UNIVERSIDAD DE COSTA RICA
SISTEMADE ESTUDIOS DE POSGRADO

CONSTRUCCIÓN DE UN MECANISMO DE CAPTURA POR VOZ PARA UN INSTRUMENTO DE EVALUACIÓN DE EXPERIENCIA DE USUARIO

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

JEAN CARLO MATA SERRANO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2022

Dedicatoria

A mis padres quienes los admiro y siempre me ha enseñado excelentes valores, como el trabajo honrado y humildad.

A mis hermanas, por su apoyo incondicional.

A mis abuelos, quienes me han enseñado que la unión y el apoyo familiar es lo principal.

A mi esposa, por creer en mí, por todo su amor y apoyo incondicional diario.

Agradecimientos

A mi director de TFIA, Ignacio Díaz Oreiro, por su guía, apoyo y conocimiento compartido durante todo el proceso de la investigación. Al grupo USING por los consejos y retroalimentación brindada en cada una de las sesiones. A Gabriela Marín Raventós por el apoyo académico en el posgrado y poder participar en la conferencia ICITS 2022 y finalmente a la UCR por haberme ayudado a crecer personal y profesionalmente durante mi participación en el posgrado.

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

LUIS JOSE QUESADA QUIROS (FIRMA)
Digitally signed by LUIS JOSE QUESADA QUIROS (FIRMA)
Date: 2022.09.26 22:39:52 -06'00'

Dr. Luis Quesada Quirós

Representante de la Decana Sistema de Estudios de Posgrado

IGNACIO DIAZ OREIRO (FIRMA)
Firmado digitalmente por IGNACIO DIAZ OREIRO (FIRMA)
Fecha: 2022.09.23 14:12:05 -06'00'

M.Sc. Ignacio Díaz Oreiro

Profesor Guía

GUSTAVO LOPEZ HERRERA (FIRMA)
Firmado digitalmente por GUSTAVO LOPEZ HERRERA (FIRMA)
Fecha: 2022.09.24 11:11:43 -06'00'

Dr. Gustavo López Herrera

Lector

LUIS ALBERTO GUERRERO BLANCO (FIRMA)
PERSONA FISICA, CPF-02-0418-0223.
Fecha declarada: 03/10/2022 08:11:47 AM
Esta representación visual no es fuente de confianza. Valide siempre la firma.

Dr. Luis A. Guerrero Blanco

Lector

GABRIELA MARIN RAVENTOS (FIRMA)
Firmado digitalmente por GABRIELA MARIN RAVENTOS (FIRMA)
Fecha: 2022.09.28 16:45:45 -06'00'

Dra. Gabriela Marín Raventós

Directora

Programa de Posgrado en Computación e Informática

JEAN CARLO MATA SERRANO (FIRMA)
PERSONA FISICA, CPF-01-1613-0895.
Fecha declarada: 27/09/2022 08:05:55 AM
Esta representación visual no es fuente de confianza. Valide siempre la firma.

Jean Carlo Mata Serrano

Sustentante

Tabla de Contenidos

Resumen

Los cuestionarios estandarizados, instrumentos ampliamente utilizados para evaluar la experiencia de usuario, cuentan con un mecanismo de captura respuestas implementado en forma escrita, ya sea en papel o en formato digital. Este trabajo propone utilizar interfaces de voz como mecanismo de recolección de las respuestas, para lo que se realizaron dos implementaciones del cuestionario UEQ (User Experience Questionnaire) en las que se varió el formato de las preguntas y la cantidad de patrones conversacionales incluidos. Estas implementaciones se probaron mediante dos casos de estudio con un total de 40 participantes en cada caso.

Los resultados de ambos casos de estudio muestran que las evaluaciones de experiencia de usuario obtenidas con el mecanismo por voz no presentan diferencias significativas de los resultados identificados con la implementación tradicional escrita del cuestionario, lo que permitiría contar con una alternativa para recolectar las respuestas de los usuarios, conservando las ventajas aportadas por los cuestionarios estandarizados y agregando la facilidad de uso y el auge de la creciente adopción de interfaces de voz conversacionales.

Adicionalmente, la usabilidad y experiencia de usuario de utilizar el mecanismo de captura implementado por voz varía en función del formato de la pregunta y de los patrones conversacionales diseñados, identificándose incluso diferencias en la cantidad de inconsistencias en las respuestas recolectadas.

Lista de Figuras

Lista de Tablas

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, Jean Carlo Mata Serrano, con cédula de identidad 1 1613 0895, en mi condición de autor del TFG titulado Construcción de un mecanismo de captura por voz para un instrumento de evaluación de experiencia de usuario.

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

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

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

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

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Jean Carlo Mata Serrano.

Número de Carné: B89880 Número de cédula: 1 1613 0895.

Correo Electrónico: jean.cms07@gmail.com.

Fecha: 15/09/2022. Número de teléfono: 8428-8898.

Nombre del Director (a) de Tesis o Tutor (a): 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 evaluación de la experiencia del usuario (UX) comprende un conjunto de métodos, habilidades y herramientas que se utilizan para descubrir las percepciones y respuestas de una persona que resultan del uso posterior y/o uso anticipado de un producto, sistema o servicio, según la definición de UX presentada por ISO [1]. Esta definición de ISO incluye las emociones, creencias, respuestas físicas y psicológicas de los usuarios y considera UX también una consecuencia de la imagen de marca, la presentación, el rendimiento del sistema, el estado interno y físico del usuario como resultado de experiencias previas, actitudes, habilidades y personalidad, entre otros.

Para evaluar la UX, los investigadores se basan en diferentes métodos e instrumentos, como la evaluación de expertos, estudios etnográficos, cuestionarios de diseño específicos para el caso de estudio, cuestionarios estandarizados, entrevistas, por nombrar algunos de los más aplicados [2-5]. El uso de cuestionarios estandarizados está muy extendido [6-7], por diferentes motivos entre los que podemos mencionar la capacidad de proporcionar una puntuación cuantitativa asociada a la evaluación que, a su vez, puede compararse con *benchmarks* o con la puntuación de otras evaluaciones realizadas. Por ejemplo, el cuestionario UEQ ("*User Experience Questionnaire*"), uno de los dos cuestionarios estandarizados de evaluación UX más utilizados en el mundo, compara los resultados obtenidos con la información de una base de datos de 452 estudios y 20,190 participantes. Adicionalmente, se necesitan pocos recursos para la administración y la recolección de los datos de estos cuestionarios, ya que no se debe invertir ningún esfuerzo en diseñarlos, dado que el conjunto de preguntas es invariable y su aplicación es sencilla, ya que la experiencia es reportada por los propios usuarios [5]. Finalmente, los cuestionarios están validados estadísticamente, por lo que se consideran fiables para medir la UX [6], [8].

A pesar de las ventajas y el uso generalizado, los cuestionarios estandarizados de evaluación de UX pueden resultar un poco aburridos o tediosos de completar, ya que todas las preguntas siguen el mismo patrón (diferencial semántico o escala Likert) y cada cuestionario está compuesto por un número considerable de preguntas, lo que podría incidir en que la experiencia de llenarlos no sea tan agradable o que eventualmente los participantes

disminuyan su interés en el proceso y brinden respuestas menos precisas o estén menos dispuestos a participar en una evaluación que utiliza estos instrumentos.

Los inconvenientes señalados anteriormente nos llevan a plantear la interrogante de si existen otras formas de capturar las respuestas de un cuestionario estandarizado, de forma que el proceso de llenarlos no se realice de manera escrita por el participante, sino utilizando alguna otra interfaz disponible que permita variar la experiencia de llenarlos, y si el utilizar otros mecanismos de captura afectaría los resultados de la evaluación de UX. En este sentido, en las últimas décadas se han desarrollado diversas interfaces no tradicionales que buscan explotar distintas formas de interacción entre personas y computadoras entre las que podemos mencionar: detección y seguimiento de la mirada o gestos, interfaces hápticas y tangibles, interfaces por voz e interfaces conversacionales [9-10]. De estas interfaces no tradicionales nos interesa resaltar las interfaces de voz conversacionales, que utilizan el lenguaje oral para interactuar con un sistema o tecnología, a través de intercambios que se asemejan a una conversación entre dos personas. Las interfaces de voz se han implementado en diferentes entornos, con gran éxito en algunos de estos entornos [11], debido tanto a la facilidad de uso como a las continuas mejoras que experimentan en el reconocimiento de patrones de voz o la expansión a diferentes idiomas además del inglés [13].

Ahora bien, el utilizar interfaces conversacionales de voz como medio para implementar el mecanismo de captura de un cuestionario estandarizado de experiencia de usuario nos plantea la siguiente pregunta de investigación:

¿Qué nivel de afectación podría darse en los resultados de la evaluación de un instrumento de evaluación UX, si la implementación del mecanismo de captura se realiza mediante una interfaz conversacional de voz en lugar de una implementación tradicional escrita?

Justificación

Como se mencionó anteriormente, los cuestionarios estandarizados de evaluación UX tienen características que hacen que sean muy utilizados, pero cuentan con algunos elementos que pueden representar inconvenientes en la experiencia de los participantes y eventualmente en la calidad de los resultados. Aunado a esto, los cuestionarios en formato de captura escrito podrían presentar limitantes para participantes con diferentes capacidades visuales o motoras,

por lo que sería útil contar con medios alternos de captura de información para estos cuestionarios. Surge entonces el interés de explorar el uso de interfaces conversacionales de voz como un medio para implementar el mecanismo de captura de un cuestionario estandarizado de evaluación de UX con el fin de aprovechar los beneficios presentes en los cuestionarios estandarizados y, al mismo tiempo, buscar alternativas de interacción para los usuarios, como por ejemplo mediante el uso del lenguaje hablado. Además, el hecho de que el uso de interfaces de voz está cada vez más extendido, la evaluación de UX de un producto o servicio podría ser realizada por los propios usuarios, siendo así una experiencia auto reportada que es una de las ventajas mencionadas anteriormente, que proporcionan los cuestionarios estandarizados. El objetivo es aprovechar el respaldo estadístico de los cuestionarios estandarizados y las ventajas que brindan en términos de diseño y aplicación en la evaluación UX y a su vez explotar las características que brindan las interfaces de voz en términos de facilidades de comunicación oral.

