impedimento físico, que les restringe o impide el uso de productos o dispositivos [12].

Actualmente, los fabricantes de los asistentes virtuales, como Amazon con Alexa, permiten realizar encuestas [30], sin embargo, estas son encuestas genéricas que funcionan como un servicio de grabación de información general ya que únicamente preguntan ¿qué producto o servicio se quiere evaluar?, ¿quién es el fabricante o proveedor? y ¿qué comentario se tiene del mismo (bueno o malo)? [30] También existe un servicio privado al que se accede desde un sitio web, para crear y aplicar

encuestas utilizando los asistentes inteligentes por voz Alexa y Google Assistant [31]. En estas encuestas se tiene como objetivo medir la satisfacción de usuario con preguntas como “en escala de 1 a 5, ¿qué tan satisfecho está con el servicio?”. Este tipo de interacción, sin embargo, no es natural dado que no se asemeja a una interacción entre dos personas, y solo se limita a copiar el formato de un cuestionario escrito.

En cuanto a las características conversacionales que debe contar una interfaz por voz, es de destacar el trabajo de Robert J. Moore y Raphael Arar [16] y Cathy Pearl [32] en el diseño de interfaces conversacionales por voz, que se abordarán nuevamente en la sección Marco Conceptual. Estos trabajos, sin embargo, no describen específicamente las características o implementación de un cuestionario en particular, ni tampoco de evaluaciones de experiencia de usuario.

De lo anterior se desprende que, a nuestro mejor entender, el tema de evaluar la experiencia de usuario por medio de interfaces conversacionales por voz, como las que proveen los asistentes inteligentes por voz, no ha sido abordado apropiadamente por otros investigadores, dado que no se encontraron en la literatura académica referencias a estudios de este tipo, y las soluciones comerciales disponibles son muy pobres en relación con la interacción que proveen.

Capítulo 3

Marco conceptual

Iniciemos entonces indicando que la experiencia de usuario (UX) es actualmente un elemento clave para determinar la calidad de un producto o servicio [3][33]. La experiencia de usuario abarca no solo conceptos llamados pragmáticos, como claridad de la estructura, comportamiento previsible o facilidad de aprendizaje, también encontrados en el concepto de Usabilidad, sino también aquellos pertenecientes a la experiencia llamada hedonística, relacionada con la estimulación y la identificación, de las que se recoge información referente a si un producto es aburrido o interesante, motivante, novedoso, entre otros [4][20-22][34].

El fin último de estudiar la experiencia de usuario es mejorar la interacción que los usuarios tengan con los productos o servicios. Para esto, un elemento imprescindible es la evaluación de la experiencia de usuario [37], que se refiere a un conjunto de métodos y herramientas que persiguen determinar la manera en que una persona percibe un sistema o producto antes de utilizarlo, durante uso y después de la utilización.

Uno de los elementos importantes a considerar en el análisis de la experiencia de usuario es qué método emplear en la medición [34]. Los métodos tienen que ver con la recolección de los sentimientos, opiniones y pensamientos conscientes de los usuarios respecto de la interacción con cierto producto. Adicionalmente, las experiencias pueden ser episódicas, medidas en un momento específico, o longitudinales, que evalúan una experiencia completa de un sistema o producto [34].

Una de las herramientas utilizadas para realizar evaluaciones de UX son los cuestionarios estandarizados, que contienen un conjunto conocido y fijo de preguntas o rubros en los que solicitan al usuario final su percepción de elementos como eficiencia del producto o sistema, claridad, novedad, entre otros [17]. Estos cuestionarios son económicos y fáciles de aplicar dado que son auto aplicados por el usuario basado en la experiencia percibida luego de utilizar un producto o servicio. Es por esta razón que su uso es extendido. Al mismo tiempo, son considerados fiables

y válidos para medir la experiencia de usuario [35].

Respecto de la estructura de cada cuestionario, los cuestionarios AttrakDiff y UEQ están contruidos por un diferencial semántico de 7 puntos [6]. En términos generales, se presenta un conjunto de adjetivos en forma bipolar respecto del mismo concepto, y se responde marcando un punto dentro de ambos extremos, en función de qué tanto se está de acuerdo con el término del extremo, por ejemplo, se presenta un ítem representado por las palabras “Inventivo – Convencional” como se muestra en la Figura 1, donde se debe marcar una de las 7 casillas que se encuentran en medio de los dos conceptos, dependiendo de qué tan cerca esté la opinión del usuario a una de esas palabras.

Inventivo O O O O O O O Convencional

Figura 1: Representación de un diferencial semántico

Las preguntas del cuestionario meCUE está formadas por escalas Likert [5]. En esta escala, ampliamente utilizada, cada rubro a medir se presenta mediante un enunciado positivo o negativo, que se responde seleccionando el nivel de acuerdo o desacuerdo que se tiene con el enunciado, por ejemplo, se puede tener el enunciado “el producto me agrada” y se selecciona el grado de acuerdo o desacuerdo entre las opciones “totalmente en desacuerdo”, “en desacuerdo”, “un poco en desacuerdo”, “ni de acuerdo ni en desacuerdo”, “un poco de acuerdo”, “de acuerdo” y “completamente de acuerdo”.

En cuanto a los conceptos relacionados con interfaces conversacionales, una de las formas de implementar estas interfaces es mediante asistentes inteligentes por voz, que se pueden definir como software que permite automatizar tareas permitiendo organizar y mantener información, mediante interacción oral con el usuario [36]. Estos asistentes funcionan como una interfaz natural dada la forma en la que los usuarios proveen información a los dispositivos durante su interacción [37].

Hay que tener presente que los asistentes inteligentes por voz se apoyan mucho en el procesamiento de lenguaje natural, el cual busca diseñar mecanismos eficaces de

comunicación entre las personas y las computadoras [38]. Este proceso está relacionado con el concepto de inteligencia artificial, es decir, la forma en que una máquina imita funciones cognitivas propias de los seres humanos [39].

Los asistentes inteligentes por voz se encuentran implementados en dispositivos específicos llamados altavoces inteligentes (*smart speakers*) y también en dispositivos de uso más general, como celulares inteligentes. Así, el asistente inteligente Alexa cuenta con hardware propio (Amazon Echo), mientras otros asistentes como Siri, se ejecutan en una gama de dispositivos como teléfonos inteligentes que cuentan con el sistema operativo iOS. Otros asistentes, como Google Assistant funcionan tanto en un hardware propio (Google Home) como en dispositivos que utilicen el sistema operativo Android, por ejemplo, en teléfonos inteligentes.

Para realizar sus tareas, los asistentes inteligentes por voz incorporan Reconocimiento de Voz, Procesamiento de Lenguaje Natural y Síntesis o Generación de Voz, que están disponibles en los altavoces o celulares inteligentes. Si se desea extender las habilidades inherentes a los dispositivos, se deben desarrollar funcionalidades específicas para cada asistente, como pueden ser los *skills* de Alexa o los *Actions* de Google Assistant. Es en este tipo de funcionalidad o programa que se implementará la interfaz conversacional por voz de recolección de información para realizar la evaluación UX.

En cuanto a las interfaces conversacionales por voz, Moore y Arar [16] y Pearl [32] describen los principios que deben cumplir las interfaces conversacionales de voz para crear una interacción apropiada y suplir toda aquella información que normalmente se presentaría visualmente en interfaces gráficas. Además, lo que se busca con una interfaz conversacional por voz es replicar de manera simplificada una conversación, en la cual se establece un contexto (tema sobre el cual trata la conversación), se toman turnos y se dan respuestas acordes al contexto [16]. Moore y Arar [16] definen tres principios del diseño conversacional para crear una conversación que ellos llaman “natural” entre un humano y un asistente, dado que no es posible realizar una conversación equivalente a la que se produce entre dos seres humanos. Estos tres principios son el diseño en función del receptor, la minimización y la reparación.

El diseño basado en el receptor indica que el asistente de voz o interfaz conversacional debe diseñarse en función del usuario esperado en la interacción, para que la misma sea natural, buscando utilizar las mismas palabras, conceptos y estructuras del usuario esperado. Minimización se refiere a que el asistente pronuncie sus enunciados (palabras, frases o sonidos que tengan sentido en sí mismos) con la mínima cantidad de palabras posible y que el permitan que el concepto expresado sea entendido por el usuario. En caso de que el usuario no entienda completamente lo que se le dijo, podría pedir que se le aclare el enunciado. Se busca con este principio que no se abrume al usuario con largas intervenciones del asistente de voz, dado que es probable que el usuario no recuerde toda la información que se le brinda, si ésta es mucha. Finalmente, el principio de reparación se refiere a la capacidad que debe implementarse para corregir enunciados que no se entendieron o escucharon bien por alguna de los participantes de la conversación. Por un lado, el Asistente de Voz podría preguntar al usuario sobre algo que no entendió, y por otro, si el usuario no entiende un enunciado del asistente, podría expresar esta situación, a lo que el asistente podría parafrasear el enunciado utilizando otras palabras o extendiéndose para aclarar su punto. En este último caso, sería válido no acatar el principio de minimización, dado que el entendimiento sería más importante.

Otro punto para describir es el cuestionario UEQ+, que se utilizó en la evaluación de la interfaz conversacional por voz. UEQ+ (A modular Extension of the User Experience Questionnaire) es una extensión de UEQ presentada en 2019 [40], en la que es posible construir modularmente un cuestionario UEQ pero orientado según lo que se quiera evaluar. Actualmente cuenta con 20 escalas para elegir (confianza, novedad, claridad, sensación háptica, entre otras) frente a las seis escalas UEQ tradicionales. En 2020, Klein et al [41] propusieron tres escalas propias para la calidad de la voz que se adaptan al formato UEQ+. Estas son: Comportamiento de la respuesta: El Asistente por Voz se comunica de forma respetuosa, paciente, educada y confiable; Calidad de las respuestas: Las respuestas del Asistente de Voz cubren las necesidades de información del usuario; Comprensibilidad: El Asistente de Voz comprende correctamente las instrucciones de los usuarios sin obligarlos a hablar de forma no natural.

Capítulo 4

Metodología

En esta sección se describe la metodología utilizada para completar cada uno de los objetivos planteados. La Figura 2 muestra los objetivos planteados, las actividades realizadas y los productos obtenidos.

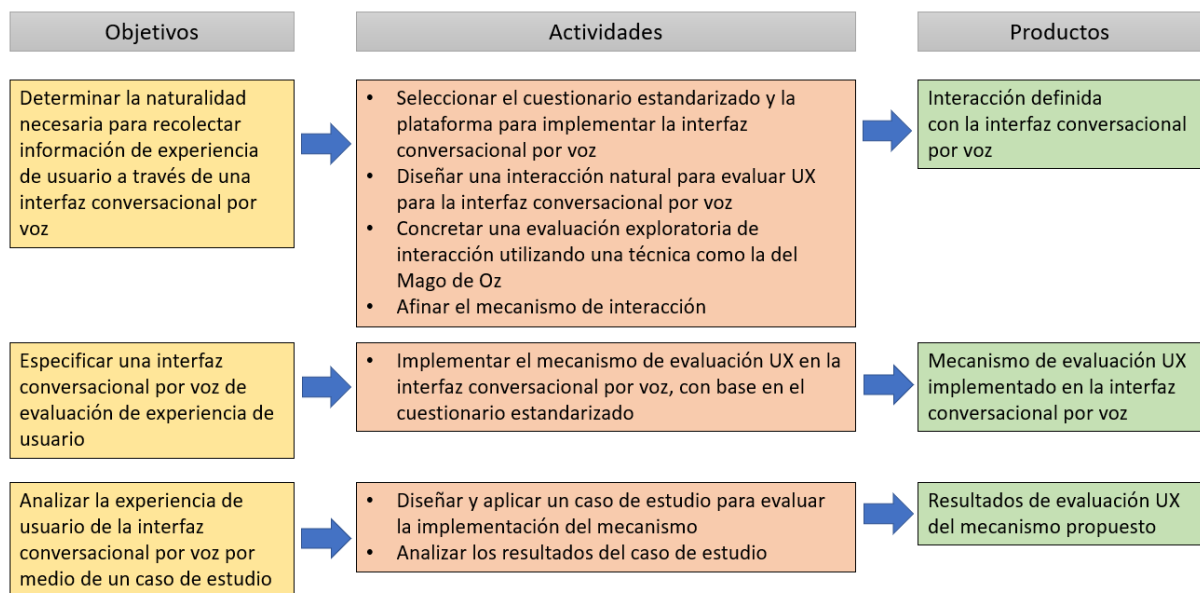


Figura 2: Metodología

Para cumplir con el primer objetivo específico, se realizó un conjunto de actividades que se describe a continuación.

Inicialmente, se estudiaron los cuestionarios estandarizados más reconocidos, así como sus características. Como se había indicado en el Marco Conceptual, los tres cuestionarios estandarizados más reconocidos en el mundo son AttrakDiff, UEQ y meCUE. AttrakDiff y UEQ son los más utilizados por los investigadores, según consta en [9], mientras que meCUE se ha utilizado muy poco. De este mismo estudio se desprende que, a pesar de que AttrakDiff fue el primero en proponerse, UEQ lo ha superado en cantidad de usos desde el año 2017. Además de este elemento, UEQ tiene a favor el hecho de que el cuestionario cuenta una traducción oficial al español, lo que hace más sencilla su aplicación.

Otro elemento a favor de UEQ es que brinda abiertamente las herramientas para analizar los datos recolectados, proveyendo un conjunto de hojas electrónicas donde se incorporan los datos de las respuestas de los usuarios, y las hojas electrónicas facilitan los resultados de las escalas UX en general, en detalle por ítem o pregunta, posibles inconsistencias detectadas en las respuestas de los usuarios, así como datos estadísticos de la consistencia de cada una de las escalas, en forma de coeficientes alfa de Cronbach y λ^2 de Guttman. Además, se pueden comparar los resultados con *benchmarks* construidos a partir de una base de datos de 452 estudios y 20190 participantes.

En cuanto a AttrakDiff, los datos deben subirse a un sitio web de donde se pueden generar resultados, hasta un máximo de 20 participantes, que puede extenderse para investigaciones académicas solicitando permiso a los responsables, por lo que su uso si es factible, pero un poco más burocrático que con UEQ.

Cabe recordar que ambos cuestionarios presentan sus preguntas en forma de diferencial semántico, a diferencia de meCUE que utiliza escalas Likert. Este elemento también fue considerado como positivo, dado que transformar un diferencial semántico, en el que solo se tienen dos conceptos opuestos, a una interfaz de voz se presentaba como un reto más interesante que el convertir escalas Likert, que ya se presentan en forma de oraciones.

Por todo lo anteriormente expuesto se decidió utilizar el cuestionario UEQ como base para la implementación del instrumento de evaluación UX que utilizara una interfaz conversacional de voz como mecanismo de captura de las respuestas de los participantes.

En cuanto a la elección del asistente inteligente por voz, se procedió a buscar aquellos que reportaran mejores resultados en cuanto a satisfacción de los usuarios en relación con las respuestas que brindan y también aquellos que facilitarían herramientas para la construcción de las interfaces. En lo referente a evaluaciones de satisfacción del usuario, se utilizó el estudio presentado en [3], donde Google Assistant y Alexa aparecen como los mejor calificados por los usuarios.