Es importante mencionar que se han realizado esfuerzos para crear otros instrumentos estandarizados de evaluación que modifican el formato de presentación de las preguntas para tratar de evitar los posibles inconvenientes de estos cuestionarios, entre los que podemos citar el número alto de preguntas o el hecho de que el participante deba leer un conjunto de conceptos de los que quizás no comprenda completamente. Las alternativas presentadas por los autores de estos esfuerzos se enfocan principalmente en formatos visuales de gestos faciales o corporales para representar emociones o actitudes con las que el participante asocia el producto o sistema [14-15]. Sin embargo, a nuestro buen entender, no hay propuestas para implementar cuestionarios estandarizados de evaluación UX a través de interfaces de voz. Esta conclusión se obtuvo como parte de una revisión sistemática de la literatura [16] desarrollada por los investigadores sobre el uso de cuestionarios estandarizados en estudios primarios, en la que en un primer filtrado inicial se identificaron variantes a los cuestionarios u otros mecanismos de evaluación.

Una implementación conversacional por voz del mecanismo de captura de información sería muy útil dado que se podría contar con un mecanismo que resulte más natural para las personas que el mecanismo actual, donde se llena el cuestionario en forma escrita, lo que podría ser percibido como una mejora en la experiencia de usuario al utilizar estos

instrumentos de evaluación. Cuando decimos “natural”, en el contexto de interfaces por voz, nos referimos a una interacción, entre la persona y la interfaz tecnológica, tal y como la definen Moore y Arar en su artículo *Natural Conversation Framework*, donde se presentan diferentes patrones conversacionales que muestran la organización y reparación de secuencias que son formas simplificadas de patrones naturales de una conversación entre personas [17]. Estos patrones conversacionales, así como la definición de una conversación “natural”, se describen en detalle en el capítulo 2 o Marco Conceptual, y servirían para establecer una interacción por voz con el mecanismo de captura del cuestionario estandarizado que emule una conversación, en lugar de la interfaz escrita usual que poseen los cuestionarios.

Luego de implementada la interfaz por voz como mecanismos de captura de respuestas del cuestionario de evaluación UX, se continuaría aprovechando la base estadística y la validación académica que proveen estos cuestionarios. Si los resultados de las evaluaciones del nuevo mecanismo no presentan diferencias significativas respecto del tradicional escrito, esta implementación por voz proporcionaría una alternativa a los instrumentos actuales respecto de la recolección de las respuestas. Se puede concluir, entonces, que la presente investigación generará conocimiento científico aplicable a la evaluación de la experiencia de usuario utilizando cuestionarios estandarizados, del que se beneficiarán los investigadores del área, que tendrán a su disposición variantes en el mecanismo de recolección de datos, aplicables a un instrumento muy utilizado y considerado confiable [6-8].

Cabe señalar que la motivación para realizar este proyecto surge del trabajo realizado en el curso “Interacción Humano Computador” del programa de Maestría en Computación e Informática de la Universidad de Costa Rica, en el cual se abordaron los temas: experiencia de usuario (UX), evaluación de UX e implementación de sistemas por medio de interfaces no tradicionales. Particularmente, en el Laboratorio de este curso, se estudió el tema de implementar el mecanismo de captura del cuestionario estandarizado UEQ (User Experience Questionnaire) mediante una interfaz conversacional por voz, por lo que este Trabajo Final de Investigación Aplicada sería una continuación del esfuerzo iniciado en el curso mencionado.

Para realizar esta investigación centrada en implementar un mecanismo de captura por voz en un cuestionario estandarizado de evaluación UX se planteó el objetivo general, así como los objetivos específicos asociados que se presentan a continuación.

Objetivo General

Determinar el efecto de un mecanismo de captura por voz en los resultados de una evaluación de UX usando el cuestionario estandarizado UEQ.

Objetivos Específicos

a. Comparar los resultados de una evaluación de UX, usando como mecanismo de captura un cuestionario estandarizado escrito y una implementación que utilice una interfaz por voz.
b. Incorporar patrones conversacionales en una interfaz por voz para la captura de respuestas de un cuestionario estandarizado.
c. Comparar los resultados de una evaluación de UX usando como mecanismos de captura un cuestionario estandarizado escrito y una implementación por voz que incluya patrones conversacionales.
d. Contrastar la experiencia de usuario de los mecanismos de captura por voz implementados.

Teniendo claros los objetivos definidos el resto del presente documento comprende los siguientes capítulos: 1. Introducción, 2. Marco Conceptual, 3. Trabajo Relacionado, 4. Metodología, 5. Resultados y finalmente 6. Conclusiones y Trabajo futuro.

En el capítulo 2. Marco Conceptual, se definen los conceptos fundamentales necesarios para entender el presente documento, mientras que en el capítulo 3. Trabajo Relacionado, se presentan referencias a trabajos asociados a la presente investigación, El capítulo 4. Metodología, muestra las actividades ejecutadas para cumplir cada uno de los objetivos planteados, en el capítulo 5. Resultados, se evidencian los resultados obtenidos de la

investigación para cada uno de los casos de estudio ejecutados. En el capítulo 6 encontraremos las conclusiones de la investigación y al final del documento se presentan los capítulos 7. Referencias bibliográficas utilizadas y 8. Anexos.

Capítulo 2 - Marco conceptual

En este capítulo se detalla un conjunto de conceptos utilizados a lo largo de este Trabajo Final de Investigación Aplicada. En primer lugar, se presentan aquellos conceptos relacionados con la evaluación de la experiencia de usuario y los cuestionarios estandarizados que se utilizan para realizar dicha evaluación y luego se describen los conceptos relacionados con interfaces de voz y patrones conversacionales de voz.

Experiencia de Usuario

Uno de los temas más importantes de la investigación, la experiencia de usuario (UX) es hoy en día un elemento clave para determinar la calidad de un producto o servicio [4], [35]. La experiencia de usuario abarca no solo conceptos llamados pragmáticos de un producto o sistema, 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 hedónica, 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 [5], [29- 31]. El fin último de estudiar la experiencia de usuario es mejorar la interacción que los usuarios tengan con los productos o servicios, por lo que la evaluación de la experiencia de usuario se vuelve un elemento imprescindible [36]. Esta evaluación se realiza por medio de un conjunto de métodos y herramientas que persiguen determinar la manera en que una persona percibe un sistema o producto antes, durante o después de utilizarlo.

Uno de los elementos importantes a considerar en la evaluación de la experiencia de usuario es qué método emplear en la medición. 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 a través del tiempo de un sistema o producto [37].

Como se mencionaba anteriormente, los cuestionarios estandarizados son instrumentos muy utilizados en evaluación de UX y, dentro de éstos, los tres más reconocidos son AttrakDiff,

UEQ y meCUE. AtrakDiff fue propuesto por Hassenzahl, Burmester y Koller en 2003 [38], mientras que el User Experience Questionnaire (UEQ) fue presentado en 2008 por Laugwitz, Held y Schrepp [39]. En cuanto a meCUE, fue propuesto en 2013 por Minge y Riedel [40].

El cuestionario AttrakDiff se basa en el modelo propuesto por Hassenzahl para describir la experiencia de usuario [41]. Está compuesto por 28 ítems clasificados en cuatro subescalas: calidad pragmática, calidad-estimulación hedónica, calidad-identificación hedónica y atractivo. Las características pragmáticas se refieren a rasgos como si un producto es predecible, confuso, simple o complicado, entre otros. Por su lado, las características hedonistas son aquellas que apelan a los sentimientos, como si un producto es aburrido, interesante, novedoso o decepcionante, que se relacionan con rasgos de estimulación, y también con rasgos de identificación y evocación, como la capacidad de un producto para que el usuario se sienta conectado con otras personas, con un sentido de pertenencia [42]. El atractivo describe el valor global del producto basado en la percepción de cualidades pragmáticas y hedónicas [43].

El cuestionario UEQ también se basa en el modelo de UX de Hassenzahl y consta de 26 ítems pertenecientes a las escalas: Atracción, Transparencia, Eficiencia, Controlabilidad, Estimulación y Novedad. Por su parte, el cuestionario meCUE se basa en el modelo de Thüring y Mahlke [44]. Se compone de 33 ítems construidos como escalas Likert de 7 puntos correspondientes a cuatro módulos, que a su vez representan sub-construcciones: percepciones del producto (Utilidad, Usabilidad, Estética Visual, Estatus y Compromiso), Emociones del Usuario (positivas y negativas), Consecuencias del Uso (Intención de Uso y Lealtad del Producto) y Evaluación General. Las percepciones del producto se refieren tanto a percepciones que en el modelo se llaman instrumentales (Utilidad y Usabilidad), como a percepciones no instrumentales (estética visual, estatus y compromiso). También incluye un diferencial semántico de 11 puntos con la pregunta "¿Cómo experimenta el producto como un todo?", y en cuyos extremos se ubican los conceptos “Malo” y “Bueno” [16].

En cuanto a la estructura de AttrakDiff y UEQ, cada ítem está constituido por un diferencial semántico de siete puntos [45]. En términos generales, el instrumento conocido como diferencial semántico, propuesto por Oswood, Suci y Tannenbaum en 1957, presenta un conjunto de adjetivos presentados en forma bipolar con respecto a un mismo concepto y se

responde marcando un punto dentro de ambos extremos, según se considera la percepción o la actitud se dirige hacia el concepto de la izquierda o de la derecha. En el caso de los cuestionarios de evaluación UX, cada ítem a responder tiene que ver con la experiencia al usar el producto. Por ejemplo, en el cuestionario UEQ se presenta un ítem representado por las palabras “Innovador - Convencional” y en medio de estas dos palabras hay siete puntos o casillas en las que el encuestado debe marcar para indicar su percepción del producto. Por un lado, está la dirección de la percepción (hacia la izquierda se considera el producto Innovador, mientras que hacia la derecha sería Convencional), y luego se marca la intensidad hacia ese concepto: cuanto más hacia la izquierda, más Innovador, y cuanto más a la derecha, sería más Convencional.

En el cuestionario meCUE, cada ítem se compone de una escala Likert de siete puntos [46]. Cabe recordar que la escala Likert toma su nombre del autor, Rensis Likert, quien la introdujo en el año 1932. En esta escala ampliamente utilizada, cada ítem a medir se presenta mediante un enunciado positivo o negativo, el cual es respondido al seleccionar el nivel de acuerdo o desacuerdo con la declaración. Los puntos extremos de la escala son: “totalmente en desacuerdo” y “totalmente de acuerdo”. Por ejemplo, en el cuestionario meCUE se presenta la siguiente afirmación: "El producto me decepciona" y el encuestado selecciona su grado de acuerdo o desacuerdo entre las opciones: "totalmente en desacuerdo", "en desacuerdo", "en general en desacuerdo", "ni de acuerdo ni en desacuerdo”, “en general de acuerdo”, “de acuerdo” y “completamente de acuerdo”.

Por su parte, UEQ+ (A modular Extension of the User Experience Questionnaire) es una extensión de UEQ presentada en 2020 [35], en la que es posible construir modularmente un cuestionario UEQ pero orientado en función de lo que se quiera evaluar. Actualmente cuenta con 20 escalas para elegir que incluyen las seis escalas del UEQ tradicional, más otras escalas como Confianza, Novedad, Claridad, Sensación Háptica, entre otras. En 2021, Klein, Hinderks, Schrepp, y Thomaschewski [47] propusieron tres escalas propias para sistemas con implementaciones de voz que se adaptan al formato UEQ+. Estas tres escalas, adoptadas oficialmente por UEQ+, son: Comportamiento de Respuesta, Calidad de Respuesta y Comprensibilidad. Cada escala está compuesta por cuatro preguntas en forma de diferencial semántico. El cuestionario UEQ+ compuesto por las escalas específicas para interfaces de

voz es uno de los instrumentos utilizados para evaluar la experiencia de usuario de los cuestionarios de evaluación UX que se implementaron en este trabajo utilizando una interfaz por voz como mecanismo de captura de las respuestas.

Interfaces de voz

En cuanto los conceptos relacionados con interfaces de voz, se debe comenzar por definir el concepto de asistente inteligente, el cual es un software que permite automatizar tareas permitiendo organizar y mantener información, mediante interacción oral con el usuario [48]. Estos asistentes funcionan como una interfaz natural dada la forma en la que los usuarios proveen información a los dispositivos durante su interacción [49].

Hay que tener presente que los asistentes 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 [50]. 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 [50].

Los asistentes inteligentes presentes en parlantes inteligentes u otros dispositivos como teléfonos, han impulsado el uso y aceptación de interfaces por voz. Sin embargo, una interacción por voz no necesariamente puede considerarse una conversación, definida como forma distintiva de uso del lenguaje natural que involucra métodos particulares para tomar turnos y ordenarlos en secuencias, la persistencia del contexto a través de los turnos y las acciones características para manejar la interacción en sí [17]. Un comando de voz que nos permite solicitar a un parlante inteligente que busque cierta información en internet, no podría considerarse una conversación.

En 2019, Moore y Arar [17] propusieron el Marco de Conversación Natural (NCF por sus siglas en inglés: *Natural Conversation Framework*) [4], un conjunto de patrones que incluyen conceptos como la organización y reparación de secuencias. Las secuencias son formas simplificadas de patrones naturales de conversación humana, identificadas previamente en la disciplina conocida como "*Conversation Analysis*", que estudia las interacciones sociales

relacionadas con las conductas verbales. Esta disciplina fue desarrollada por los sociólogos Sacks, Schegloff y Jefferson [51].

Formalmente, el marco NCF es un lenguaje de patrones conversacionales genéricos y reutilizables para diseñar interacciones conversacionales por voz que permitan una interacción natural entre el usuario y el agente conversacional, definiendo "natural" como una interacción basada en tres principios tomados de "*Conversation Analysis*": diseño en función del destinatario, minimización y reparación.

El diseño en función del destinatario se implementa adaptando las respuestas de los agentes a la audiencia objetivo y proporcionando múltiples rutas a través del mismo espacio de conversación, lo que permite a un usuario experto, por ejemplo, lograr sus metas de manera más eficiente. Minimización se refiere al hecho de que los hablantes diseñan sus enunciados para que los destinatarios puedan entenderlos con la menor cantidad de palabras posible. Los detalles innecesarios aumentan la carga cognitiva, ya que escuchar las respuestas del agente requiere tiempo y esfuerzo. Si el usuario necesita detalles o aclaraciones, ya que la minimización le impidió comprender al agente, el concepto de reparación proporcionaría el mecanismo para obtener la información que falta o parafrasear el enunciado con una explicación más larga, aunque el principio de minimización se omita temporalmente. Reparación se refiere a la capacidad de rehacer todo o parte de un enunciado que plantea dificultades al hablar, oír o comprender y por lo tanto impide que la conversación avance [17].

Los conceptos descritos anteriormente para interfaces conversacionales son de gran utilidad para crear un mecanismo de recolección de respuestas de los cuestionarios estandarizados de evaluación de UX, dotando a estos instrumentos de un medio alternativo de implementación. Este nuevo método de implementación amplía las posibilidades de uso y suma las ventajas de una interfaz de voz, como el poder realizar una evaluación UX con las manos libres o sin tener que estar frente a una computadora al realizar la evaluación.

Capítulo 3 - Trabajo Relacionado

Los tres cuestionarios de evaluación de UX estandarizados más reconocidos son AttrakDiff, UEQ y meCUE [5], que se describieron en el capítulo 2 o Marco Conceptual. Es importante retomar en este punto cómo están construidos los cuestionarios estandarizados, en cuanto a su formato, para entender el alcance de otros trabajos realizados en relación con estos cuestionarios.

Como se había explicado, los cuestionarios AttrakDiff y UEQ están compuestos de escalas en forma de diferencial semántico con cada pregunta compuesta por dos conceptos opuestos respecto de la experiencia de usuario y en la que el participante debe marcar una de las siete casillas presentes, con las que indica la dirección de su experiencia o actitud respecto del producto y la intensidad de esa actitud, acercándose más o menos al concepto elegido. Por ejemplo, en la Figura 1 se muestran los siete primeros ítems o preguntas que forman parte del cuestionario UEQ.

desagradable	○	○	○	○	○	○	○	agradable
no entendible	○	○	○	○	○	○	○	entendible
creativo	○	○	○	○	○	○	○	sin imaginación
fácil de aprender	○	○	○	○	○	○	○	difícil de aprender
valioso	○	○	○	○	○	○	○	de poco valor
aburrido	○	○	○	○	○	○	○	emocionante
no interesante	○	○	○	○	○	○	○	interesante

Figura 1. Primeras siete preguntas del cuestionario UEQ.

Attrakdiff está compuesto de 28 preguntas mientras que UEQ está construido con 26 preguntas que siguen el mismo formato de diferencial semántico. Por otro lado, el cuestionario meCUE toma un enfoque distinto ya que no utiliza el diferencial semántico, sino que está construido con 33 preguntas en forma de escala Likert, como se puede ver en la Figura 2 y una pregunta final construida con un diferencial semántico de 11 puntos [16].

	strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
The product is easy to use.	○	○	○	○	○	○	○
The functions of the product are exactly right for my goals.	○	○	○	○	○	○	○
It is quickly apparent how to use the product.	○	○	○	○	○	○	○

Figura 2. Primeras tres preguntas del cuestionario meCUE.

Habiendo descrito de forma general la estructura de AttrakDiff, UEQ y meCUE es conveniente indicar algunas propuestas de modificaciones a estos cuestionarios encontradas en la literatura académica.

En lo que se refiere a modificaciones en el formato de las preguntas se mencionó anteriormente la presentación de instrumentos que utilizan elementos visuales, en particular imágenes o caricaturas de personas para representar las emociones o actitudes hacia un producto. El usuario debe entonces marcar la imagen que más se acerca a la emoción que siente al utilizar un producto o sistema.

Figura 3. Escala PSIUS para evaluación de una máquina para hacer café.

Por ejemplo, en [14] se propone PSIUS (“Pictorial Single-Item Usability Scale”), una escala para medir el componente de Satisfacción del constructo Usabilidad, que se basa en un único ítem gráfico que reúne 5 elementos: la satisfacción con el sistema (representado por una mano con un pulgar hacia arriba o abajo), las emociones experimentadas durante el uso del sistema (una cara con diferentes expresiones) y el sistema a ser evaluado (un dibujo simplificado del sistema), una escala numérica con valores entre -4 y 4 y color de fondo (que va desde el rojo en un extremo hasta el verde en el otro). La Figura 3 presenta la escala PSIUS construida para evaluar una máquina para hacer café que aparece en la esquina derecha de cada dibujo.