Para el tema de disponibilidad y facilidad de herramientas de desarrollo, uno de los

parámetros que influyó en la elección del asistente por voz fue la curva de aprendizaje, es decir, que en términos de programación fuera el más rápido y sencillo de utilizar, además de que no requiriera software o licencias adicionales para cualquier implementación. De dicha evaluación resultó electo Google Assistant, utilizando como plataforma de programación Voiceflow, la cual es una plataforma que permite programar el flujo del diálogo que tendrá la interacción final en Google Assistant. La plataforma Voiceflow proveyó una ventaja adicional al facilitar la aplicación del caso de estudio, dado que cuenta con la facilidad de poner la interfaz de voz implementada a disposición de los participantes a través de un sitio web. Esto facilitó la aplicación del caso de estudio, dado que fue posible distribuir el instrumento de evaluación de forma remota, sin necesidad de requerir que los participantes se desplazaran a un sitio específico a interactuar con un altavoz inteligente, que tuvieran que contar con uno propio o que debieran contar con un teléfono inteligente específico para poder utilizar la interfaz implementada en el Google *action*.

Para diseñar la interfaz y determinar su naturalidad se realizaron dos actividades complementarias: por un lado, identificar el tipo de formato de pregunta a la que mejor se adaptara el diferencial semántico presente en las preguntas de UEQ, y por otro, incorporar elementos que constituyan una conversación natural en la interfaz de voz.

Como se mencionó en el Marco Conceptual, Moore y Arar [16] consideran una interfaz conversacional de voz como natural si cumple los principios de diseño en función del receptor, la minimización y la reparación. Dentro de estos principios, estos autores definen las acciones básicas de navegación que debe tener una interfaz de voz natural, acciones que se fueron incorporando en los diferentes prototipos y versiones del cuestionario implementado por el Asistente de Voz. Estas acciones se consideran generales a todas las interfaces de voz y son las siguientes:

- Entender las capacidades del asistente, qué puede hacer.
- Solicitar que repita una pregunta o respuesta.
- Solicitar que explique o reformule un enunciado.
- Cerrar una pregunta o secuencia, por ejemplo, dando las gracias.
- Abortar una secuencia.
- Cerrar una conversación.

Para el diseño de la interacción se utilizó la técnica de diseño y evaluación “Mago de Oz”. Dicha técnica consiste en prototipar y simular el comportamiento final de un sistema computacional [42], pero obviando ciertas funcionalidades del producto final, con el objetivo de analizar la interacción del usuario lo antes posible en el desarrollo.

El principal aspecto por evaluar con el Mago de Oz era la forma en la que se debería formular la pregunta, ya que las preguntas en formato de diferencial semántico tienden a ser más visuales dado que se ubica la respuesta de acuerdo con la afinidad a uno de los dos términos. Para esto, se aplicaron dos formatos de preguntas en los cuales se pueda efectivamente implementar un diferencial semántico sin necesidad de ayuda visual. El primero de esos formatos son preguntas cuyas respuestas eran dadas en números del 1 al 7, donde el número 1 se asocia al primer término o concepto (el que está a la izquierda en el diferencial semántico) y 7 se asocia al concepto de la derecha. A este formato le llamaremos preguntas “1 a 7”. El segundo formato utilizado es uno donde las preguntas se dividen en dos partes, la primera pregunta responde a cuál de los términos del diferencial semántico se inclina la persona, es decir, la dirección de la respuesta, si es hacia el concepto de la izquierda, al de la derecha o al centro y, en una segunda pregunta se determina el nivel de afinidad con el término entre “un poco”, “mucho” o “en extremo”. Posteriormente esas respuestas se transforman en un único valor numérico para su posterior cuantificación y análisis. A este formato les llamaremos preguntas “segmentadas”.

Cabe resaltar que la ejecución del Mago de Oz se hizo de forma iterativa, es decir, se modeló e implementó un primer prototipo utilizando la mitad de la totalidad de preguntas de un cuestionario UEQ estándar, en las cuales se utilizaron los formatos descritos anteriormente. Dicho prototipo fue sometido a uso por un grupo de personas donde la evaluación que hacía el agente no era el punto principal sino la retroalimentación que pudieran dar los participantes respecto del uso del agente, dado que en este punto lo que se buscaba era establecer el nivel de naturalidad con la que debe interactuar el agente con los usuarios.

La mecánica implementada en la aplicación del Mago de Oz consistió en dividir a los participantes en dos grupos, donde el primer grupo interactuó con el primer prototipo del agente mientras que el segundo grupo interactuó con el prototipo después aplicar las mejoras sugeridas. Dichas interacciones se realizaron en sesiones individuales

con los participantes, donde cada una de las sesiones estaba dividida en tres partes; una primera parte donde se explicó los contenidos del documento de consentimiento informado, así como el objetivo a lograr con la interacción con la interfaz conversacional por voz y contestar cualquier pregunta previo al uso del prototipo. La segunda parte de las sesiones era la interacción del participante con la interfaz conversacional por voz en donde se simulaba el escenario en el participante observa la publicidad de un producto y luego procedía a realizar la evaluación de este. Finalmente, la tercera parte de las sesiones consistía en contestar un cuestionario donde unas preguntas eran en formato Likert y otras de respuesta libre sobre la interacción que el participante recién experimentó.

El cuestionario aplicado en la segunda iteración tiene una variación respecto al cuestionario aplicado en la primera iteración. Dicha variación consistió en eliminar las dos preguntas referentes a cuál formato es más fácil de contestar y reemplazarla por una única pregunta, buscando la misma información. Este cambio se hizo para evitar confusiones y dar una respuesta directa en cuanto a cuál de los dos formatos el participante piensa es más fácil de contestar o utilizar. Concluida las iteraciones del Mago de Oz y utilizando la retroalimentación recibida para la implementación final, se considera cumplido el primer objetivo de este trabajo.

Tomando como punto de partida la retroalimentación obtenida de la ejecución de las iteraciones del Mago de Oz, se decidió implementar la versión completa del cuestionario UEQ cuyos elementos serán explicados en el capítulo 5, para el caso de estudio que será descrito más adelante, utilizando el formato de preguntas segmentadas. Definido el formato de preguntas a utilizar y tomando en consideración la retroalimentación recibida tanto en la primera como en la segunda iteración del Mago de Oz, se aplicaron las mejoras al prototipo tanto en términos de interacción con los usuarios, así como mejoras a nivel de funcionalidad.

Concluido el proceso de diseño e implementación final del agente basado en la retroalimentación recibida durante la ejecución del Mago de Oz, y de las recomendaciones de qué debe tener una interfaz conversacional por voz para considerarse natural, se considera cumplido el segundo objetivo de este trabajo.

Para el tercer objetivo de este trabajo se diseñó y aplicó un caso de estudio, con el fin de evaluar el cuestionario estandarizado UEQ cuyo mecanismo de captura está implementado por una interfaz conversacional de voz. Para este caso de estudio se determinó estudiar la experiencia de usuario de estudiantes de segundo año de la carrera de Bachillerato en Ciencias de la Computación e Informática de la Universidad de Costa Rica, al utilizar la plataforma colaborativa MURAL. Para ello se identificaron estudiantes que habían utilizado esta plataforma colaborativa anteriormente.

El caso de estudio consistió en dos partes: en la primera parte los participantes evaluaron MURAL utilizando el cuestionario UEQ implementado con la interfaz conversacional por voz. Para ello se les suministró una dirección a un sitio web donde accedían a dicha implementación. En esta dirección no hay elementos visuales asociados a la conversación, únicamente la interacción por voz. En la segunda parte del ejercicio, los participantes evaluaron la interfaz conversacional por voz, también llamada Asistente de Voz. Esta evaluación se realizó por medio del cuestionario UEQ+ y también por medio de un conjunto de preguntas de usabilidad, definidas específicamente para este caso.

En el cuestionario UEQ+, que se construye modularmente, se utilizaron las tres escalas propuestas por Klein [41] para interfaces de voz: comportamiento de respuesta, calidad de respuesta y comprensibilidad. Cada escala está compuesta por 4 preguntas en forma de diferencial semántico, que responden a las preguntas e ítems que se muestran en la Tabla 1.

Además de estas tres escalas propuestas específicamente para la calidad de la voz, se incluyeron en la evaluación las escalas Uso intuitivo (el uso del Asistente de Voz no requiere mayor explicación o capacitación para su uso) y Originalidad (el Asistente por Voz representa un enfoque nuevo y original), que se describen seguidamente en la Tabla 2.

Tabla 1: Escalas de evaluación propuestas por Klein

Escala	Pregunta	Item izquierdo	Item derecho
Comportamiento de la respuesta	En mi opinión, el comportamiento de respuesta del Asistente de Voz es:	Artificial	Natural
		Desagradable	Agradable
		Antipático	Simpático
		Aburrido	Entretenido
Calidad de las respuestas	Las preguntas y respuestas del Asistente de Voz son:	Inapropiadas	Adecuadas
		Inútiles	Útiles
		Proveen ayuda	No proveen ayuda
		Ignorantes	Inteligentes
Comprensibilidad	En mi opinión, el Asistente de Voz ha entendido mis comandos de voz de forma:	Complicada	Simple
		Ambigua	Inequívoca
		Imprecisa	Precisa
		Enigmática	Explicable

Tabla 2: Escalas de evaluación propuestas por el investigador

Escala	Pregunta	Item izquierdo	Item derecho
Uso intuitivo	En mi opinión, utilizar el Asistente de Voz es:	Difícil	Fácil
		Ilógico	Lógico
		Equívoco	Evidente
		Incoherente	Coherente
Novedad	En mi opinión, la idea detrás del Asistente de Voz y su diseño son:	Faltos de imaginación	Creativos
		Convencionales	Originales
		Tradicionales	Novedosos
		Conservadores	Innovadores

Las preguntas de Usabilidad se agregaron para complementar las anteriores escalas descritas para evaluar UX en interfaces de voz, ya que están un poco más orientadas a la evaluación de elementos como el reconocimiento de voz, que es propio de las características que provee la plataforma de desarrollo VoiceFlow, y no parte del diseño del Asistente de Voz presentado en este estudio. Se presentaron como ítems de escala Likert de cinco puntos, donde las posibles respuestas fueron: “Totalmente en desacuerdo”, “En desacuerdo”, “Ni en desacuerdo ni de acuerdo”, “De acuerdo” y “Totalmente de acuerdo”.

Seguidamente se presentan las preguntas incluidas en la evaluación del Asistente de Voz:

- La forma de interactuar con el Asistente es clara desde la primera pregunta.
- La extensión del cuestionario realizado es adecuada.
- Las preguntas evalúan los mismos conceptos más de una vez.
- Se entiende fácilmente cuántas preguntas se han hecho y cuántas faltan para terminar.
- Todos los conceptos incluidos en las preguntas son sencillos de entender.
- Es sencillo saber si el Asistente de Voz entendió mi respuesta.
- Es sencillo saber si el Asistente de Voz entendió la pregunta que le hice.

- El Asistente de Voz hablaba más de lo necesario.
- Era común que, a la hora de responder, se me hubiera olvidado qué era lo que me había preguntado.
- Si me equivoqué al dar una respuesta, es posible corregirla en el momento.
- Es sencillo solicitar ayuda al Asistente de Voz.

Capítulo 5

Implementación del cuestionario por voz

Esta sección describe el proceso de implementación de la interfaz conversacional por voz que permitirá la ejecución del cuestionario UEQ durante el desarrollo del caso de estudio que fue descrito en el capítulo 4. Este proceso incluye desde las iteraciones ejecutadas utilizando el Mago de Oz hasta la implementación final que será utilizada en el caso de estudio.

Para el desarrollo de la primera iteración del Mago de Oz, en la primera iteración participaron 6 personas (3 mujeres y 3 hombres) con edades cuya mediana es de 31 años donde la mayoría (5 personas) trabajan o trabajaban en áreas relacionadas con Tecnología de la Información (TI). La primera ejecución tenía como objetivo determinar la afinidad o predilección hacia un tipo de redacción de pregunta (con escala de 1 a 7 o preguntas segmentadas). De dicha interacción se obtuvieron los siguientes resultados descritos en la Tabla 3. En esta primera iteración se comenzaron a evidenciar ciertos problemas en cuanto al reconocimiento de voz, ciertos usuarios tenían problemas para que la interfaz implementada les reconociera su respuesta de manera correcta mientras que otros no tuvieron problema alguno. Esto se puede deber a problemas ya sea en el “hardware” de los usuarios o problemas con el reconocimiento de voz de Voiceflow en español.

Las preguntas de respuesta libre fueron las que dieron mayor retroalimentación respecto a la interacción y la usabilidad de la interfaz implementada. De esta primera iteración, se tomó toda la retroalimentación que la plataforma Voiceflow permite implementar con tal de mejorar el prototipo para una segunda aplicación. Dichas preguntas permitieron hacer que la interacción con el agente fuera adquiriendo más naturalidad, según la definición de conversación natural definida anteriormente. Las respuestas más destacables se centraron en que el agente permitiera la opción de repetir la pregunta después de cierto tiempo o a pedido del usuario, además de permitir la opción de preguntar por el significado de una palabra o pedir una aclaración.

Tabla 3: Resultados de encuesta de primera iteración del Mago de Oz

Participante	Satisfacción		Naturalidad		Facilidad		Preferencia
	Escala 1 a 7	Segmentadas	Escala 1 a 7	Segmentadas	Escala 1 a 7	Segmentadas	
1	3	3	2	3	3	3	Segmentadas
2	3	4	4	3	5	4	Escala 1 a 7
3	2	4	2	2	2	5	Segmentadas
4	4	4	4	3	5	5	Escala 1 a 7
5	4	3	3	3	5	4	Escala 1 a 7
6	2	3	3	2	4	4	Escala 1 a 7
Promedio	3	3.5	3	2.7	4	4.5	

Uno de los participantes mencionó que sería bueno permitir interrumpir al agente para dar una respuesta, en lugar de esperar a que se enunciara toda la pregunta. Otro participante sugirió la creación de un “avatar” para el agente, esto con la finalidad de hacer la interacción más natural, además de capturar el nombre del encuestado. Uno de los participantes indicó que la redacción de las preguntas, en esta iteración podría inducir al “error”, ya que la interfaz implementada indicaba realizar una evaluación “del sistema” aun cuando lo que se debía evaluar eran productos.

Producto de la retroalimentación recibida de este primer grupo de personas, se implementan las sugerencias que podrían tener un impacto en la interacción del agente con los participantes y este nuevo prototipo fue sometido a uso por otro grupo de personas, donde nuevamente la evaluación que hacia el agente no era el punto principal sino la retroalimentación que pudieran dar los participantes del uso del agente.

Una vez aplicados los cambios sugeridos en los cuestionarios de la primera iteración se procedió a ejecutar la segunda iteración con una participación de 6 personas (5 mujeres y 1 hombre) con edades con una mediana de 28.5 años y cuyas profesiones varían desde psicología hasta administración de negocios. En esta iteración el objetivo sigue siendo determinar afinidad o predilección hacia un tipo de redacción de pregunta (con escala de 1 a 7 o preguntas segmentadas) pero esta vez evaluando el impacto de si se captura el nombre del usuario, si se crea un “avatar” al asistente por voz, cambiar la redacción de las preguntas después de 8 preguntas (simplificación de la redacción), permitir la opción de pedir que se repita la pregunta actual. Para esta iteración se hizo una variación en el cuestionario aplicado a los participantes después de la interacción, descrito en el capítulo 4 y cuyos resultados se pueden observar en la Tabla 4.

Tabla 4: Resultados de encuesta de la segunda iteración del Mago de Oz

Participante	Satisfacción		Naturalidad		Facilidad	Preferencia
	Escala 1 a 7	Segmentadas	Escala 1 a 7	Segmentadas		
1	4	3	5	2	Escala 1 a 7	Escala 1 a 7
2	4	4	3	3	Escala 1 a 7	Segmentadas
3	4	4	5	5	Escala 1 a 7	Escala 1 a 7
4	5	3	2	2	Segmentadas	Segmentadas
5	4	3	5	5	Escala 1 a 7	Escala 1 a 7
6	4	3	4	4	Escala 1 a 7	Escala 1 a 7
Promedio	4.14	3.33	4	3.5		

Una vez ejecutada esta segunda iteración, se obtuvo retroalimentación de los participantes referentes a las mejoras implementadas en el prototipo resaltando la claridad con la que habla el agente y el hecho de que capture el nombre del participante, esto hizo sentir a la persona que estaba en una conversación y no en un interrogatorio. Además, varios participantes resaltaron el hecho de que esta es una implementación inclusiva, ya que permite que poblaciones con algún tipo de discapacidad visual o motora, puedan ser parte de estos cuestionarios. Uno de los participantes recalcó el hecho que notó el cambio en la redacción de las preguntas, lo cual hizo que estuviera más atento, reduciendo el factor de monotonía.

Adicionalmente, un participante resaltó el hecho de que esta implementación le da la libertad realizar una evaluación en cualquier momento y lugar (siempre y cuando tenga acceso a un teléfono celular o un Asistente por Voz) y no lo limita a tener que realizar la evaluación en una computadora con la posibilidad de no encontrar el sitio web para poder completarlo. Sin embargo y a pesar de que los resultados de las Tablas 1 y 2 muestran que los participantes prefieren las preguntas redactadas con la escala de 1 a 7 (más que todo por facilidad para responder), uno de los participantes dijo que a pesar de que es más sencillo de responder preguntas con la escala de 1 a 7, prefiere las preguntas segmentadas, ya que le da la sensación de que forma parte de una conversación y puede dar su opinión de forma más expresiva.

La siguiente interacción entre el participante y el Asistente de Voz, ejemplifica la forma en que se implementó la escala de diferencial semántico, utilizando una pregunta inicial que establece la dirección de la actitud, y una segunda pregunta que identifican la intensidad de esa actitud.

Asistente: Pregunta 1 ¿Usted calificaría Mural como: desagradable, agradable, o ni uno ni otro?
Participante: Agradable
Asistente: ¿Y qué tan agradable? ¿Un poco, mucho o en extremo?
Participante: Mucho
Asistente: Entendido

Ambas preguntas, en conjunto, recogen la información que normalmente estaría en el diferencial semántico, y la convierten en un número entre uno y siete, equivalente a la casilla seleccionada dentro del diferencial semántico.

Completada la fase de diseño y una vez decidido el formato con el cual se van a formular las preguntas, se procedió con la implementación de la interfaz conversacional por voz, que será utilizada para la implementación del cuestionario por voz. El primer cambio implementado fue la modularización de la lectura de las preguntas. Durante la ejecución de las primera y segunda iteración del mago de oz, las preguntas eran una tras otra de forma serial, lo cual hacía que el prototipo fuera poco mantenible, ya que un cambio pequeño en la redacción de las preguntas debe aplicarse en diversos puntos del código. La modularización se hizo aprovechando una funcionalidad de Voiceflow, que permite conectar el prototipo con hojas de cálculo de Google en un "Google Drive". De esta forma se generalizó el código para leer una pregunta, reduciendo considerablemente la complejidad de mantenimiento y centralizando la funcionalidad de la lectura en un solo punto.

Debido a la posibilidad de leer y escribir en una hoja de cálculo desde Voiceflow, se procedió a la automatización de la captura de las respuestas de los usuarios. Dicha captura se hizo tomando el identificador único aleatorio para cada participante, la hora en que dio su respuesta, el número de respuesta y la cuantificación de la respuesta. Se capturó además la cantidad de veces que el usuario solicitaba repetir la pregunta, así como la cantidad de veces que solicitó el significado de las palabras del diferencial semántico que se estaba preguntando.

El siguiente ejemplo muestra un segmento de la evaluación, en la que el participante solicita al Asistente que le repita la pregunta:

Asistente: Pregunta 8 ¿Usted calificaría Mural como: impredecible, predecible, o ni uno ni otro?
 Participante: ¿Podría repetir la pregunta?
 Asistente: Aquí va de nuevo la pregunta. ¿Usted calificaría Mural como: impredecible, predecible, o ni uno ni otro?
 Participante: Ni uno ni otro
 Asistente: ¡Perfecto!

Además de solicitar la repetición de la pregunta inicial, se agregó la capacidad de solicitar la repetición de la segunda pregunta, referente a la intensidad de la actitud del participante.

Asistente: Pregunta 20 ¿Califica a Mural como: ineficiente, eficiente, o ni uno ni otro?
 Participante: Eficiente
 Asistente: ¿Y qué tan eficiente? ¿Un poco, mucho, o en extremo?
 Participante: ¿Podría repetir?
 Asistente: ¿Y qué tan eficiente? ¿Un poco, mucho, o en extremo?
 Participante: Un poco
 Asistente: Listo. Sigamos.

Otra funcionalidad implementada, pensando en la interacción que tendrán los usuarios con el agente es la posibilidad de pedir el significado de las palabras que son parte del diferencial semántico que actualmente se está preguntando, de forma tal que antes de que el usuario de su respuesta, este puede preguntar y estar seguro del significado de cada una de las palabras o frases. El siguiente ejemplo muestra este tipo de interacción:

Asistente: Pregunta 23 ¿Califica a Mural como: ordenado, sobrecargado, o ni uno ni otro?
 Participante: ¿Me podría explicar los conceptos?
 Asistente: Ordenado se refiere a que está organizado apropiadamente y los elementos son claros. Sobrecargado implica que tiene muchos elementos o los presenta de forma desordenada.
 Asistente: ¿Califica a Mural como: ordenado, sobrecargado, o ni uno ni otro?
 Participante: Me parece sobrecargado
 Asistente: ¿Y qué tan sobrecargado? ¿Un poco, mucho, o en extremo?
 Participante: Un poco
 Asistente: Listo. Sigamos

También se agregó al Asistente la posibilidad de que el usuario pregunte el tiempo transcurrido y las preguntas restantes por contestar, con miras a que el usuario tenga un mayor control de tiempo invertido aplicando el cuestionario y sepa en qué punto del proceso está, en relación con la cantidad de preguntas. Por ejemplo:

Asistente: Pregunta 5 ¿Usted calificaría Mural como: valioso, de poco valor, o ni uno ni otro?
 Participante: ¿Cuánto falta?
 Asistente: Han transcurrido 2 minutos y faltan 22 preguntas por responder
 Asistente: Volvamos a las preguntas...
 Asistente: ¿Usted calificaría Mural como: valioso, de poco valor, o ni uno ni otro?
 Participante: Ninguno
 Asistente: Bien

Con la finalidad de darle mayor control al usuario en la interacción con el agente, se agregó la función de una salida forzada, es decir, que si el usuario decide terminar el cuestionario antes de completar las 26 preguntas pueda hacerlo. Una interacción válida de esta funcionalidad sería:

Asistente: Pregunta 20 ¿Califica a Mural como: ineficiente, eficiente, o ni uno ni otro?
 Participante: Quiero terminar el cuestionario
 Asistente: Se van a perder las respuestas dadas... ¿Está seguro de que quiere terminar aquí?
 Participante: Así es
 Asistente: ¡Muchas gracias Marco por participar!

Por otro lado, el usuario podría solicitar terminar la evaluación antes de finalizar, pero luego arrepentirse. Por ejemplo:

Asistente: Pregunta 7 ¿Usted calificaría Mural como: no interesante, interesante, o ni uno ni otro?
 Participante: Quiero finalizar
 Asistente: Se van a perder las respuestas dadas... ¿Está seguro de que quiere terminar aquí?
 Participante: Quiero seguir
 Asistente: Entendido, continuemos con el cuestionario
 Asistente: ¿Usted calificaría Mural como: no interesante, interesante, o ni uno ni otro?

Ahora bien, al igual que el participante puede solicitar se le repita algo que no entendió, el Asistente de Voz también podría tener dificultades para comprender los enunciados del participante, o que las respuestas del participante no formen parte de las opciones válidas que el Asistente está esperando, por lo que se implementaron interacciones en las que el Asistente es quien solicita una repetición.

En la pregunta inicial, esta solicitud podría ser:

Asistente: Pregunta 15. ¿Califica a Mural como: convencional, novedoso, o ni uno ni otro?
 Participante: Me parece extravagante
 Asistente: Disculpe, no le entendí. ¿Convencional, novedoso, o ni uno ni otro?
 Participante: Diría que novedoso
 Asistente: ¿Y qué tan novedoso? ¿Un poco, mucho, o en extremo?
 Participante: En extremo
 Asistente: ¡Perfecto!

Si lo que el Asistente no entendió fue la intensidad, una interacción válida sería:

Asistente: Pregunta 6 ¿Usted calificaría Mural como: aburrido, emocionante, o ni uno ni otro?
 Participante: Aburrido
 Asistente: ¿Y qué tan aburrido? ¿Un poco, mucho, o en extremo?
 Participante: Es difícil de explicar
 Asistente: Disculpe, no le entendí. ¿Un poco, mucho, o en extremo?
 Participante: Creo que mucho
 Asistente: Listo. Sigamos

Otras interacciones que forman parte del diseño del Asistente de Voz son la bienvenida y el cierre de la evaluación. En la bienvenida, el asistente se identifica con su nombre, Emilia, le pide el nombre al participante y le da unas breves instrucciones del cuestionario y de sus capacidades. Por ejemplo:

- Asistente: ¡Hola! mi nombre es Emilia. Antes de continuar, ¿me podría decir su nombre?
- Participante: Marco
- Asistente: ¡Gracias Marco! Este cuestionario cuenta con 26 preguntas acerca de Mural. Si desea que le repita la pregunta o que le explique alguno de los conceptos, por favor indíquelo.
¡Comencemos!

Cuando el participante respondió las 26 preguntas y la evaluación ha finalizado, el Asistente se despide y le agradece al participante, de la siguiente forma:

- Asistente: Pregunta 26 ¿Califica a Mural como: conservador, innovador, o ni uno ni otro?
- Participante: Ninguno de los dos
- Asistente: Hemos terminado. ¡Muchas gracias Marco por contestar este cuestionario!

Un punto a mencionar es que, con el propósito de poder identificar diferentes posibles enunciados o formas de expresar instrucciones que pueda dar un usuario ante una tarea como solicitar repetir una pregunta, el significado de las palabras, de terminar la interacción o de solicitar el tiempo transcurrido, se aplicó un ejercicio de lluvia de ideas con los estudiantes del curso Diseño de Interacción Persona Computador, del posgrado en Computación de la Universidad de Costa Rica.

En este ejercicio se recolectó un cuerpo importante de expresiones, que se detallan en el Anexo 1. Estas expresiones fueron analizadas y las recomendadas de forma más frecuente fueron agregadas al agente en Voiceflow en forma de declaraciones o enunciados (utterances).

En las siguientes figuras se muestra el diseño final la interfaz conversacional por voz implementado en VoiceFlow. La Figura 3 muestra la implementación el mensaje de bienvenida, así como las instrucciones generales que da el Asistente al inicio de la interacción.



Figura 3: Implementación de mensaje de bienvenida e instrucciones generales del Asistente

La Figura 4 muestra la implementación de las preguntas segmentadas. Se nota como en una primera parte se define hacia que lado del diferencial semántico se inclina la respuesta y después se determina el grado de intensidad de esta.

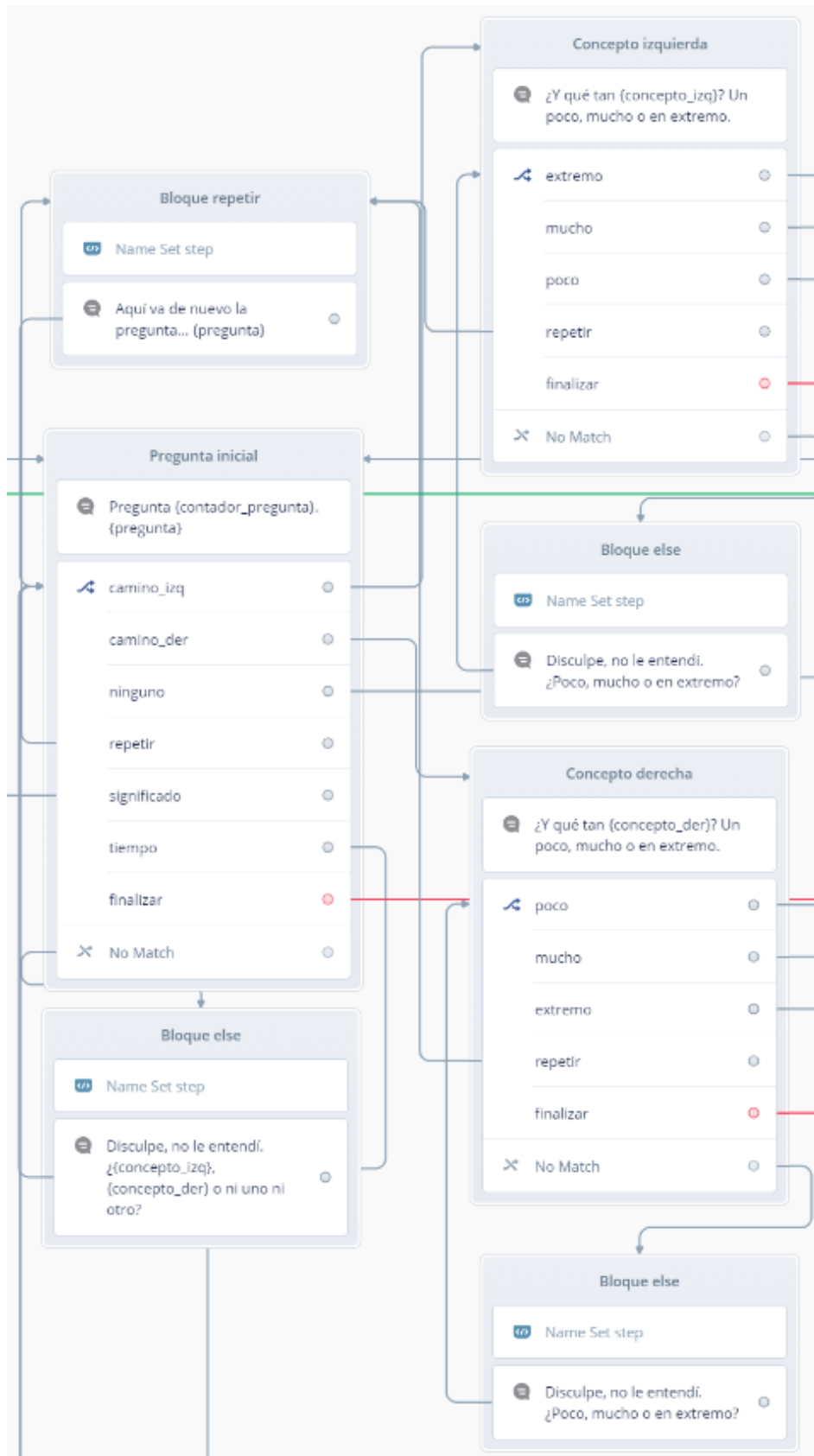


Figura 4: Implementación de las preguntas segmentadas.

La forma en la que se transforma la respuesta del usuario a un formato numérico y su posterior escritura en una hoja de cálculo se puede observar en la Figura 5.