Sin embargo, esta escala no se aplica a otros rubros de la experiencia de usuario como por ejemplo si el producto es convencional o inventivo, complicado o simple, confuso o estructurado, entre otros.

También existe la escala SAM, acrónimo de Self-Assessment Manikin [15] que puede verse en la Figura 4 y que presenta tres escalas en diferentes filas para medir el placer, la excitación y el dominio asociados con la reacción afectiva de una persona a un estímulo.

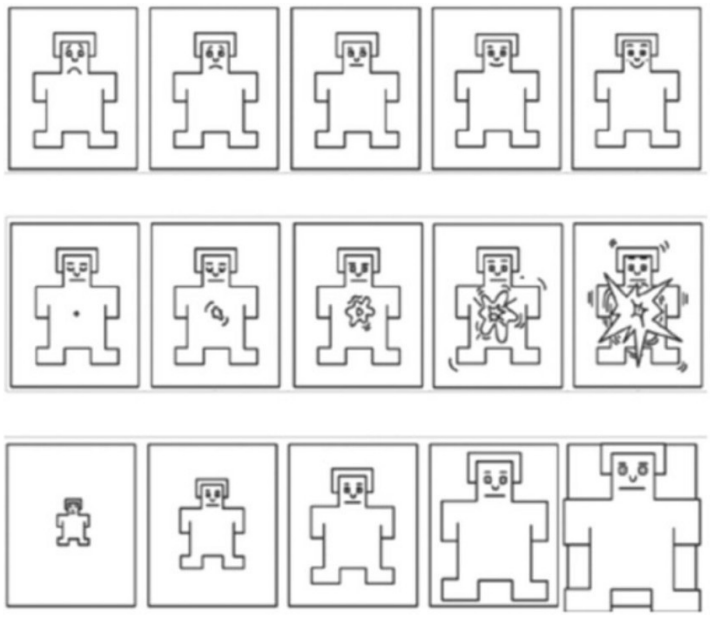


Figura 4. Escala SAM para medir la reacción afectiva a un estímulo.

Se han propuesto también otras variantes a los cuestionarios estandarizados, aunque sin presentar diferencias en el formato de las preguntas. Por ejemplo, en [18] se presenta una variación al UEQ adaptado a la cultura taiwanesa y en [19] los autores realizan una traducción de AttrakDiff al francés mediante un proceso formal de traducción propuesto en [20].

Adicionalmente, se han definido cuestionarios específicos para ciertos dominios que toman ejemplo de los cuestionarios estandarizados, pero utilizan sus propios elementos de evaluación. Por ejemplo, en [21] se propone un cuestionario para medir UX de la televisión interactiva. En [22] los autores construyen el cuestionario BUZZ para medir la experiencia de interfaces auditivas, recolectando información referente a la comprensión de lo que escuchan

los participantes y la estética de los sonidos, entre otros factores. En [23] se propone SUPR-Qm, un cuestionario para evaluar la experiencia de usuario de aplicaciones móviles.

En cuanto a las interfaces conversacionales por voz, se puede observar que los asistentes de voz inteligentes se han vuelto extremadamente populares, hasta el punto de que alrededor del 20% de las consultas de Internet en los Estados Unidos en el 2019 se realizan por voz [24]. Otras estadísticas importantes muestran que el 65% de las personas de entre 25 y 49 años hablancon sus asistentes de voz inteligentes al menos una vez al día, que el 93% de los consumidores están satisfechos con sus asistentes de voz [25] y que el 70% de las personas prefieren realizar consultas en internet por voz que escribirlas en un dispositivo [26]. Adicionalmente, cabe destacar que se espera que el mercado de asistentes de voz inteligentes pase de 2,500 millones de dólares, en 2019, a 8,000 millones de dólares en 2024, ya sea en asistentes presentes en teléfonos inteligentes, tabletas, computadoras o en altavoces o parlantes inteligentes como Amazon Echo o Google Home [26-27].

En relación con los asistentes inteligentes por voz, se identificaron estudios respecto de comparar diferentes asistentes virtuales realizando las mismas tareas, así como de evaluar distintos elementos de seguridad y privacidad [27]. Otros investigadores evalúan la experiencia de usuario desde un punto de vista emocional al utilizar estos dispositivos [28], o se enfocan en evaluar qué tan correcta o qué tan buena es la respuesta dada por un asistente inteligente por voz [29], o en la naturalidad con la que estos asistentes contestan [30]. Por su parte, en [31] se muestra cómo los asistentes inteligentes por voz ayudan a reducir la brecha tecnológica que puede existir entre usuarios con algún tipo de impedimento físico, que les restringe el uso de ciertos dispositivos [31].

En cuanto a publicaciones que utilicen patrones de diseño para interfaces conversacionales utilizando los principios expresados por Moore y Arar en el Marco de Conversación Natural [17], encontramos estudios con distintos propósitos, entre los que podemos citar a [32] con un Sistema de Conversación para aplicaciones en el dominio de Business Intelligence, [33] con interfaces conversacionales para búsqueda de información, y [34] sobre el diseño de un *chatbot* malicioso que buscaba engañar a los usuarios.

A pesar del auge identificado en los asistentes inteligentes por voz y en la formalización de las interfaces conversacionales por voz (por ejemplo con los estudios de Moore y Arar [17]) no se identificaron propuestas de implementar un cuestionario estandarizado de experiencia de usuario usando una interfaz por voz como mecanismo de captura de las respuestas de los participantes.

Capítulo 4 - Metodología

En este capítulo se define el marco metodológico, así como las actividades realizadas para cumplir con los objetivos planteados anteriormente. Se utilizó Ciencias del Diseño como marco metodológico, muy útil en campos como la Ingeniería de Software dado que su razón de estudio son los objetos y fenómenos creados por seres humanos llamados artefactos y en un contexto particular [52], en contraposición con los fenómenos naturales estudiados por las ciencias naturales y con los fenómenos humanos estudiados por las ciencias sociales. Ciencias del Diseño permite diseñar artefactos de software y obtener nuevo conocimiento a partir de la investigación del problema y la evaluación de los artefactos en la que las decisiones sobre el diseño se basan en la evidencia obtenida. Este marco metodológico establece que cuando nos planteamos un problema de investigación en Ingeniería de Software, nos enfrentamos a un conjunto de problemas de diseño, que pertenecen al dominio ingenieril, y un conjunto de preguntas de investigación, que corresponden al dominio del conocimiento [53].

Siguiendo la categorización para diseño de la investigación de Saunders y Tosey [54], en la selección metodológica se utilizaron métodos cuantitativos de obtención de datos, que se aplicaron en forma de cuestionarios estandarizados de evaluación UX, tanto en su forma estándar (UEQ) como en la versión modular (UEQ+) como parte de casos de estudio. Estos casos de estudio fueron la estrategia para responder a la pregunta de investigación y se aplicaron de forma transversal, es decir, abordaron el problema en un momento determinado de tiempo, en oposición a lo que serían casos longitudinales, en los que se recopilan datos en un período de tiempo prolongado [54]. Estos tres elementos (selección metodológica, estrategias y horizonte de tiempo) se explican a continuación como parte de las etapas de Desarrollo y Evaluación.

Vaishnavi y Kuelcher definen cinco etapas en las que se divide el proyecto en Ciencias del Diseño [52]: Conciencia del Problema, Sugerencia, Desarrollo, Evaluación y Conclusión. Seguidamente se explica el trabajo desarrollado en esta investigación para cada una de ellas.

Conciencia del Problema

La conciencia del problema se refiere a la identificación de un problema de investigación interesante o la aplicación de nuevos hallazgos dentro de una disciplina de referencia. En esta investigación, este punto se aborda en la justificación del capítulo 1. Introducción, y se refiere a la intención de buscar nuevas formas de interacción para la recolección de datos de cuestionarios estandarizados de evaluación UX, dado que el formato usual escrito podría acarrear algunos inconvenientes a los participantes en el momento de llenarlos.

Sugerencia

La Sugerencia es la etapa que sigue inmediatamente a la Conciencia del Problema, y es un paso creativo en el que se visualiza una nueva funcionalidad basada en una configuración novedosa de elementos existentes o nuevos. En este proyecto de investigación, la Sugerencia también ha sido descrita en detalle en los capítulos anteriores de este documento, en las que se explicó la intención de utilizar interfaces por voz, particularmente los elementos conversacionales de estas interfaces, como nuevo mecanismo de captura de un cuestionario conversacional por voz.

Las etapas de Conciencia del Problema y Sugerencia están muy ligadas en Ciencias del Diseño, y tienen como salidas usuales una propuesta de investigación y un diseño tentativo.

Desarrollo

La siguiente etapa es la de Desarrollo, que tiene como salida un artefacto. Este artefacto tiene una implementación simple y no necesita involucrar una novedad más allá del estado de la práctica para el artefacto dado, puesto que la novedad está principalmente en el diseño y no en la construcción del artefacto.

En este trabajo se diseña una interfaz conversacional por voz que pueda utilizarse como mecanismo de captura de las respuestas de un cuestionario estandarizado de evaluación de experiencia de usuario. Inicialmente se diseñó una interacción que reflejara lo más directamente posible el formato de las preguntas del cuestionario UEQ (User Experience

Questionnaire), que se presentan en forma de diferencial semántico, así como la estructura de 26 preguntas del cuestionario.

En esta implementación que determinamos como "directa" una pregunta presentada de forma escrita mediante un diferencial semántico como la que se muestra en la Figura 5, sería implementado por la pregunta siguiente: en la escala de 1 a 7, donde 1 significa "desagradable" y 7 significa "agradable", ¿cómo calificaría este producto?