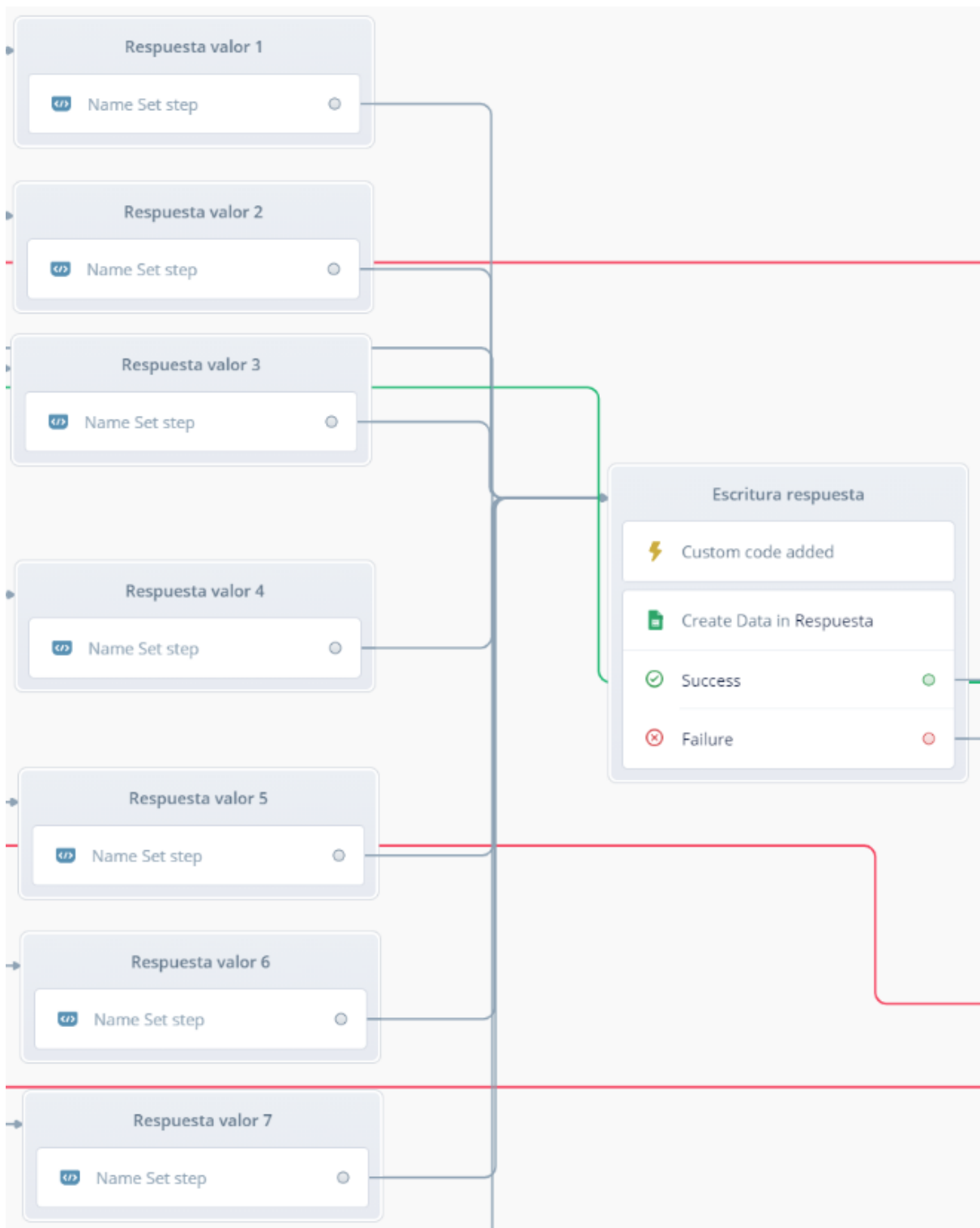


Figura 5: Implementación de la conversión y escritura de respuestas

La implementación del mecanismo de salida de la interacción se muestra en la Figura 6. En dicha imagen se observa la salida al completar las 26 preguntas del cuestionario UEQ o cuando el usuario decide salir de la interacción.

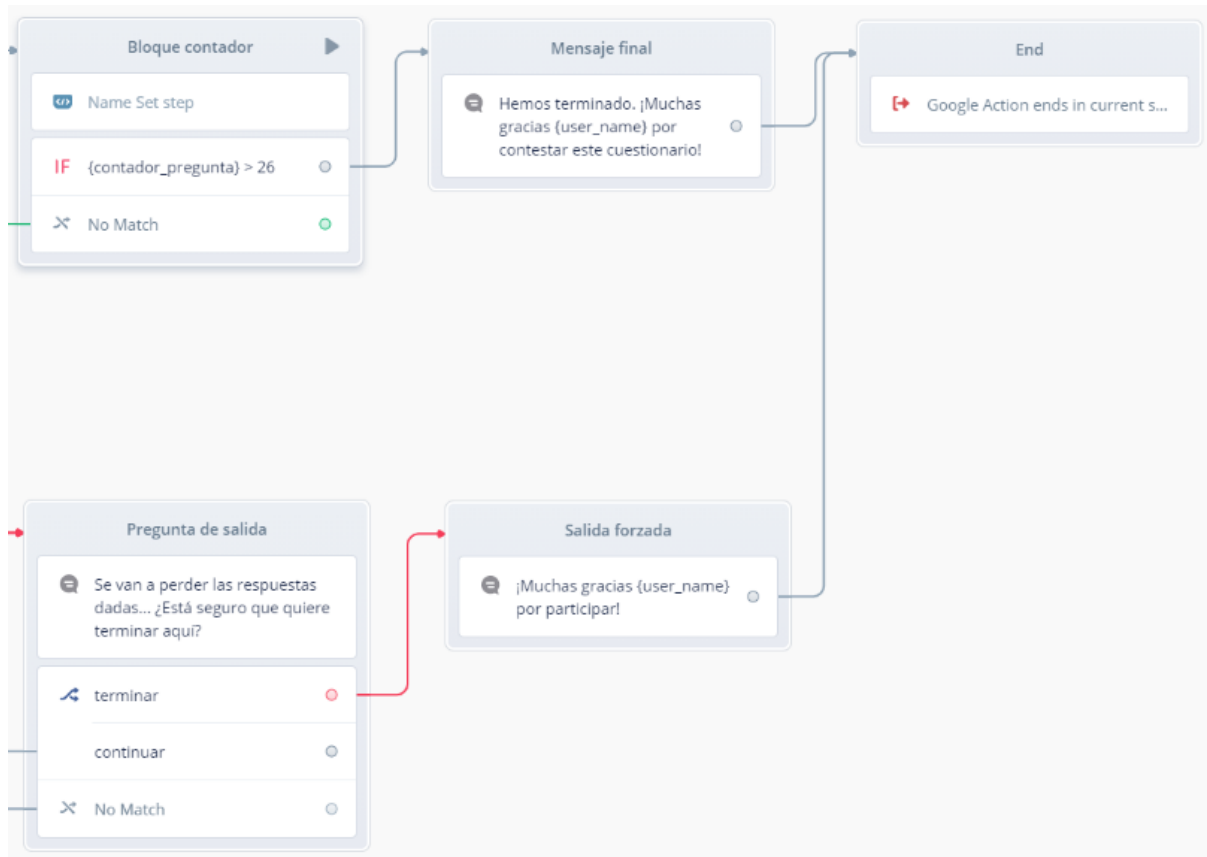


Figura 6: Implementación del mecanismo de salida de la interacción

Capítulo 6

Caso de Estudio: Mural

Esta sección describe los resultados obtenidos al aplicar el caso de estudio con 44 participantes, donde 33 son hombres, 9 son mujeres y 2 participantes prefirieron no revelar su género, con edades que oscilan entre los 19 y los 36 años. Para el caso de estudio los participantes estaban divididos en dos grupos, el Grupo 1 con 20 participantes y el Grupo 2 con 24 participantes.

Una vez ejecutado el caso de estudio se obtienen dos evaluaciones. La primera de ellas es la evaluación realizada a la plataforma colaborativa MURAL utilizando la interfaz conversacional, mientras que la segunda es la evaluación aplicada a la interfaz conversacional utilizando UEQ+ y un conjunto de preguntas adicionales sobre la Usabilidad del asistente.

6.1 Evaluación de la plataforma MURAL mediante el cuestionario UEQ implementado con una interfaz conversacional

La evaluación de la plataforma MURAL utilizando una interfaz conversacional que aplica UEQ, se realizó con los dos grupos de participantes. Cabe resaltar que ambos grupos de participantes ya habían utilizado en el pasado dicha plataforma, por lo que, para esta ejecución, no se tuvo que realizar una ronda de utilización de esta previo a la evaluación utilizando la interfaz conversacional.

En ambos grupos, los participantes ejecutaron el Asistente de Voz, implementado en Voiceflow, a través de un enlace que se les facilitó y dicha ejecución debía realizarse desde una computadora con acceso a internet. Las preguntas realizadas por el Asistente de Voz seguían el formato de preguntas segmentadas, es decir, cada una de las preguntas estaba dividida en dos partes donde la primera parte corresponde a ubicar la respuesta del participante en cualquiera de los dos términos del diferencial semántico, mientras que la segunda pregunta determina el grado de intensidad de dicha respuesta (poco, mucho o en extremo). Las respuestas dadas por cada uno de los participantes se graban automáticamente en una hoja de cálculo de Google,

utilizando la integración provista por Voiceflow. Posteriormente, estos resultados se introdujeron en la hoja de cálculo que provee la herramienta de UEQ, que permite analizar la experiencia de usuario de los participantes, clasificando la experiencia en seis subescalas: atracción, transparencia, eficiencia, controlabilidad, estimulación y novedad.

En la Figura 7 se muestran los resultados de la evaluación de la plataforma MURAL, de los participantes del Grupo 1, para cada una de las subescalas de UEQ. Dicho gráfico muestra que la característica de transparencia es la mejor evaluada por este grupo mientras que controlabilidad es la peor evaluada, sin embargo, esta evaluación representa una evaluación neutral y no una evaluación negativa según UEQ.

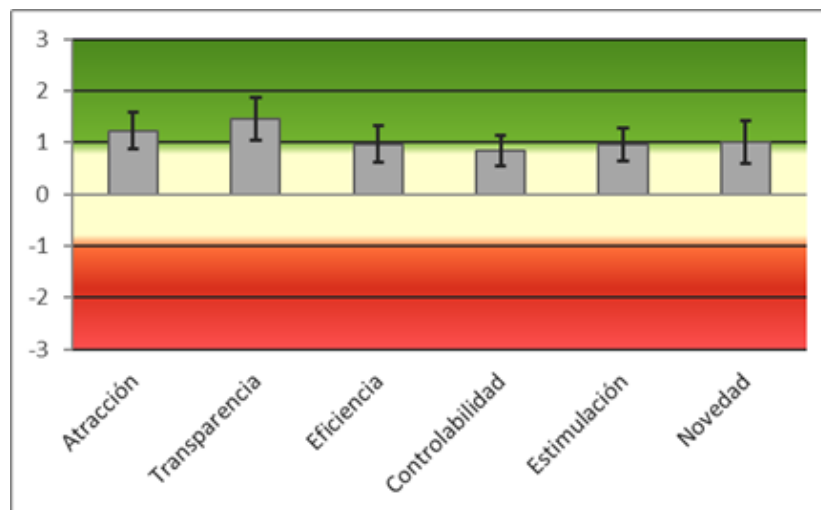


Figura 7: Gráfico de resultados UEQ MURAL Grupo 1.

Observando los resultados en la Tabla 5 de medias y varianzas para las características que se evalúan en UEQ (atracción, transparencia, eficiencia, controlabilidad, estimulación y novedad), tanto en la Figura 6 como en la Tabla 5, los valores presentados, se dan en una escala de -3 a 3, en donde los valores mayores a 0.8 se catalogan como una evaluación positiva, los valores entre -0.8 y 0.8 se catalogan como una evaluación neutral y los valores menores a -0.8 se catalogan como una evaluación negativa. En ambos casos, los datos muestran que la evaluación hecha por el Grupo 1 a la plataforma MURAL, se puede catalogar como positiva, ya que todos los valores obtenidos, en términos de medias, son mayores a 0.8.

Tabla 5: Escala UEQ (media y varianza) evaluación de MURAL Grupo 1.

Escala UEQ (Media y Varianza)		
Atracción	1.23	0.66
Transparencia	1.45	0.90
Eficiencia	0.96	0.65
Controlabilidad	0.84	0.44
Estimulación	0.96	0.55
Novedad	1.00	0.90

En la Figura 8 se muestran los resultados de las evaluaciones de los participantes del Grupo 2. Dicho gráfico muestra que la característica de controlabilidad es la mejor evaluada por este grupo mientras que transparencia es la peor evaluada. Sin embargo, esta evaluación representa una evaluación neutral y no una evaluación negativa según UEQ.

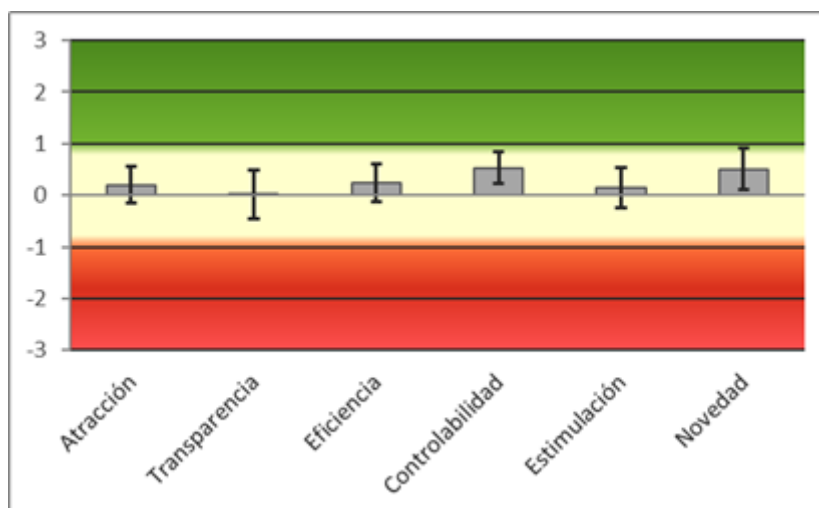


Figura 8: Gráfico de resultados UEQ MURAL grupo 2.

La Tabla 6 muestra la escala UEQ de media y varianza. Tanto la Figura 7 como la Tabla 6 muestran que la evaluación realizada por el Grupo 2 de participantes se puede catalogar como una evaluación neutral, ya que todos sus valores, en términos de media, se ubican en el rango de -0.8 a 0.8.

Tabla 6: Escala UEQ (media y varianza) evaluación de MURAL Grupo 2.

Escalas UEQ (Media y Varianza)		
Atracción	0.20	0.76
Transparencia	0.01	1.41
Eficiencia	0.24	0.86
Controlabilidad	0.53	0.60
Estimulación	0.15	0.95
Novedad	0.50	1.03

Cabe destacar que al hacer una comparación de los resultados de ambos grupos se observa como las características mejor y peor evaluadas por el Grupo 1, se invirtieron en la evaluación realizada por el Grupo 2. Lo anterior se evidencia en el hecho en que para el Grupo 1 la característica mejor evaluada es la transparencia mientras que para el Grupo 2, la transparencia es la característica peor evaluada. Además, para el Grupo 1 la característica de controlabilidad es la peor evaluada mientras que para el Grupo 2 esa es la característica mejor evaluada.

Al hacer una evaluación conjunta, utilizando los resultados de ambos grupos, se evidencia que en general, la evaluación realizada a la plataforma MURAL, se puede catalogar como mala, tal y como lo muestran el Figura 9 y la Tabla 7 de la escala UEQ de media y varianza.

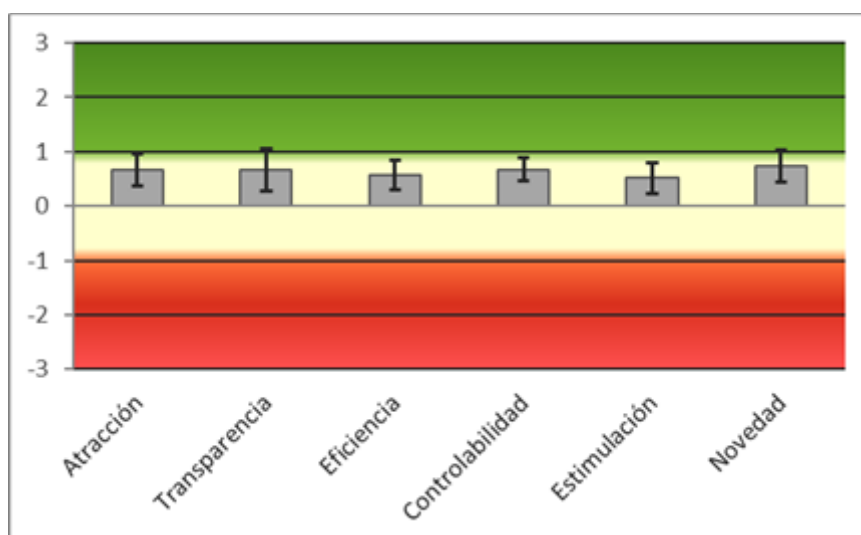


Figura 9: Gráfico de resultados conjuntos UEQ MURAL Grupo 1 y Grupo 2.

Tabla 7: Escala UEQ (media y varianza) evaluación conjunta de MURAL Grupo 1 y Grupo 2.

Escala UEQ (Media y Varianza)		
Atracción	0.67	0.97
Transparencia	0.66	1.68
Eficiencia	0.57	0.88
Controlabilidad	0.67	0.54
Estimulación	0.52	0.92
Novedad	0.73	1.01

La Figura 10 muestra la evaluación realizada por ambos grupos de participantes comparada con un conjunto de 452 estudios previos de UEQ. Puede notarse que los participantes catalogan a MURAL por debajo del promedio e incluso como un producto malo. Esto coincide con algunas opiniones expresadas por los participantes al finalizar la aplicación de las evaluaciones, en las que los participantes expresaron que la herramienta en sí no era muy buena.

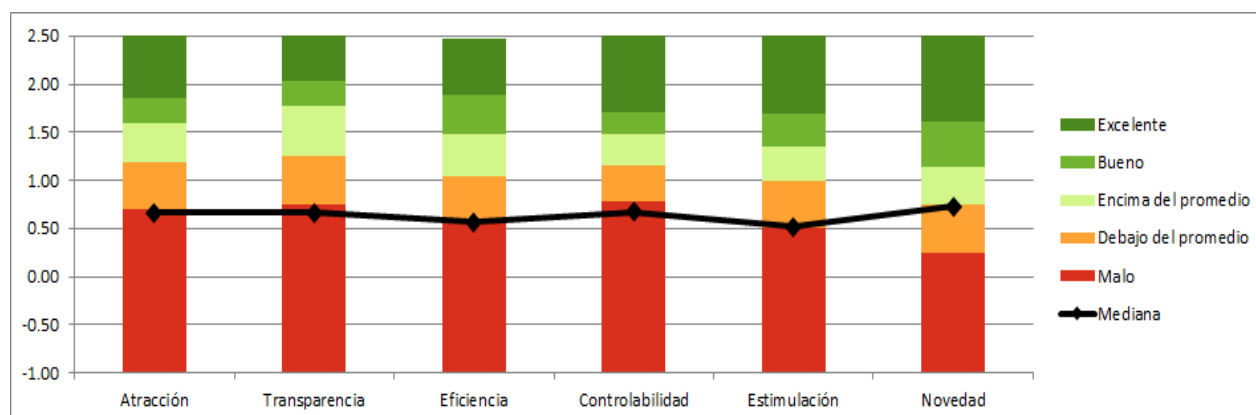


Figura 10: "Benchmark" de evaluación de MURAL comparado a 452 estudios previos de UEQ.

Estas opiniones se dieron espontáneamente y no pretenden ser una confirmación formal de los resultados del UEQ sobre la experiencia de los participantes con MURAL. No se había planificado una comparación de UEQ con otro instrumento de evaluación dado que el objetivo del caso de estudio no era evaluar MURAL, sino la implementación del cuestionario UEQ con una interfaz conversacional por voz en la recolección de las respuestas de los participantes.

Otro elemento que se identifica en las hojas de cálculo de UEQ son las inconsistencias de las respuestas. Una respuesta se cataloga como inconsistente cuando la diferencia en las características evaluadas es mayor a 3, aunque tales inconsistencias se

pueden deber a respuestas aleatorias dadas por el usuario o por malentender una pregunta.

En la Tabla 8 se listan los resultados de los 12 participantes que presentaron inconsistencias en sus respuestas, donde cada fila representa a cada participante. Los restantes 32 participantes no mostraron respuestas inconsistentes en ninguna de las seis subescalas. De los 12 participantes con inconsistencias puede verse que dos de ellos tuvieron dos inconsistencias y 10 solamente una subescala considerada como inconsistente.

Los autores del cuestionario UEQ recomiendan que, si la respuesta de un participante es inconsistente en tres o más subescalas, dato que se calcula en la columna Criticidad, lo apropiado es no considerar esa respuesta en el análisis de los datos. Puede verse que, en este caso, ninguna de las 44 respuestas alcanza las tres inconsistencias, por lo que la totalidad de 44 respuestas se considera apropiada para el análisis. Este dato refuerza la confianza en la implementación de la interfaz conversacional por voz como mecanismo de captura para UEQ, implementado en este trabajo.

Tabla 8: Escala con respuesta inconsistentes

Atracción	Transparencia	Eficiencia	Controlabilidad	Estimulación	Novedad	Criticidad
1		1				2
		1	1			2
				1		1
1						1
1						1
1						1
			1			1
			1			1
			1			1
	1					1
	1					1
	1					1

6.2 Evaluación del asistente por voz implementado mediante UEQ+

En la aplicación del caso de estudio, se decidió implementar una variante entre los grupos de participantes. La variante consistió en que al Grupo 1 no se le dieron instrucciones detalladas de las funcionalidades del Asistente de Voz, más allá de las instrucciones que brinda el propio asistente al inicio de la interacción, mientras que al Grupo 2 sí se le dieron instrucciones detalladas de las diferentes funcionalidades del asistente y cómo utilizarlas.

Al iniciar el Asistente de Voz, ésta indica: “Este cuestionario cuenta con 26 preguntas acerca de Mural. Si desea que le repita la pregunta o que le explique alguno de los conceptos, por favor indíquelo. ¡Comencemos!”.

Esas fueron las únicas instrucciones que se le dieron a los participantes del Grupo 1. Por otro lado, el Grupo 2 escuchó estas instrucciones del Asistente de Voz, pero también fueron instruidos por la persona que facilitó la evaluación, que el Asistente de Voz tenía varias características conversacionales que podían explotar. En esta explicación se mencionaron las capacidades del Asistente de Voz de repetir una pregunta, explicar el significado de los conceptos que forman cada pregunta (correspondientes al diferencial semántico), identificar la cantidad de preguntas que se han realizado y cuántas faltan por realizar, así como el tiempo que ha transcurrido desde que se inició la evaluación, y también la capacidad de terminar el cuestionario antes de responder a todas las preguntas si así lo solicita el participante.

Dicha variante se aplicó con la intención de determinar, en términos de usabilidad y experiencia de usuario, si impacta positiva o negativamente la evaluación que hacen los participantes del Asistente de Voz al entender mejor las capacidades de esta.

En esta evaluación del Asistente de Voz implementado, utilizando UEQ+ como herramienta, se evaluaron los aspectos de comprensibilidad, comportamiento de respuesta, calidad de respuesta, definidos en [41] y además se evaluaron los aspectos de uso intuitivo y originalidad como parte de esta investigación. Dicha evaluación se realizó a través de un cuestionario implementado por medio de Google Forms, en total eran 20 preguntas (4 por cada aspecto a evaluar), de las cuales se obtuvieron los resultados que se muestran a continuación.

El Grupo 1 de participantes, a los cuales no se les brindó mayores instrucciones de uso más allá de las que brinda la propia interfaz al inicio de la interacción (como la posibilidad de solicitar repetir una pregunta por ejemplo) y tal y como lo muestra la Figura 10, se evidencia como el uso intuitivo es la característica mejor evaluada por el Grupo 1, a pesar de no recibir mayores instrucciones de uso del asistente, mientras que la característica de comportamiento de respuesta es la peor calificada. En la Tabla 9, se evidencia como en las diferentes preguntas que componen la subescala de uso intuitivo, presenta las medias más altas de los aspectos evaluados, mientras que los valores de las medias que componen la subescala de comportamiento de respuesta son las más bajas. Presenta además el valor de la media de la escala mostrada en la Figura 11.

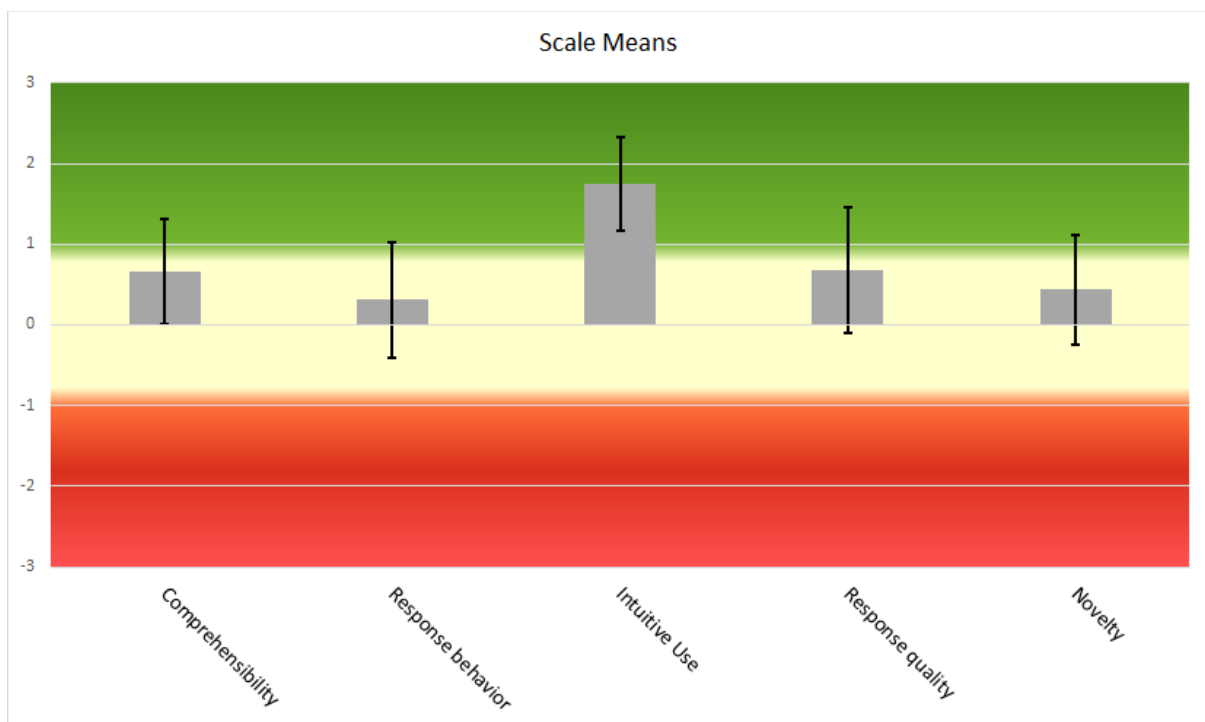


Figura 11: Escala de medias evaluación UEQ+ Grupo 1

La Figura 11 utiliza los mismos colores que los presentes en el cuestionario UEQ clásico, donde los valores entre -0,8 y 0,8 representan una evaluación más o menos neutral de la escala correspondiente, los valores mayores a 0,8 representan una evaluación positiva y los valores menores a -0,8 representan una evaluación negativa. Estas valoraciones son determinadas por los creadores de UEQ y se basan en los resultados históricos que almacenan de diferentes estudios sobre las seis escalas de UEQ: Atracción, Transparencia, Eficiencia, Controlabilidad, Estimulación y Novedad. Sin embargo, UEQ+ no cuenta con registros históricos, por lo que lo que se considera

una evaluación positiva, neutra o negativa, para las escalas seleccionadas en este estudio, podría variar.

Cabe destacar que únicamente 3 personas solicitaron 9 veces en total al Asistente de Voz que les repitiera una pregunta, sin embargo, ninguno de ellos solicitó que se les dijera el significado de alguna de las palabras que componen el diferencial semántico. Además, ninguno de los participantes preguntó en qué punto o por cual pregunta de cuestionario estaban, rescatando el hecho de que el agente indica al principio cuál pregunta es.

Es importante mencionar que el Asistente de Voz indicó en 181 veces en total que no entendía la respuesta del participante lo cual puede generar inconformidad en el participante. Sin embargo, esto no puede ser atribuible a la implementación del Asistente de Voz en sí, ya que el reconocimiento de voz no es parte de la programación implementada, sino que se hace uso de las capacidades con las que cuenta la plataforma Voiceflow en esta área

Tabla 9: Medias y Varianzas en las subescalas Grupo 1

Media y Varianza por ítem				
Escala	Media de escala	Item izquierdo	Item derecho	Media
Comprensibilidad	0.66	Complicado	Simple	0.40
		Ambiguo	Inequívoco	0.30
		Inexacto	Exacto	0.60
		Enigmático	Explicable	1.35
Comportamiento de respuesta	0.31	Artificial	Natural	-0.80
		Desagradable	Agradable	1.05
		Antipático	Simpático	1.15
		Aburrido	Entretenido	-0.15
Uso intuitivo	1.75	Difícil	Fácil	2.10
		Ilógico	Lógico	1.75
		Equívoco	Evidente	1.60
		Incoherente	Coherente	1.55
Calidad de respuesta	0.68	Inapropiado	Adecuado	1.70
		Inútil	Útil	1.20
		No provee ayuda	Provee ayuda	-0.65
		Ignorante	Inteligente	0.45
Novedad	0.44	Falta imaginación	Creativo	1.00
		Convencional	Original	0.25
		Tradicional	Novedoso	0.35
		Conservador	Innovador	0.15

Además de la evaluación de UEQ+, se realizaron una serie de preguntas en formato de escala Likert, esto con el fin de determinar aspectos de Usabilidad que no se incluyen en el UEQ+, como por ejemplo si perciben que el cuestionario es muy largo o si se repiten preguntas. En la Figura 12 se evidencia como la pregunta “La forma de interactuar con el Asistente es clara desde la primera pregunta” ocupa un gran nivel aceptación por parte de los participantes del Grupo 1, resaltando aún más el hecho de que el uso intuitivo fue la característica mejor evaluada por este grupo. Además, se nota como la pregunta “Es sencillo solicitar ayuda al Asistente de Voz” ocupa la peor calificación, resaltando así que la característica de comportamiento de respuesta es la peor evaluada por este grupo.