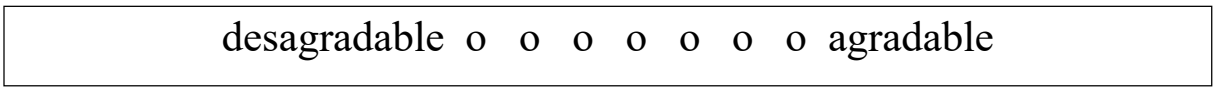


Figura 5. Diferencial semántico para la pregunta "desagradable/agradable".

En una segunda iteración, se modificó la estructura de cada pregunta de forma que la recolección de información se pareciera más un intercambio propio de una conversación. Así, por ejemplo, para realizar la captura de la evaluación del diferencial semántico descrito en la Figura 5, se realizan dos preguntas. Primero se le consulta al participante por la dirección de la actitud: ¿Usted considera que el producto es desagradable, agradable o ni uno ni otro? Una vez establecida la dirección se le pregunta por la intensidad, por ejemplo: ¿Y qué tan agradable, un poco, mucho o en extremo?

Además, se identificaron patrones conversacionales propios del *framework* NCF de Moore y Arar [17] que se adapten al cuestionario UEQ, de forma que la captura de respuestas del cuestionario buscara una conversación "natural", es decir, incluyendo un diseño en función del destinatario, minimización y reparación de secuencias.

Las interfaces definidas como mecanismo de captura de respuestas del cuestionario UEQ fueron implementadas mediante la plataforma VoiceFlow que permite programar los flujos conversacionales en los que se presentó el cuestionario a los participantes. Esta plataforma provee la capacidad de generar programas que se ejecutan en parlantes inteligentes (como *skills* de Alexa o *actions* de Google Assistant) y también brinda la opción de poner a disposición de los participantes la interfaz de voz implementada a través de un sitio web, por lo que para ejecutar la evaluación de un producto o servicio basta con proveer a los

participantes la dirección URL de la interacción creada. Esto facilita la aplicación de los casos de estudio, dado que es posible distribuir el instrumento de evaluación de forma remota, sin necesidad de requerir que los participantes se desplacen a un sitio específico a interactuar con un altavoz inteligente o que deban contar con uno personal.

Evaluación

Para validar las interfaces creadas se ejecutaron dos casos de estudio. En el primero se evaluó el primer diseño creado (implementación por voz directa, sin patrones conversacionales) comparando los resultados obtenidos de evaluar la anticipación de experiencia de usuario de dos productos, con una versión equivalente del cuestionario pero que utiliza el formato escrito usual. Los productos por evaluar fueron seleccionados considerando que se tratara de productos con un componente tecnológico interesante, y se presentaron a los participantes en forma de video, con una duración similar y en idioma español. Un par de zapatillas deportivas y una billetera fueron los productos seleccionados. Las zapatillas tienen características como: proyectar imágenes y transmitir impulsos al que las portaba. La billetera permite cargar dispositivos vía USB y cuenta con GPS para poder recuperarla en caso de que el dueño la extravíe.

La comparación de los resultados de la evaluación UX de estos productos se realizó utilizando los promedios de las seis escalas del cuestionario UEQ: Atracción, Transparencia, Eficiencia, Controlabilidad, Estimulación y Novedad, evaluando la semejanza de estas escalas en las implementaciones conversacional por voz y escrita tradicional.

En este primer caso de estudio, también se evaluó la experiencia de usuario de la implementación conversacional por voz utilizando el cuestionario UEQ+ (*A modular Extension of the User Experience Questionnaire*) utilizando las tres escalas propuestas por Klein et al [47] para la evaluar interfaces de voz: 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) y Comprensibilidad (el Asistente de Voz comprende correctamente las instrucciones de los usuarios sin obligarlos a hablar de forma no natural). También se agregaron en esta evaluación los módulos Uso Intuitivo y Novedad que, aunque no son

específicas para interfaces de voz, se presentaban como interesantes de evaluar dadas las características de las implementaciones realizadas. Se incluyó también en esta evaluación un conjunto de preguntas propias del concepto de Usabilidad, que, aunque está contenido en el concepto de UX, podrían quedar menos cubiertas dada la selección de módulos realizada en UEQ+. Las preguntas incluidas en formato de afirmación pueden verse en la Tabla 1.

Elementos consultados de Usabilidad
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.

Tabla 1. Elementos consultados de Usabilidad

Se efectuó posteriormente un segundo caso de estudio en el que la implementación del mecanismo de captura de la interfaz conversacional incluyó patrones conversacionales de voz. Se compararon nuevamente los resultados de las seis escalas de UEQ con la implementación tradicional del cuestionario UEQ cuyo mecanismo de captura es escrito. Al igual que en el primer caso de estudio, también se evaluó la experiencia de usuario de la interfaz conversacional por voz utilizando el cuestionario UEQ+ con las escalas propias de calidad de voz, junto con los módulos Uso Intuitivo y Novedad, así como con las preguntas de Usabilidad señaladas en la Tabla 1.

Finalmente, se contrastaron las evaluaciones obtenidas con UEQ+ para ambas implementaciones conversacionales, para determinar el impacto de la incorporación de patrones conversacionales en la usabilidad y evaluación de experiencia de usuario del instrumento.

En cada caso de estudio se utilizó una población de 40 participantes, para un total de 80 personas en total. Los participantes pertenecen a la carrera Bachillerato en Computación de la Universidad de Costa Rica. Cabe señalar que la tendencia identificada en [16] respecto de la cantidad de participantes en estudios primarios cuya evaluación utiliza cuestionarios estandarizados de evaluación UX señala que la media es de 20 participantes para estudios presentados en conferencias y 30 participantes para estudios publicados en revistas.

Actividades

Como punto adicional de la metodología antes descrita, en la Figura 6 se presentan las actividades desarrolladas para cada objetivo definido en el capítulo 1 de Introducción.

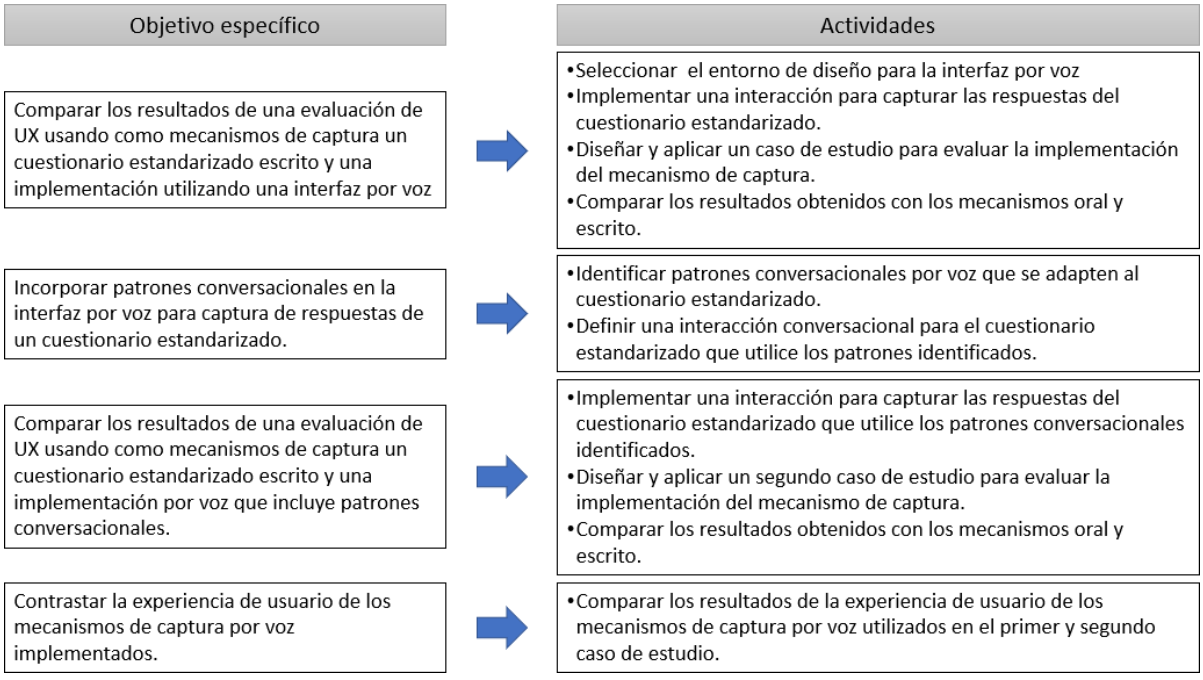


Figura 6. Actividades por Objetivo específico.

Para el objetivo específico 1 se ejecutaron las siguientes actividades:

Seleccionar el entorno de diseño para la interfaz por voz. Se identificó la interfaz por voz que se utilizó para implementar el cuestionario.

Implementar una interacción para capturar las respuestas del cuestionario estandarizado: Una vez seleccionado el entorno para la interfaz por voz se procedió a familiarizarse con el ambiente y a desarrollar los diferentes pasos del flujo conversacional, replicando los mismos pasos para el método de captura escrito.

Diseñar y aplicar un caso de estudio para evaluar la implementación del mecanismo de captura: Completada la implementación de cuestionario en los métodos de captura escrito y por voz se diseñó y ejecutó el primer caso de estudio con un grupo inicial de 40 participantes.

Comparar los resultados obtenidos con los mecanismos por voz y escrito: Ejecutado el caso de estudio se realizó el procesamiento de las respuestas de los participantes que permitió comparar los resultados de la evaluación UX obtenidos por medio de cada uno de los métodos de captura.

Para el objetivo específico 2 se ejecutaron las actividades siguientes:

Identificar diferentes patrones conversacionales por voz que se adapten al cuestionario estandarizado: Se realizó una identificación de patrones que pudieran utilizarse en interacciones conversacionales y que se adaptaran al estudio ejecutado.