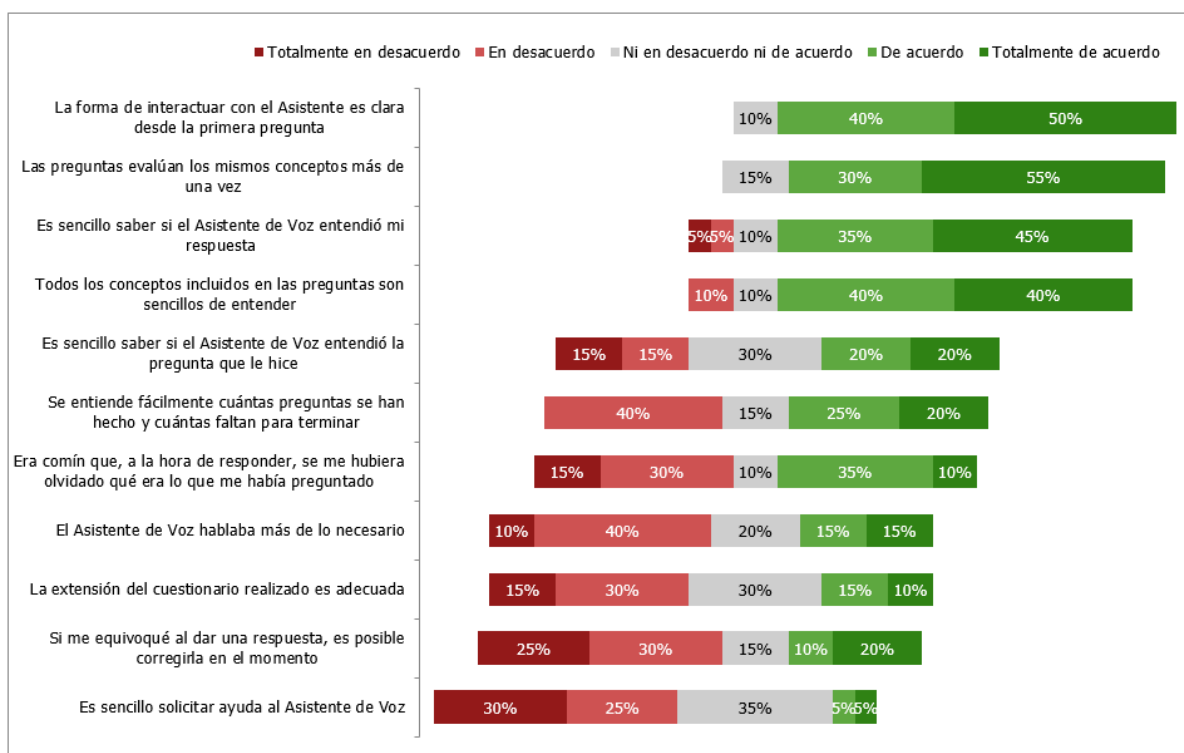


Figura 12: Respuesta de escala Likert Grupo 1

Adicionalmente, como parte de la evaluación del Asistente de Voz, se les solicitó a los participantes, describir el Asistente de Voz con tres palabras, la Figura 13 muestra la nube de palabras obtenida de dichas respuestas. Resalta el hecho de que las palabras “lento” y “repetitivo” tengan tanta presencia como la palabra “interesante”, demostrando desde otra perspectiva el hecho de que la característica “uso intuitivo” tenga la mejor valoración en este grupo y la característica “comportamiento de respuesta” tenga la peor valoración en este grupo. Resulta también interesante como

las palabras “innovador” y “monótono” tienen una presencia similar, al igual que “tedioso” e “intuitivo”, conceptos que de alguna manera se pueden considerar como polos opuestos entre sí.



Figura 13: Nube de palabras con las respuestas del Grupo 1

Al igual que el Grupo 1, a los participantes del Grupo 2 se les solicitó evaluar la interacción que tuvieron con el Asistente de Voz utilizando una evaluación UEQ+ implementada mediante un Google Forms. En dicho formulario se evaluaron nuevamente los aspectos de comprensibilidad, comportamiento de la respuesta, calidad de las respuestas, uso intuitivo y originalidad. Sin embargo, en esta ocasión se tiene una diferencia significativa en la interacción, ya que previo a que los participantes hicieran uso del Asistente de Voz, se les instruyó sobre las diferentes capacidades y las diferentes formas de interacción que pueden tenerse con el asistente, es decir se les indicó qué tipo de solicitud de ayuda pueden hacer (repetir una pregunta, solicitar el significado de las palabras) y las diferentes maneras de verbalizar dichas solicitudes.

Concluida la explicación, los participantes comenzaron a utilizar el agente y como se muestra en la Figura 14 todos los aspectos evaluados, mejoraron considerablemente respecto a los resultados del Grupo 1. Cabe destacar que al igual que en el Grupo 1, Uso intuitivo es el mejor evaluado por el grupo 2, mientras que Comprensibilidad y Calidad de respuesta comparten la peor evaluación.

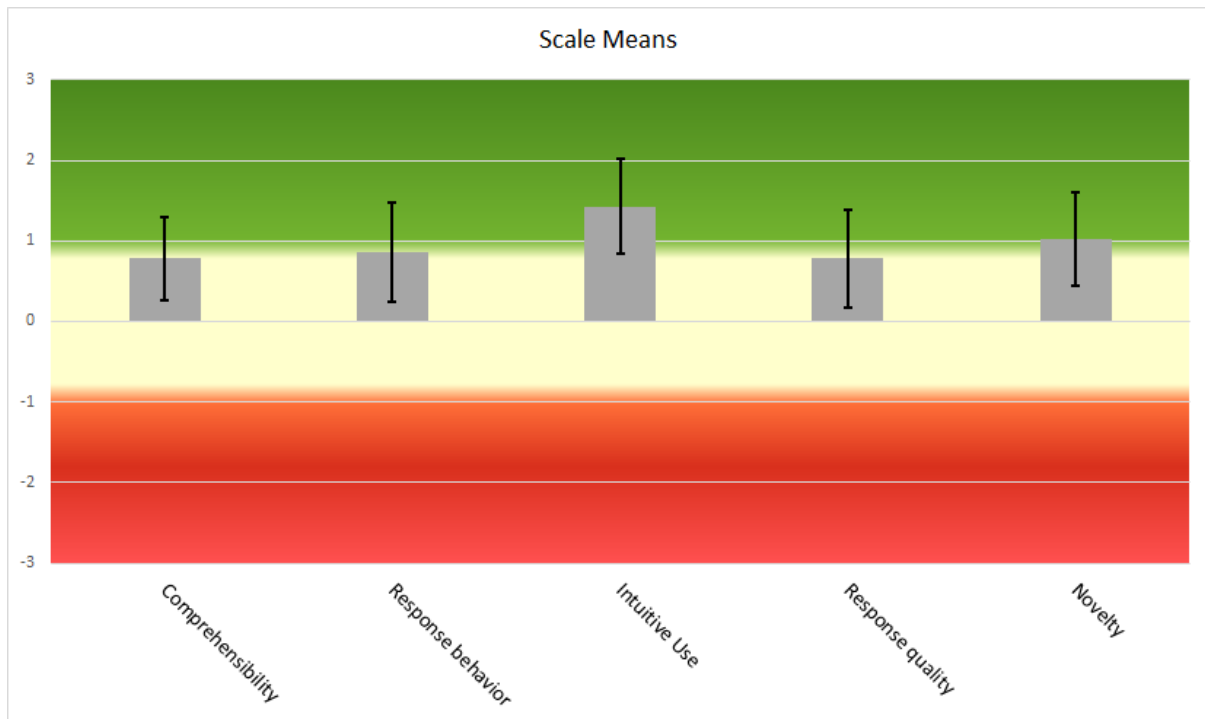


Figura 14: Escala de medias evaluación UEQ+ Grupo 2

La Figura 14 utiliza los mismos colores que los presentes en el cuestionario UEQ clásico, donde los valores entre -0,8 y 0,8 representan una evaluación más o menos neutral de la escala correspondiente, los valores mayores a 0,8 representan una evaluación positiva y los valores menores a -0,8 representan una evaluación negativa. Estas valoraciones son determinadas por los creadores de UEQ y se basan en los resultados históricos que almacenan de diferentes estudios sobre las seis escalas de UEQ: Atracción, Transparencia, Eficiencia, Controlabilidad, Estimulación y Novedad. Sin embargo, UEQ+ no cuenta con registros históricos, por lo que lo que se considera una evaluación positiva, neutra o negativa, para las escalas seleccionadas en este estudio, podría variar.

En la Tabla 10, se evidencia aún más la mejora significativa en la evaluación de los diferentes aspectos que se abordaron en el UEQ+, siendo particularmente interesante, como los aspectos peor evaluados por el Grupo 2 tiene una mejor

calificación que el segundo aspecto mejor evaluado por el Grupo 1. Puede verse como en las diferentes preguntas que componen la subescala de uso intuitivo, se presentan las medias más altas de los aspectos evaluados, mientras que los valores de las medias que componen las subescalas de Comprensibilidad y Calidad de respuesta son las más bajas. Presenta además el valor de la media de la escala mostrada en la Figura 14.

Tabla 10: Medias y Varianzas en las subescalas Grupo 2

Media y Varianza por ítem				
Escala	Media de escala	Item izquierdo	Item derecho	Media
Comprensibilidad	0.78	Complicado	Simple	0.88
		Ambiguo	Inequívoco	0.54
		Inexacto	Exacto	0.58
		Enigmático	Explicable	1.13
Comportamiento de respuesta	0.86	Artificial	Natural	0.50
		Desagradable	Agradable	1.38
		Antipático	Simpático	1.13
		Aburrido	Entretenido	0.46
Uso intuitivo	1.43	Difícil	Fácil	1.46
		Ilógico	Lógico	1.71
		Equívoco	Evidente	1.13
		Incoherente	Coherente	1.42
Calidad de respuesta	0.78	Inapropiado	Adecuado	1.54
		Inútil	Útil	1.25
		No provee ayuda	Provee ayuda	-0.21
		Ignorante	Inteligente	0.54
Novedad	1.02	Falto imaginación	Creativo	1.38
		Convencional	Original	0.67
		Tradicional	Novedoso	1.08
		Conservador	Innovador	0.96

De forma similar al igual que el Grupo 1, cinco participantes distintos del Grupo 2 solicitaron repetir una pregunta en 11 ocasiones (en contraste con las tres personas del grupo 1). En este segundo grupo, el Asistente de Voz indicó en 243 ocasiones al participante que no entendía su respuesta, resaltando el hecho de que todos los participantes no les entendió una respuesta al menos una vez. De estas 243 solicitudes de repeticiones, 44 corresponden a un único participante, en donde el Asistente le indicó que no le entendía en 14 de las 26 preguntas del cuestionario, lo que parecería indicar que el sistema de micrófono que disponía era deficiente.

Como se indicó anteriormente, los errores por parte del Asistente al entender una respuesta de un participante no pueden ser atribuibles a la programación del Asistente, ya que simplemente se utilizaron las capacidades de reconocimiento de voz que provee la plataforma Voiceflow. Lo elevado de estos números se puede deber

a factores como ruido ambiente o la imposibilidad que tiene el Asistente para entender la pronunciación del participante y no por no haber programado las verbalizaciones utilizadas por el asistente.

Al igual que con los participantes del Grupo 1, además de la evaluación de UEQ+, se realizaron una serie de preguntas en formato de escala Likert, esto con el fin de determinar aspectos que no se incluyen en el UEQ+.

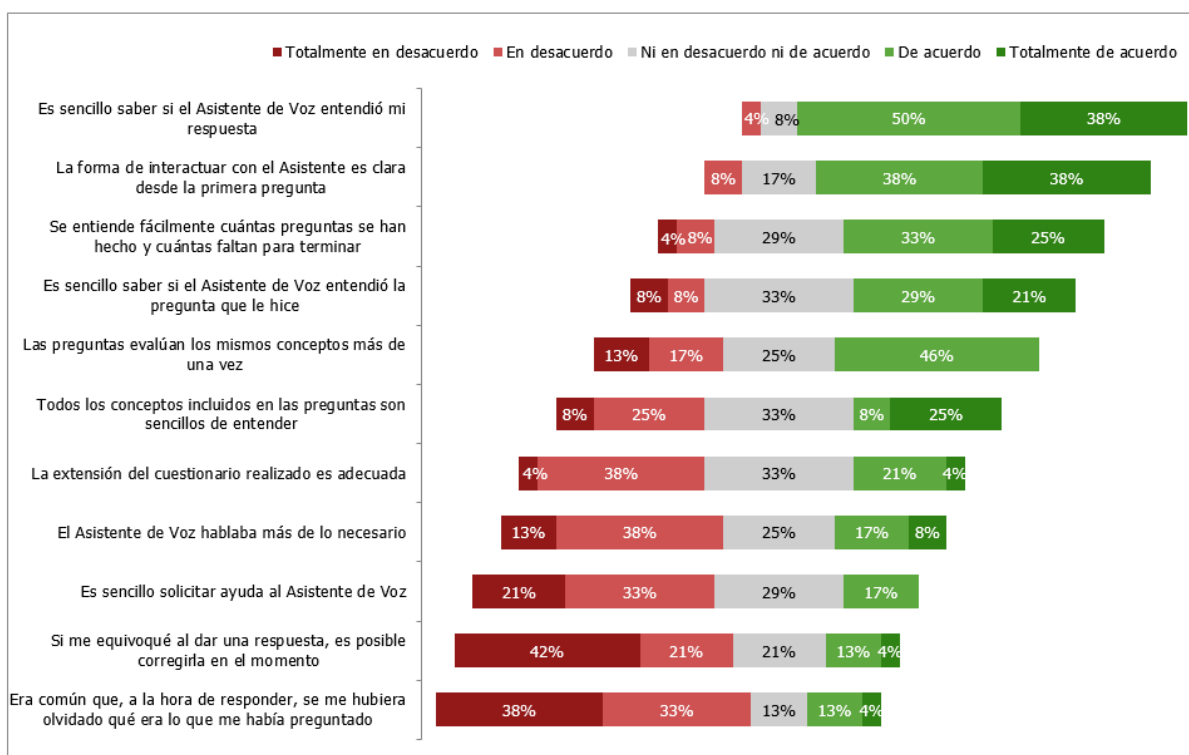


Figura 15: Respuesta de escala Likert Grupo 2

En la Figura 15 se evidencia como la pregunta “Es sencillo saber si el Asistente de Voz entendió mi respuesta” ocupa un gran nivel aceptación por parte de los participantes del Grupo 2, resaltando el hecho de que el uso intuitivo fue la característica mejor evaluada por este grupo y contrastando con la gran cantidad de veces que el Asistente no le entendió una respuesta a los participantes. Además, se nota como la pregunta “Era común que, a la hora de responder, se me hubiera olvidado qué era lo que me había preguntado” ocupa la peor calificación, esto se puede deber a falta de concentración al momento de responder el cuestionario, pero también podría estar influenciado por el hecho de que si el cuestionario tiene implementada la opción de Repetir pregunta, y se les hizo énfasis en este y otros conceptos antes de iniciar la prueba, los participantes quedarán con la idea de que

recordar las preguntas puede ser un inconveniente de las interfaces de voz, y que esto se reflejara en sus respuestas.

Al igual que el Grupo 1, el Grupo 2 contestó con tres palabras cómo describiría al asistente, la Figura 16 muestra el resultado de dichas respuestas, en una nube de palabras. En dicha figura se observa como las palabras “entendible” y “sencillo” resaltan del resto, comprobando desde otra perspectiva el hecho de que la característica de “uso intuitivo” sea la mejor evaluada por los participantes del Grupo 2. Resulta también interesante, como la cantidad de palabras que puede ser catalogadas como positivas (innovador, creativo, interesante, inteligente) aumentan en comparación con el Grupo 1, evidenciando desde otro ángulo la mejora significativa en la evaluación por parte de los participantes, tal y como se mostró en la Figura 14 y la Tabla 9.



Figura 16: Nube de palabras con las respuestas del Grupo 2

Al hacer una comparación de los resultados de las evaluaciones del Grupo 1 y Grupo 2 realizadas con el cuestionario UEQ+, tal y como se observa en la Tabla 11, en cuatro de las cinco escalas, la media del grupo 2 fue superior a la media del grupo 1.

Tabla 11: Comparación de media de las subescalas de UEQ+ para el Grupo 1 y Grupo 2

Pregunta	Grupo 1 (N = 20)				Grupo 2 (N = 24)			
	Media	Desviación estándar	Intervalo de confianza		Media	Desviación estándar	Intervalo de confianza	
Comprensibilidad	0.66	1.01	0.22	1.11	0.78	1.15	0.32	1.24
Comportamiento de respuesta	0.31	1.04	-0.14	0.77	0.86	1.3	0.34	1.39
Uso intuitivo	1.75	1.09	1.27	2.23	1.43	1.39	0.87	1.98
Calidad de respuesta	0.68	0.73	0.36	0.99	0.78	0.75	0.48	1.08
Novedad	0.44	1.26	-0.12	0.99	1.02	1.34	0.48	1.56

En cuanto a las escalas de comprensibilidad, comportamiento de respuesta, calidad de respuesta y novedad, muestran un aumento en las opiniones favorables de los participantes, situación que podría estar influenciada por la explicación detallada de las capacidades del asistente, que se realizó de previo a su uso. La escala de novedad, por ejemplo, aumenta la media de forma considerable, indicando la posibilidad de que los participantes encontrasen el Asistente de Voz más novedoso cuando tenían más información sobre las capacidades del mismo.

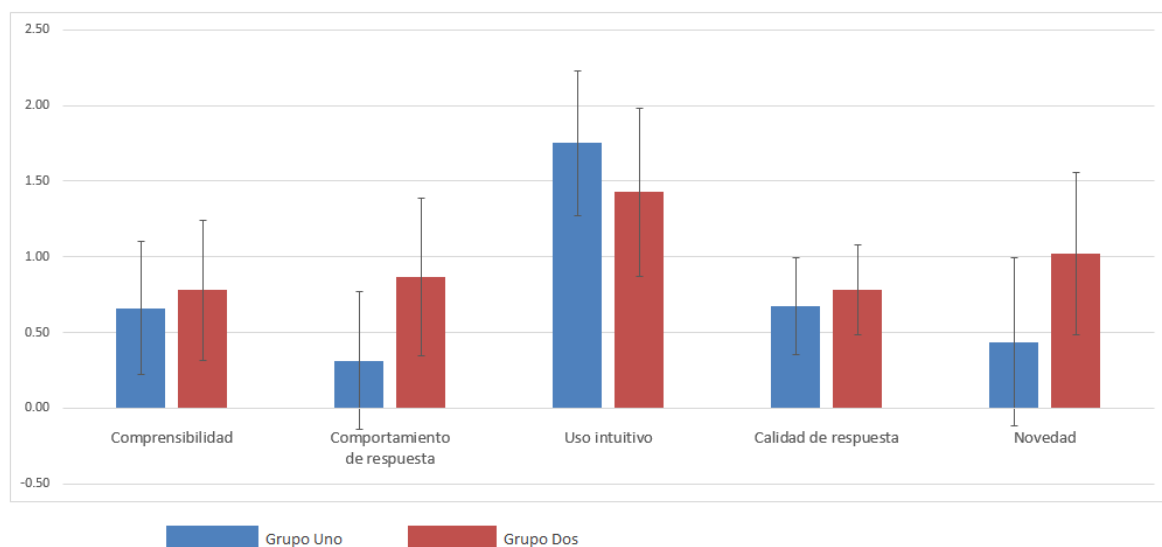


Figura 17: Comparación de las medias del Grupo 1 y Grupo 2, en las 5 subescalas evaluada en UEQ+

La Figura 17 muestra el comportamiento mencionado en la Tabla 11, además de mostrar el hecho de que la escala de uso intuitivo descendió en términos de apreciación entre los participantes del Grupo 2 respecto a los participantes del Grupo 1. El conocer de manera previa las capacidades con las que cuenta el asistente pudo haber influido en que la percepción de los participantes penalizara que varias de las

capacidades implementadas no son intuitivas para el usuario, y que quedan de alguna manera, escondidas en el Asistente de Voz, pero no visibles al usuario. Los participantes del Grupo 1 perciben al Asistente de Voz como más intuitivo, dado que lo consideran con un conjunto menor de características, y este conjunto es más evidente para ellos.

Comparando ahora las preguntas adicionales de Usabilidad realizadas utilizando escalas Likert, en 6 de las preguntas el Grupo 1 tuvo medias superiores, mientras que en los 5 restantes fue el Grupo 2 el que tuvo medias mayores. Dichas medias se calcularon asignando un valor numérico entre 1 y 5 a las respuestas posibles de la siguiente forma: “Totalmente en desacuerdo”, 1; “En desacuerdo”, 2; “Ni en desacuerdo ni de acuerdo”, 3; “De acuerdo”, 4 y “Totalmente de acuerdo”, 5.

Tabla 12: Preguntas adicionales con medias superiores para el Grupo 1

Pregunta	Grupo 1 (N = 20)				Grupo 2 (N = 24)			
	Media	Desviación estándar	Intervalo de confianza		Media	Desviación estándar	Intervalo de confianza	
La forma de interactuar con el Asistente es clara desde la primera pregunta	4.40	0.68	4.10	4.70	4.04	0.95	3.66	4.42
Las preguntas evalúan los mismos conceptos más de una vez	4.40	0.75	4.07	4.73	3.04	1.08	2.61	3.47
Todos los conceptos incluidos en las preguntas son sencillos de entender	4.10	0.97	3.68	4.52	3.17	1.31	2.64	3.69
El Asistente de Voz hablaba más de lo necesario	2.85	1.27	2.29	3.41	2.71	1.16	2.24	3.17
Era común que, a la hora de responder, se me hubiera olvidado qué era lo que se me había preguntado	2.95	1.32	2.37	3.53	2.13	1.19	1.65	2.60
Si me equivoqué al dar una respuesta, es posible corregirla en el momento	2.70	1.49	2.05	3.35	2.17	1.24	1.67	2.66

En la Tabla 12 pueden observarse las preguntas en las que la opinión del Grupo 1 fue más favorable que la del Grupo 2, aclarando que al Grupo 2 fue el grupo que tuvo una explicación detallada de las diferentes capacidades del agente y la manera de utilizarlas.

El hecho de haber recibido esta explicación puede haber influido en la manera de responder dichas preguntas, tal y como puede apreciarse en la Figura 18.

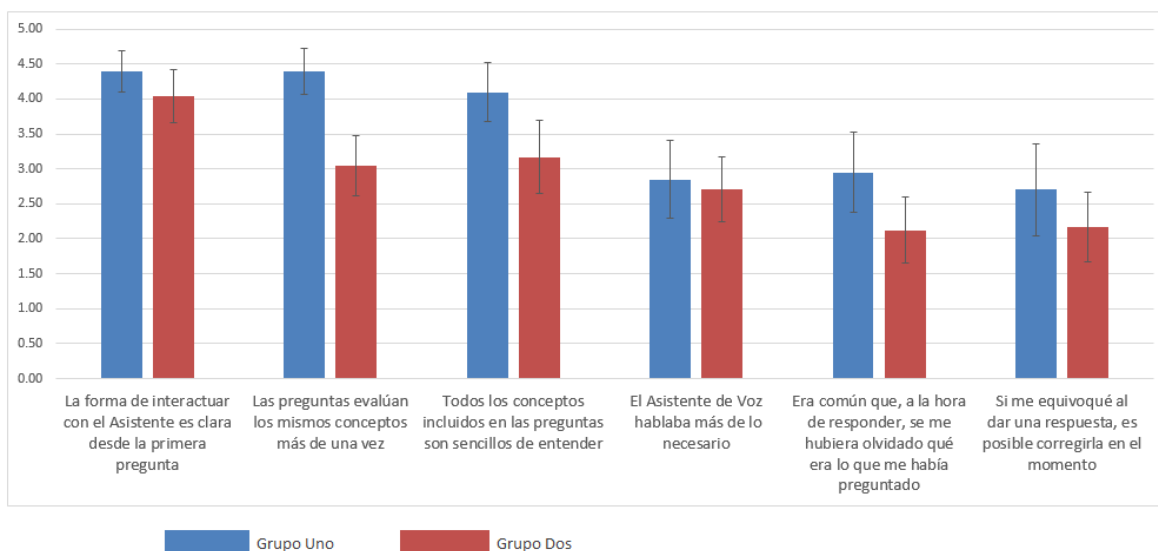


Figura 18: Comparación de las respuestas adicionales con medias superiores para el Grupo 1 sobre el Grupo 2

Por ejemplo, ante la afirmación “todos los conceptos incluidos en las preguntas son sencillos de entender”, el hecho de implementar la capacidad de proporcionar definiciones o explicaciones de los conceptos pudo sugerir a los participantes que algunos de ellos no eran de fácil comprensión. De la misma forma, el conocer las diferentes capacidades del asistente pudo afectar la percepción de la afirmación “La forma de interactuar con el Asistente es clara desde la primera vez”.

Tabla 13: Preguntas adicionales con medias superiores en el Grupo 2

Pregunta	Grupo 1 (N = 20)				Grupo 2 (N = 24)			
	Media	Desviación estándar	Intervalo de confianza		Media	Desviación estándar	Intervalo de confianza	
La extensión del cuestionario realizado es adecuada	2.75	1.21	2.22	3.28	2.83	0.96	2.45	3.22
Se entiende fácilmente cuántas preguntas se han hecho y cuántas faltan para terminar	3.25	1.21	2.72	3.78	3.67	1.09	3.23	4.10
Es sencillo saber si el Asistente de Voz entendió mi respuesta	4.10	1.12	3.61	4.59	4.21	0.78	3.90	4.52
Es sencillo sabe si el Asistente de Voz entendió la pregunta que le hice	3.15	1.35	2.56	3.74	3.46	1.18	2.99	3.93
Es sencillo solicitar ayuda al Asistente de Voz	2.30	1.13	1.81	2.79	2.42	1.02	2.01	2.82

La Tabla 13 muestra las preguntas en las que las medias de las respuestas del Grupo 2 son mayores a las del Grupo 1, destacando que en este grupo de preguntas las

medias son bastante similares. La Figura 19 muestra gráficamente la información representada en la Tabla 13.

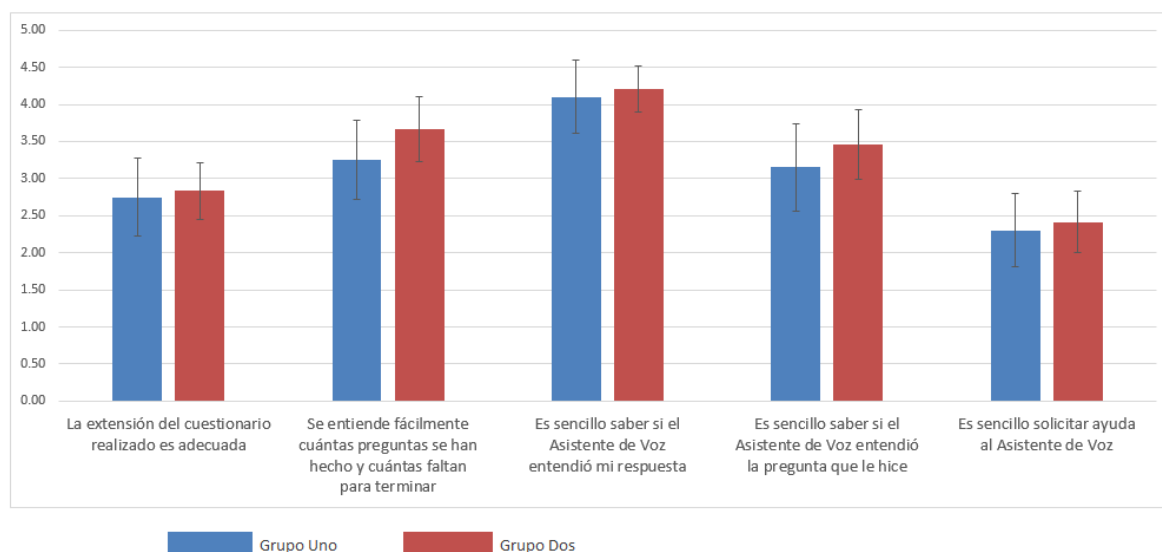


Figura 19: Comparación de las respuestas adicionales con medias superiores para el Grupo 2 sobre el Grupo 1

Resulta interesante destacar que la afirmación “Se entiende fácilmente cuántas preguntas se han hecho y cuántas faltan para terminar” se podría haber visto afectada positivamente por la explicación dada al Grupo 2, en la cual se les indicó a los participantes que podían solicitar dicha información al Asistente en cualquier momento.

Cabe resaltar que, en promedio, los participantes duraron 10 minutos con 25 segundos en completar los cuestionarios estandarizados utilizando el Asistente por Voz.

Al hacer una Prueba T para determinar si las diferencias que se encontraron en los resultados de la evaluación de UEQ+ son significativas o no, se tiene que tomando alfa como 0.05 ninguna de las escalas presenta una diferencia significativa, tal y como se muestra en la Tabla 14.

Del mismo modo, al hacer una Prueba T para determinar si las diferencias que se encontraron en las preguntas de usabilidad son significativas o no, se tiene que tomando alfa como 0.05 los ítems de “Las preguntas evalúan los mismos conceptos más de una vez”, “Todos los conceptos incluidos en las preguntas son sencillos de

entender” y “Era común que, a la hora de responder, se me hubiera olvidado qué era lo que me había preguntado” presentaron una diferencia significativa, mientras que el resto de ítems no, tal y como se muestra en la Tabla 15.

Tabla 14: Resultados Pruebas T en resultados de UEQ+

Escala	Valor T	Tipo de diferencia
Comprensibilidad	0.7179	Sin diferencia significativa
Comportamiento de respuesta	0.1253	Sin diferencia significativa
Uso intuitivo	0.3918	Sin diferencia significativa
Calidad de respuesta	0.6356	Sin diferencia significativa
Novedad	0.1458	Sin diferencia significativa

Tabla 15: Resultados Pruebas T en resultados de preguntas de usabilidad

Items	Valor T	Tipo de diferencia
La forma de interactuar con el Asistente es clara desde la primera pregunta	0,1548	Sin diferencia significativa
La extensión del cuestionario realizado es adecuada	0,8045	Sin diferencia significativa
Las preguntas evalúan los mismos conceptos más de una vez	0,0000	Diferencia significativa
Se entiende fácilmente cuántas preguntas se han hecho y cuántas faltan para terminar	0,2412	Sin diferencia significativa
Todos los conceptos incluidos en las preguntas son sencillos de entender	0,0096	Diferencia significativa
Es sencillo saber si el Asistente de Voz entendió mi respuesta	0,7172	Sin diferencia significativa
Es sencillo saber si el Asistente de Voz entendió la pregunta que le hice	0,4291	Sin diferencia significativa
El Asistente de Voz hablaba más de lo necesario	0,7034	Sin diferencia significativa
Era común que, a la hora de responder, se me hubiera olvidado qué era lo que me había preguntado	0,0370	Diferencia significativa
Si me equivoqué al dar una respuesta, es posible corregirla en el momento	0,2104	Sin diferencia significativa
Es sencillo solicitar ayuda al Asistente de Voz	0,7231	Sin diferencia significativa

Capítulo 7

Conclusiones y trabajo futuro

Al término de este trabajo de investigación y luego de la participación de 56 personas (12 en la fase de diseño y 44 en la implementación del caso de estudio) se puede llegar a la conclusión de que es posible implementar un instrumento de evaluación de experiencia de usuario que utilice una interfaz no tradicional como lo es una interfaz conversacional por voz. En el presente trabajo se utilizó como base un cuestionario estandarizado de evaluación, concretamente el UEQ, y se implementó el mecanismo de captura de respuestas a través de una interfaz conversacional por voz. Los resultados obtenidos son coherentes con lo que se espera de una evaluación mediante un cuestionario estandarizado, lo que quedó reflejado en la cantidad baja de inconsistencias presentadas en las respuestas de los participantes.