Definir una interacción conversacional para el cuestionario estandarizado que utilice los patrones conversacionales identificados: Una vez identificados los patrones a utilizar, se procedió a afinar la definición de la interacción por voz implementada inicialmente incorporando los patrones conversacionales identificados.

En cuanto al objetivo específico 3, se ejecutaron las siguientes actividades:

Implementar una interacción para capturar las respuestas del cuestionario estandarizado que utilice los patrones conversacionales identificados: Se implementa una vez más el cuestionario estandarizado, esta vez realizando los cambios respectivos aplicando los patrones conversacionales identificados.

Diseñar y aplicar un segundo caso de estudio para evaluar la implementación del mecanismo de captura: Completada la implementación del cuestionario aplicando patrones conversacionales se procede a diseñar un segundo caso de estudio que se aplicó a un grupo nuevo de 40 participantes.

Comparar los resultados obtenidos con los mecanismos oral y escrito: Ejecutado el caso de estudio se realizó el procesamiento de las respuestas de los participantes, en la que se comparan los resultados obtenidos con el mecanismo escrito tradicional y el mecanismo implementado con la interfaz conversacional por voz.

Finalmente, se realizaron las siguientes actividades para cumplir con el objetivo específico 4:

Ejecutados el primer y el segundo caso de estudio se procede a realizar un análisis y comparación de los resultados obtenidos del primer caso de estudio con el segundo, en relación con la experiencia de usuario (UX) y Usabilidad de ambos mecanismos por voz implementados y así obtener las conclusiones del estudio respecto del impacto en la experiencia de usuario al utilizar los patrones conversacionales.

Capítulo 5 - Resultados

En el primer caso de estudio, se realizó una implementación por voz, pero sin patrones conversacionales, donde los ítems de diferencial semántico del cuestionario escrito se transforman en preguntas del tipo: en la escala de 1 a 7, donde 1 significa “desagradable” y 7 significa “agradable”, ¿cómo calificaría este producto?

En este caso se ubicó a los participantes en cuatro grupos, combinando producto y mecanismo de captura. Como se indicó en la sección de Metodología se seleccionaron dos productos que tuvieran algún componente tecnológico atractivo. Los elegidos fueron una billetera que puede cargar dispositivos USB y tiene GPS, y unas zapatillas que proyectan imágenes y generan impulsos hápticos al que las porta. En cuanto a las implementaciones de los cuestionarios, se utilizó VoiceFlow para la versión por voz, mientras que para la versión escrita se utilizó un cuestionario Google Forms.

En una primera parte, 20 personas evaluaron los productos indicados, separados en 4 grupos de 5 personas cada uno. En la Tabla 2 se lista la configuración de los grupos.

Producto evaluado	Mecanismo de captura
Billetera	Google Forms
Billetera	VoiceFlow
Zapatillas	Google Forms
Zapatillas	VoiceFlow

Tabla 2. Organización de los grupos por producto evaluado y mecanismo utilizado.

Posteriormente se realizó una segunda ronda de evaluación, con otros 20 participantes siguiendo el mismo patrón de distribución de producto y cuestionario indicado arriba.

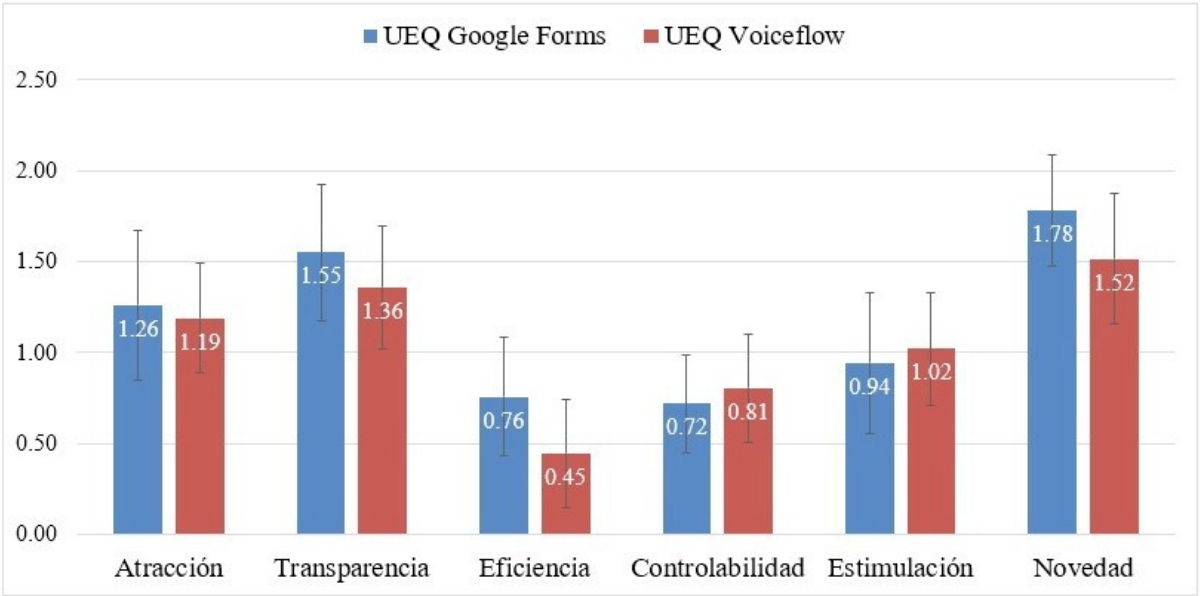
La descripción de este caso de estudio y los resultados obtenidos se presentaron en la conferencia ICITS 2022 (*International Conference on Information Technology & Systems*) y se publicaron como parte de los libros de la serie *Lecture Notes in Networks and Systems de Springer*: *Comparing Written and Voice Captured Responses of the User Experience*

Questionnaire (UEQ)". Mata-Serrano, Díaz-Oreiro, López y Guerrero. DOI 10.1007/978-3-030-96293-7_43 2.

En el Anexo 1 se encuentra el artículo completo tal y como fue publicado.

Entre los resultados más importantes de esta primera implementación, conviene resaltar el hecho de que no hubo diferencias significativas en las evaluaciones de ambos productos cuando se agrupan las evaluaciones de los 40 participantes. Esto se comprobó realizando una prueba t de Student que provee UEQ dentro de su conjunto de herramientas de análisis, que permite comparar las evaluaciones de dos productos distintos o dos versiones de un mismo producto. Cuando solamente se tomaban 20 participantes por mecanismo de evaluación, los resultados sí mostraban diferencias significativas en las escalas de Estimulación y Novedad.

La Figura 7 muestra los resultados de las evaluaciones comparadas de los cuestionarios UEQ cuyo mecanismo de captura está implementado por escrito (Google Forms) o por voz (VoiceFlow).



	Atracción		Transparencia		Eficiencia		Controlabilidad		Estimulación		Novedad	
Media	1.26	1.19	1.55	1.36	0.76	0.45	0.72	0.81	0.94	1.02	1.78	1.52
Desv.Std	1.33	0.97	1.21	1.09	1.05	0.96	0.88	0.96	1.26	1.00	0.99	1.16
Confianza	0.41	0.30	0.37	0.34	0.33	0.30	0.27	0.30	0.39	0.31	0.31	0.36

Figura 7. Evaluación UX comparando método de captura (N=40).

La Tabla 3 presenta la prueba t de Student que muestra que no hay diferencias significativas en los resultados de ninguna de las seis escalas de UEQ, al comparar solo por método de captura, donde la cantidad de participantes es 40.

Escala UEQ	Valor-t	Diferencia significativa
Atracción	0.7763	No
Transparencia	0.4632	No
Eficiencia	0.1751	No
Controlabilidad	0.6722	No
Estimulación	0.7623	No
Novedad	0.2766	No
		Alpha level: 0.05

Tabla 3. Prueba t de Student por método de captura en cada escala UEQ (N=40).

Otro resultado interesante para mencionar es la cantidad de inconsistencias encontradas usando cada formato de captura. La hoja de cálculo de procesamiento de UEQ advierte que puede suceder que no todos los participantes respondan todos los ítems con seriedad y, para detectar respuestas aleatorias o no serias, la hoja de cálculo de la UEQ utiliza una heurística simple. Todos los ítems en una escala deben medir un aspecto de calidad UX similar, por lo que la hoja de cálculo verifica cuánto difiere la mejor y la peor evaluación de un ítem en una escala. Si hay una gran diferencia (más de tres), esto se ve como un indicador de un patrón de datos problemático. UEQ considera críticas las respuestas de los participantes con tres o más inconsistencias y sugiere eliminarlas del análisis.

Formato	Inconsistencias	Respuestas críticas
Google Forms	45 / 240 (18.75%)	4 / 40 (10.00%)
VoiceFlow	60 / 240 (25.00%)	7 / 40 (17.50%)

Tabla 4. Inconsistencias detectadas por método de captura (N=40).

La Tabla 4 muestra el número total de inconsistencias detectadas por formato de captura, así como el número de respuestas críticas que son aquellas respuestas de un participante con tres o más escalas marcadas como inconsistentes. Se puede observar que existen más inconsistencias en los cuestionarios utilizando el formato conversacional implementado en

VoiceFlow ya sea por inconsistencias totales o por respuestas consideradas críticas, en esta primera implementación del mecanismo de voz.

Como se indicó en el capítulo de metodología, el cuestionario UEQ implementado con la interfaz conversacional se evaluó con el cuestionario UEQ+, formado con las escalas Comprensibilidad, Comportamiento de Respuesta y Calidad de Respuesta desarrolladas para interfaces de voz junto con Uso Intuitivo y Novedad.

UEQ+ Escala	Media de la escala	Ítem Izquierda	Ítem Derecha	Media del ítem
Comprensibilidad	0.72	complicado	simple	0.63
		ambiguo	inequívoco	0.45
		impreciso	preciso	0.53
		enigmática	explicable	1.28
Comportamiento de la respuesta	0.28	artificial	natural	-0.25
		desagradable	agradable	1.03
		antipático	simpático	0.78
		aburrido	entretenido	-0.45
Uso Intuitivo	1.36	difícil	fácil	1.68
		ilógico	lógico	1.38
		equívoco	evidente	1.10
		incoherente	coherente	1.28
Calidad de la respuesta	0.89	inapropiado	adecuado	2.13
		inútil	útil	1.80
		no proveen ayuda	proveen ayuda	-1.18
		ignorante	inteligente	0.83
Novedad	0.36	falto de imaginación	creativo	0.88
		convencional	original	0.00
		tradicional	novedoso	0.15
		conservador	innovador	0.43

Tabla 5. Valores de la media por escala e ítems evaluados con UEQ+.

La Tabla 5 muestra que las medias para cada escala, que varían entre -3 y 3, indican que esta implementación de UEQ no es muy buena en términos de evaluación de UX. La escala mejor evaluada es Uso Intuitivo, seguida de Calidad de Respuesta. Las medias individuales de cada ítem de la escala indica que los participantes consideran el cuestionario adecuado, útil, fácil, lógico, concluyente, explicable y plausible. Al mismo tiempo, los ítems peor calificados revelan que los participantes consideran que esta implementación conversacional no es útil, es aburrida, artificial, convencional, habitual y conservadora.

Además del UEQ+, se evaluaron aspectos de Usabilidad para la versión conversacional del UEQ mediante un conjunto de 10 preguntas en formato de escala Likert. Los resultados obtenidos, que se muestran en la Figura 8, muestran algunos aspectos positivos del Asistente, tales como: la interacción es clara desde la primera pregunta, es fácil saber si el asistente entendió mi respuesta, así como saber cuántas preguntas se han hecho. Entre los aspectos negativos: no es posible corregir una respuesta dada, no es fácil pedir ayuda al Asistente, el Asistente habla más de lo necesario y era común olvidar lo que había pedido el Asistente. Algunas otras características negativas identificadas podrían atribuirse a la implementación del mecanismo de captura de datos de voz, pero también a las características inherentes de UEQ, como preguntar por el mismo concepto más de una vez y tener demasiadas preguntas.

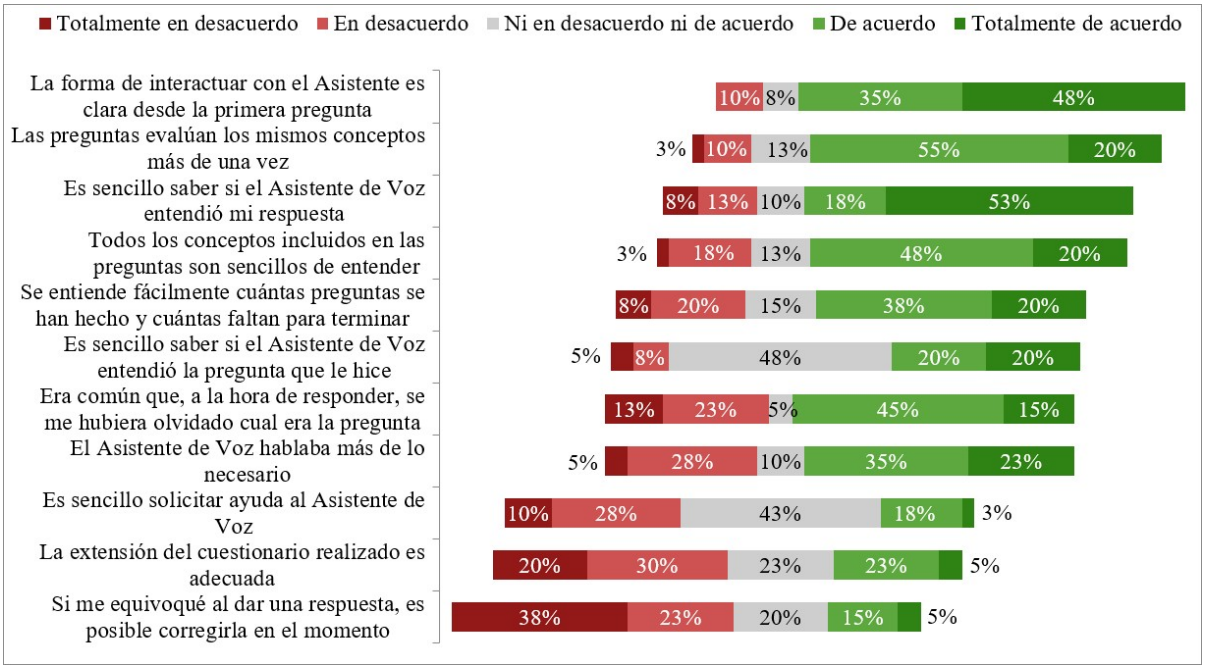


Figura 8. Evaluación de Usabilidad de la interfaz por voz.

Como es mencionó anteriormente, los resultados muestran que no habría diferencia significativa al utilizar cualquiera de los dos formatos de captura (escrito o por voz) para evaluar la UX anticipada de dos productos, para las seis escalas del cuestionario UEQ. Sin embargo, existen muchas oportunidades para mejorar la usabilidad y la experiencia de usuario del asistente de voz que implementa UEQ.

Con estos resultados obtenidos de la primera implementación por voz del mecanismo de captura de UEQ, se implementó una segunda versión del citado mecanismo, pero usando un formato diferente para formular las preguntas de forma que la interacción se perciba más natural. Como se indicó antes, el ítem diferencial semántico se modificó para capturar la información mediante dos preguntas: una pregunta inicial en la que se identifica la dirección de la actitud y una segunda pregunta en la que se mide la intensidad en la dirección ya seleccionada.

Otras modificaciones al mecanismo de captura por voz incluyeron crear flujos de conversación más amigables e interactivos donde el participante pueda estar al tanto del progreso del cuestionario en cualquier momento, tener un mayor control y libertad de flujo y evitar errores. Por ejemplo, se implementó en el Asistente de voz la capacidad de brindar la definición de los conceptos que forman cada diferencial semántico y el que el asistente pueda repetir preguntas o solicitar al participante que repita una pregunta que no entendió. También se incluyó que el Asistente de voz confirme que entendió la respuesta obtenida mediante un conjunto de frases que alterna aleatoriamente, y la capacidad de informar al participante en qué pregunta va del cuestionario y cuántas faltan. Adicionalmente, el participante puede solicitar cancelar el proceso de llenar el cuestionario, a lo que el Asistente accede, pero no sin antes advertirle al participante que las respuestas no están completas, y lo que a su vez permite al participante decidir si de verdad termina anticipadamente el proceso o no. Todos estos patrones conversacionales fueron identificados en el Trabajo Final de Investigación Aplicada de Marco Chacón Chaves [55].

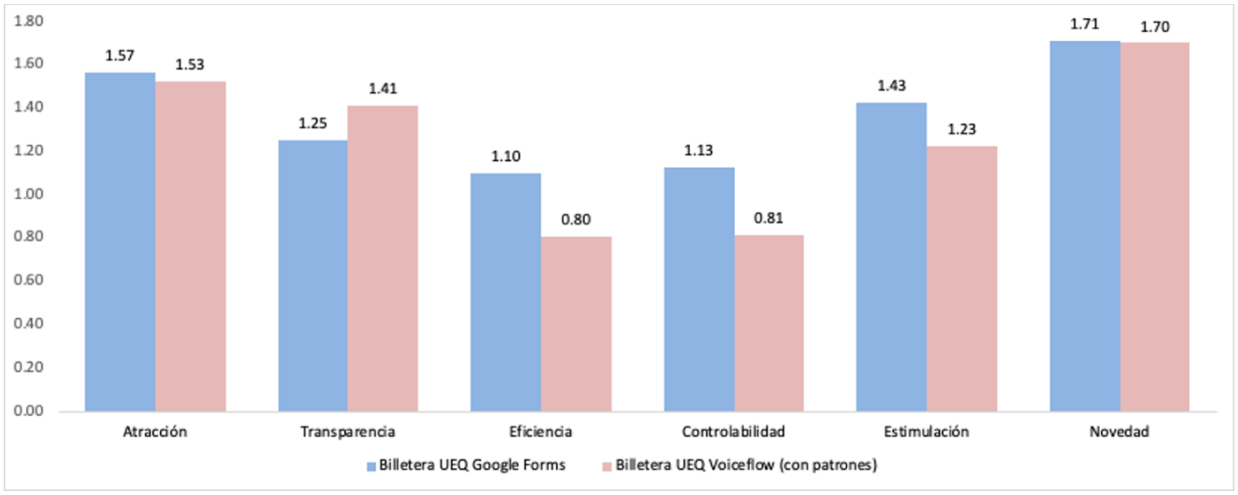
Además de los patrones conversacionales antes citados, se agregó también la posibilidad de corregir una respuesta luego de la pregunta inicial, y la capacidad de reconocer la pregunta inicial (dirección de la actitud o sentimiento) y la pregunta complementaria (intensidad) en una sola. Así, ante la pregunta “¿usted considera el producto desagradable, agradable, o ni

uno ni otro?", el participante puede contestar: "me parece que es en extremo desagradable", y el Asistente reconoce la respuesta de la segunda pregunta y entonces ya no la formula. Todos estos patrones y la conformación del diferencial semántico en dos preguntas se formalizaron en una investigación complementaria que se presentó en la conferencia ICITS 2022 y que se publicó en el artículo *Conversational Design Patterns for a UX Evaluation Instrument Implemented by Voice* [56].