Al evaluar los resultados de la aplicación del caso de estudio, tanto la evaluación UEQ+ así como las preguntas adicionales, se comienza a notar un patrón de que la evaluación del agente como tal mejora significativamente cuando los participantes tienen una explicación amplia de las diferentes capacidades y funciones del agente. En ambos grupos del caso de estudio, el aspecto de “uso intuitivo” fue el mejor evaluado, sin embargo, el resto de los aspectos evaluados fueron significativamente mejores en el Grupo 2, grupo en el cual se dio una explicación de las capacidades del agente.

El contar con un mecanismo alternativo de aplicación de un instrumento de evaluación de experiencia de usuario resulta útil a los investigadores que utilizan instrumentos de distribución masiva y en los que los participantes auto aplican directamente el instrumento de evaluación. Lo anterior se debe a la facilidad de aplicación, ya que no se necesitaría más que una conexión a internet y un dispositivo (teléfono celular o asistente personal) para utilizarlo. Además, brinda una plataforma más inclusiva principalmente para personas que cuentan con algún tipo de impedimento ya sea motriz o visual. Claramente, si el participante tiene algún problema para expresarse verbalmente, una interfaz de voz no sería útil, pero el objetivo es contar con más instrumentos que puedan utilizarse en diferentes situaciones o contextos.

El hecho de utilizar una interfaz no tradicional brinda cierto sentido de ubicuidad, ya que las evaluaciones se podrían realizar desde diferentes lugares y contextos, situación que no ocurre con los medios tradicionales, donde se está limitado a un medio y un lugar físico determinado. Sin embargo, quedan ciertos aspectos por mejorar en términos de la implementación como tal, basado en la retroalimentación dada por los participantes, como son el tipo de voz utilizado por el asistente y el reconocimiento de voz de la captura de las respuestas, que en este caso estuvo limitado a las capacidades de la plataforma Voiceflow, que podría verse beneficiado al exportar el Asistente a un parlante inteligente que cuente con mejores capacidades de reconocimiento de voz.

Otro elemento que puede trabajarse en mayor grado es lograr que la interacción no sea ejecutada tan mecánicamente como si fuera un guion, sino procurando una conversación más natural, en el entendido de que el participante entiende que está interactuando con un Asistente y no con otro ser humano. Aunque fueron implementados los elementos identificados como básicos para una conversación natural con el Asistente de Voz, y aquellos necesarios a nivel de navegación, a los participantes les fue difícil encontrar algunas de estas características o suponer que estaban implementadas en el asistente.

Como parte del trabajo futuro, también se contempla realizar un caso de estudio en el que los resultados obtenidos de la evaluación mediante UEQ, implementado con una interfaz conversacional por voz, sean comparados con una evaluación en el que el mecanismo de captura se realiza mediante una interfaz gráfica tradicional, por ejemplo, un cuestionario escrito distribuido digitalmente.

Finalmente, se deben buscar formas de que el participante entienda qué capacidades tiene el Asistente de Voz o que se atreva a explorar más activamente los alcances de la conversación, por ejemplo, implementar una ayuda que en todo momento pueda ser activada por el participante por ejemplo mediante las preguntas ¿qué puedes hacer?, ¿de qué eres capaz? o similares. Esto, sin embargo, también requiere que los usuarios estén más acostumbrados a que las conversaciones por voz, no solamente a las interfaces gráficas, y que también entiendan que las interfaces conversacionales por voz son cada vez más complejas y con mayores capacidades, y que no solamente sirven para comandos o preguntas simples.

Referencias

- [1] Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014, June). Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In *International Conference of Design, User Experience, and Usability* (pp. 383-392). Springer, Cham.
- [2] Vermeeren, A. P. O. S., Lai-Chong Law, E., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User Experience Evaluation Methods: Current State and Development Needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries: NordiCHI '10* (pp. 521-530).
- [3] Wallach, D., Conrad, J., & Steimle, T. (2017, July). The UX metrics table: A missing artifact. In *International Conference of Design, User Experience, and Usability* (pp. 507-517). Springer, Cham.
- [4] Lallemand, C., & Gronier, G. (2018). *Méthodes de design UX: 30 méthodes fondamentales pour concevoir des expériences optimales*. (2 ed.) Ed. Eyrolles.
- [5] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- [6] Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological bulletin*, 49(3), 197.
- [7] Klammer, J., & van den Anker, F. W. (2018, July). A Platform to Connect Swiss Consumers of Fair Trade Products with Producers in Developing Countries: Needs and Motivations. In *International Conference of Design, User Experience, and Usability* (pp. 664-681). Springer, Cham.
- [8] Maia, C. L. B., & Furtado, E. S. (2016, July). A systematic review about user experience evaluation. In *International Conference of Design, User Experience, and Usability* (pp. 445-455). Springer, Cham.
- [9] Díaz-Oreiro, I., López, G., Quesada, L., & Guerrero, L. A. (2019). Standardized questionnaires for user experience evaluation: A systematic literature review. In *Multidisciplinary Digital Publishing Institute Proceedings* (Vol. 31, No. 1, p. 14).

- [10] (2019, octubre) Global Voice Assistant Market, By Technology (Speech Recognition, Text to Speech Recognition, Voice Recognition & Natural Language Processing), By Application, By End User, By Region, Competition, Forecast & Opportunities, 2024. Tomado de: <https://www.businesswire.com/news/home/20190916005535/en/Global-Voice-Assistant-Market-Projected-Grow-1.2>
- [11] van Beurden, M. H., Ijsselsteijn, W. A., & de Kort, Y. A. (2011, May). 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.
- [12] Pradhan, A., Mehta, K., & Findlater, L. (2018, April). " Accessibility Came by Accident" Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on human factors in computing systems* (pp. 1-13).
- [13] Borgia, E. (2014). The Internet of Things vision: Key features, applications and open issues. *Computer Communications*, 54, 1-31.
- [14] Perez, S. (2019). Report: Voice assistants in use to triple to 8 billion by 2023. Tomado de: <https://techcrunch.com/2019/02/12/report-voice-assistants-in-use-to-triple-to-8-billion-by-2023/>
- [15] López, G., Quesada, L., & Guerrero, L. A. (2017, July). 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.
- [16] Moore, R. J., & Arar, R. (2019). *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. Morgan & Claypool.
- [17] Lallemand, C., & Koenig, V. (2017, September). How Could an Intranet be Like a Friend to Me? Why Standardized UX Scales Don't Always Fit. In *Proceedings of the European Conference on Cognitive Ergonomics 2017* (pp. 9-16).
- [18] Baumgartner, J., Sonderegger, A., & Sauer, J. (2019). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Studies*, 122, 78-89.

- [19] Forster, Y., Hergeth, S., Naujoks, F., & Krems, J. F. (2018, September). How usability can save the day-methodological considerations for making automated driving a success story. In Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications (pp. 278-290).
- [20] Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In Mensch & computer 2003 (pp. 187-196). Vieweg+ Teubner Verlag.
- [21] Laugwitz, B., Held, T., & Schrepp, M. (2008, November). Construction and evaluation of a user experience questionnaire. In Symposium of the Austrian HCI and usability engineering group (pp. 63-76). Springer, Berlin, Heidelberg.
- [22] Minge, M., & Riedel, L. (2013). meCUE–Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. In Mensch & Computer 2013–Tagungsband (pp. 89-98). Oldenbourg Wissenschaftsverlag.
- [23] Chien, Y. H., Chang, W. T., Chuang, C. C., & Chen, S. H. (2017). A Taiwanese user experience questionnaire. In Advances in Ergonomics Modeling, Usability & Special Populations (pp. 347-355). Springer, Cham.
- [24] Vallerand, R. J. (1989). Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue française. Canadian Psychology/Psychologie Canadienne, 30(4), 662.
- [25] Bernhaupt, R., & Pirker, M. (2013, September). Evaluating user experience for interactive television: towards the development of a domain-specific user experience questionnaire. In IFIP Conference on Human-Computer Interaction (pp. 642-659). Springer, Berlin, Heidelberg.
- [26] Tomlinson, B. J., Noah, B. E., & Walker, B. N. (2018, April). Buzz: An auditory interface user experience scale. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-6).
- [27] Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. Medical reference services quarterly, 37(1), 81-88.
- [28] Yang, X., Aurisicchio, M., & Baxter, W. (2019, May). Understanding affective experiences with conversational agents. In proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-12).

- [29] Berdasco, A., López, G., Diaz, I., Quesada, L., & Guerrero, L. A. (2019). User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana. In *Multidisciplinary Digital Publishing Institute Proceedings* (Vol. 31, No. 1, p. 51).
- [30] Amazon Inc. Tomado de: <https://www.amazon.com/Voice-Surveys-SurveyLine/dp/B07LFXV7DF>
- [31] SurveyLine. Tomado de: <https://www.surveybyvoice.com>
- [32] Pearl, C. (2016). *Designing voice user interfaces: Principles of conversational experiences*. " O'Reilly Media, Inc."
- [33] Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014, June). Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In *International Conference of Design, User Experience, and Usability* (pp. 383-392). Springer, Cham.
- [34] Roto, V., Obrist, M., & Väänänen-Vainio-Mattila, K. (2009, August). User experience evaluation methods in academic and industrial contexts. In *Proceedings of the Workshop UXEM* (Vol. 9, pp. 1-5).
- [35] Gronier, G., Lallemand, C., & Chauvet, A. (2015). Mesurer la formation de la première impression d'une interface à l'aide du test des 5 secondes.
- [36] Chaudhri, V. K., Cheyer, A., Guilli, R., Jarrold, W., Myers, K. L., & Niekarsz, J. (2006, November). A Case Study in Engineering a Knowledge Base for an Intelligent Personal Assistant. In *SemDesk*.
- [37] Kaushik, D., & Jain, R. (2014). Natural user interfaces: Trend in virtual interaction. *arXiv preprint arXiv:1405.0101*.
- [38] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE), 2493-2537.
- [39] Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: a modern approach*. 3rd ed. Upper Saddle River, NJ. Prentice Hall
- [40] Schrepp, M., & Thomaschewski, J. (2019). Design and Validation of a Framework for the Creation of User Experience Questionnaires. *International Journal of Interactive Multimedia & Artificial Intelligence*, 5(7).

- [41] Klein, A. M., Hinderks, A., Schrepp, M., & Thomaschewski, J. (2020, September). Construction of UEQ+ scales for voice quality: measuring user experience quality of voice interaction. In Proceedings of the Conference on Mensch und Computer (pp. 1-5).
- [42] Dahlbck, N., Jdnsson, A., & Ahrenberg, L. (1993). Wizard of oz studies—why and how. In Proceedings of the 1993 International Workshop on Intelligent User Interfaces: January (pp. 4-7).

Anexo 1.

Cuerpo de expresiones obtenidas de actividad de lluvia de ideas

Formas de solicitar que repitan la pregunta:

- Decir la frase "repetir pregunta"
- Decir la frase "repetir pregunta más despacio"
- Asistente, repetir
- ¿Cómo?
- ¿Cómo fue?
- ¿Qué dijo?
- Repetir pregunta
- ¿Me lo puede repetir?
- ¿Cuál fue la pregunta?
- ¿Qué?
- De nuevo
- Decir una palabra clave como "repeat"
- Repetir toda la oración
- No entendí, ¿me podría repetir?
- ¿Puede decirlo de nuevo?
- Necesito que me repita eso último
- Repita lo que acaba de decir
- No escuché
- Decir "Repetir"
- Decir "No entendí"
- Decir "De nuevo"
- Repetir enunciado/pregunta
- Repetir pregunta #X
- Preguntar de nuevo
- Repetir la última pregunta
- Volver a preguntar
- ¿Me puede repetir la pregunta?
- ¿Podría repetir lo que acaba de decir?
- No entendí lo que dijo, ¿será que lo puede volver a decir?
- Estoy perdido, ¿puede decirlo nuevamente?
- De nuevo
- ¿Cuál es la pregunta?
- ¿Que acaba de decir?
- ¿Cuál fue la última pregunta?
- ¿Me podría repetir la última oración?
- ¿Me podría repetir la última pregunta?
- ¿Qué fue lo que dijo?
- ¿Cual fue la pregunta?
- ¿Podría repetir, por favor?
- No recuerdo qué me preguntó
- ¿Otra vez?

- Me distraje, ¿qué me preguntó?
- ¿Qué me preguntó?
- ¿Me repite la pregunta?
- ¿Me repite, por favor?
- ¿Cómo era?
- ¿Cuáles eran las palabras?
- ¿Cuáles eran las ideas?
- ¿Cuáles eran los conceptos?
- Repita la pregunta
- ¿Me podría por favor repetir la pregunta?
- Por favor, repita la pregunta
- No entendí, de nuevo por favor
- ¿Cómo era?

Formas de solicitar explicar un significado:

- Describa palabra
- Sinónimo palabra
- Traduzca a X idioma
- Traduzca palabra a X idioma
- ¿Podría explicarme palabra?
- ¿Qué quiere decir?
- ¿Cuál es el significado?
- Defina
- Ejemplificar palabra
- ¿Qué significa <palabra>?
- ¿Cuál es sinónimo de <palabra>?
- ¿Puede usar <palabra> en una frase?
- Describa <palabra>
- No entiendo el significado, ¿me lo puede repetir/explicar?
- ¿Qué quiso decir con <palabra>?
- Traduzca <palabra> al idioma <idioma>
- Describir conceptos

Formas de solicitar el tiempo transcurrido y la cantidad de preguntas faltantes

- ¿Quiero terminar?
- ¿Cuántas preguntas faltan?
- ¿Cuál es el número de esta pregunta?
- ¿Cuánto he completado del cuestionario?
- ¿Cuántos ítems me faltan?
- ¿Ya terminé?
- ¿Cuánto falta?
- ¿Cuántas preguntas faltan?
- ¿Cuál es el número de esta pregunta?
- ¿Cuánto tiempo queda?
- ¿Cuál es el tiempo restante?
- ¿Falta mucho?

- ¿Número de pregunta sobre el total?
- ¿Tiempo aproximado para terminar?
- Preguntas restantes
- ¿Cuántas preguntas llevo del total?
- ¿Llevo más de la mitad?
- ¿Ya casi se acaba?
- Tiempo
- ¿Cuánto tiempo queda?
- ¿Cuántas preguntas son?
- ¿Falta mucho para terminar?
- ¿Cuántas preguntas quedan?
- ¿Qué hora es?
- ¿Cuánto tiempo llevamos?
- ¿Cuánto tiempo hemos invertido en contestar el cuestionario?
- ¿Cuántas preguntas tiene este cuestionario?
- ¿Cuántos ítems tiene este cuestionario?
- ¿Cuántas preguntas quedan pendientes?
- ¿Cuántas preguntas quedan sin responder?
- ¿Cuál es el progreso?
- ¿Cuál es el avance?
- ¿Cuánto falta para terminar?
- ¿Cuántas preguntas quedan?
- ¿Cuánto tiempo llevamos hasta aquí?
- ¿En qué número de pregunta estoy?