Con la nueva implementación del mecanismo de captura por voz, esta vez con patrones conversacionales, se llevó a cabo un segundo caso de estudio, diseñado de la misma forma que el primer caso descrito anteriormente. Es decir, se distribuyeron los participantes en cuatro grupos, combinando producto y mecanismo de captura: Billetera / Google Forms, Billetera / Voiceflow, Zapatillas / Google Forms y Zapatillas / Voiceflow.

Al igual que con la primera implementación del mecanismo por voz, no se encontraron diferencias significativas entre las evaluaciones de los dos productos, con la diferencia de que, en esta segunda implementación con patrones conversacionales, ya desde los 20 participantes se notó que no había diferencia en ninguna de las 6 escalas de UEQ. Cabe recordar que en la primera implementación, se necesitó una muestra mayor (40 participantes) para llegar a este mismo punto de convergencia.

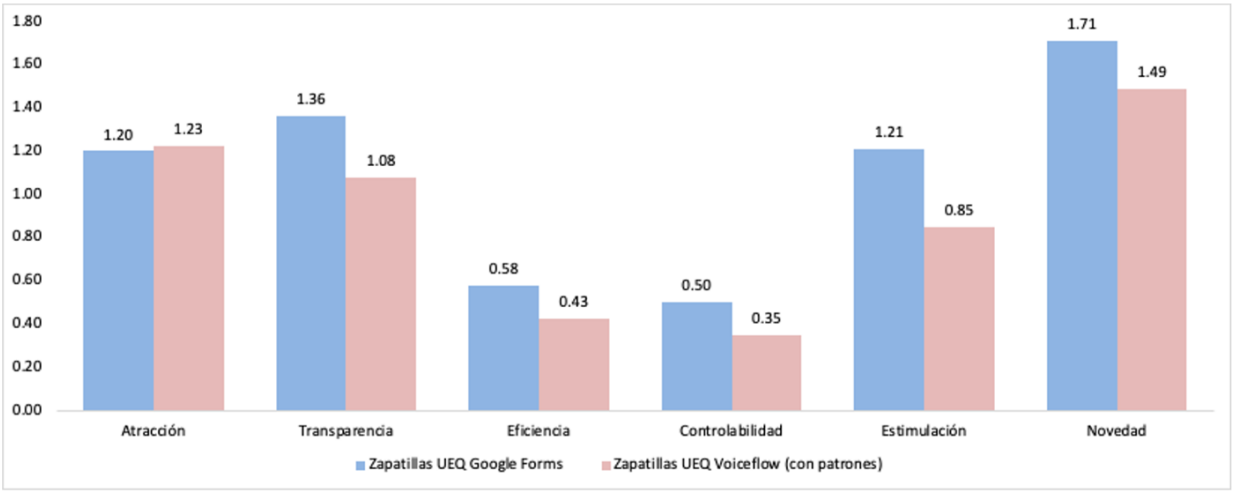
Por ejemplo, en la Figura 9 muestra las medias, la desviación estándar y el alfa que se define como la probabilidad a priori de que el intervalo de confianza a calcular contenga el verdadero valor del parámetro [57], correspondientes a cada escala UEQ obtenida de los formatos de captura de Google Forms y Voiceflow para la billetera.



Mean	1.57	1.53	1.25	1.41	1.10	0.80	1.13	0.81	1.43	1.23	1.71	1.70
Std. Dev.	1.05	0.73	1.16	0.99	0.92	0.87	0.99	0.70	1.35	0.78	1.32	0.92
Confidence	0.46	0.32	0.51	0.43	0.40	0.38	0.43	0.31	0.59	0.34	0.58	0.40

Figura 9. Evaluación UX para el producto billetera por método de captura (N=20).

De la misma forma, la Figura 10 presenta las medias, la desviación estándar y los rangos de confianza, esta vez para las zapatillas, comparando el cuestionario con el mecanismo de captura escrito con el que utiliza patrones conversacionales.



Mean	1.20	1.23	1.36	1.08	0.58	0.43	0.50	0.35	1.21	0.85	1.71	1.49
Std. Dev.	1.19	0.81	1.27	0.62	1.14	0.61	1.13	0.56	1.27	0.86	1.06	0.83
Confidence	0.52	0.36	0.55	0.27	0.50	0.27	0.49	0.25	0.56	0.38	0.46	0.37

Figura 10. Evaluación UX para el producto zapatillas por método de captura (N=20).

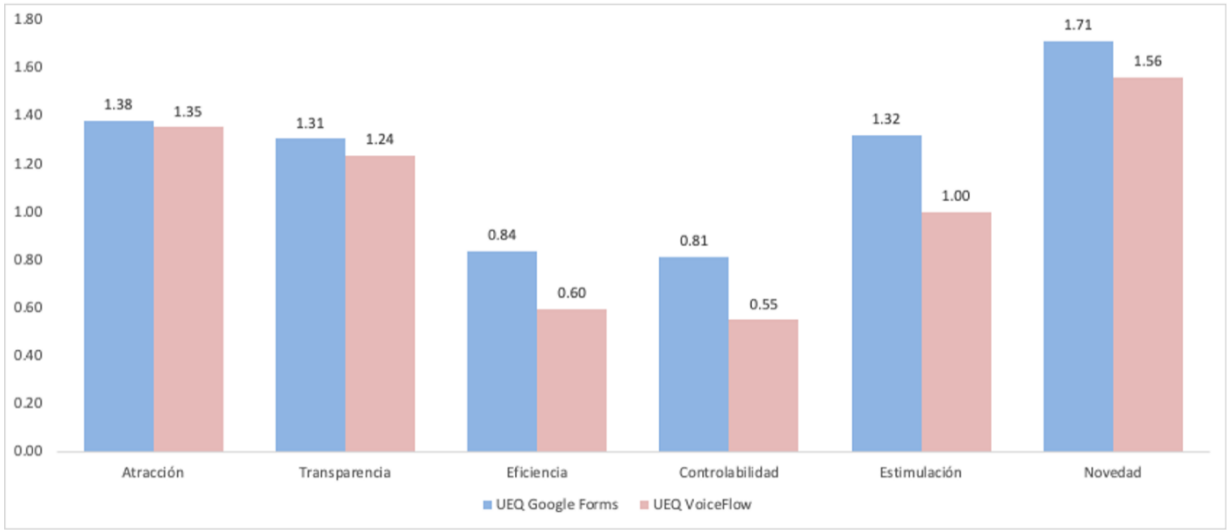
Para comprobar que no hay diferencia en las evaluaciones de los dos mecanismos, se realizó una prueba t de Student en cada escala comparando los métodos de captura escrito y de voz aplicando patrones conversacionales.

Escala UEQ	Billetera Valor t	¿Diferencia significativa?	Zapatillas Valor t	¿Diferencia significativa?
Atracción	0.8853	No	0.9385	No
Transparencia	0.6369	No	0.3693	No
Eficiencia	1.3033	No	0.6086	No
Controlabilidad	0.2580	No	0.5986	No
Estimulación	0.5696	No	0.2982	No
Novedad	0.9725	No	0.4599	No

Tabla 6. Prueba t de Student por método de captura para billetera y zapatillas, con patrones conversacionales.

Como se puede observar en la Tabla 6, tanto para el producto de la billetera como para el producto de zapatillas no existen diferencias significativas.

Seguidamente, en la Figura 11 se muestran las medias, la desviación estándar y los rangos de confianza comparados de los dos métodos de captura (escrito en Google Forms y por voz implementado en VoiceFlow con patrones conversacionales), esta vez agrupando los datos de los 40 participantes.



Mean	1.38	1.35	1.31	1.24	0.84	0.60	0.81	0.55	1.32	1.00	1.71	1.56
Std. Dev.	1.12	0.78	1.20	0.84	1.06	0.77	1.09	0.65	1.29	0.81	1.18	0.87
Confidence	0.35	0.24	0.37	0.26	0.33	0.24	0.34	0.20	0.40	0.25	0.37	0.27

Figura 11. Evaluación UX comparando método de captura (N=40), con patrones conversacionales.

Escala UEQ	Valor-t	Diferencia significativa
Atracción	0.8953	No
Transparencia	0.7676	No
Eficiencia	0.2537	No
Controlabilidad	0.2007	No
Estimulación	0.1931	No
Novedad	0.5252	No
		Alpha level: 0.05

Tabla 7. Prueba t de Student por método de captura en cada escala UEQ (N=40), con patrones conversacionales.

En este caso, al tener una muestra aún más grande (N = 40), la Tabla 7 presenta la prueba t de Student que muestra que no hay diferencias significativas en los resultados de ninguna de las seis escalas de UEQ, al comparar solo por método de captura, pero esta vez utilizando patrones conversacionales.

De igual forma que en la primera implementación del mecanismo de voz, donde no se utilizaron patrones, se evaluaron aspectos de Usabilidad para la versión conversacional del UEQ, al incorporarle patrones, mediante un conjunto de 10 preguntas en formato de escala Likert. Los resultados obtenidos, que se muestran en la Figura 12, muestran algunos aspectos positivos del Asistente, tales como: es fácil preguntarle al Asistente por ayuda, todos los conceptos incluidos en el cuestionario son fáciles de entender. Entre los aspectos negativos con el Asistente habla más de lo necesario, es posible corregir una respuesta incorrecta en el momento y la extensión del cuestionario no es adecuada. En la última pregunta en la Figura 12 los participantes indican que no era común que olvidaran la pregunta que se les había formulado, lo que es un rasgo positivo de la implementación por voz con patrones conversacionales. La versión sin patrones fue peor calificada en este punto.

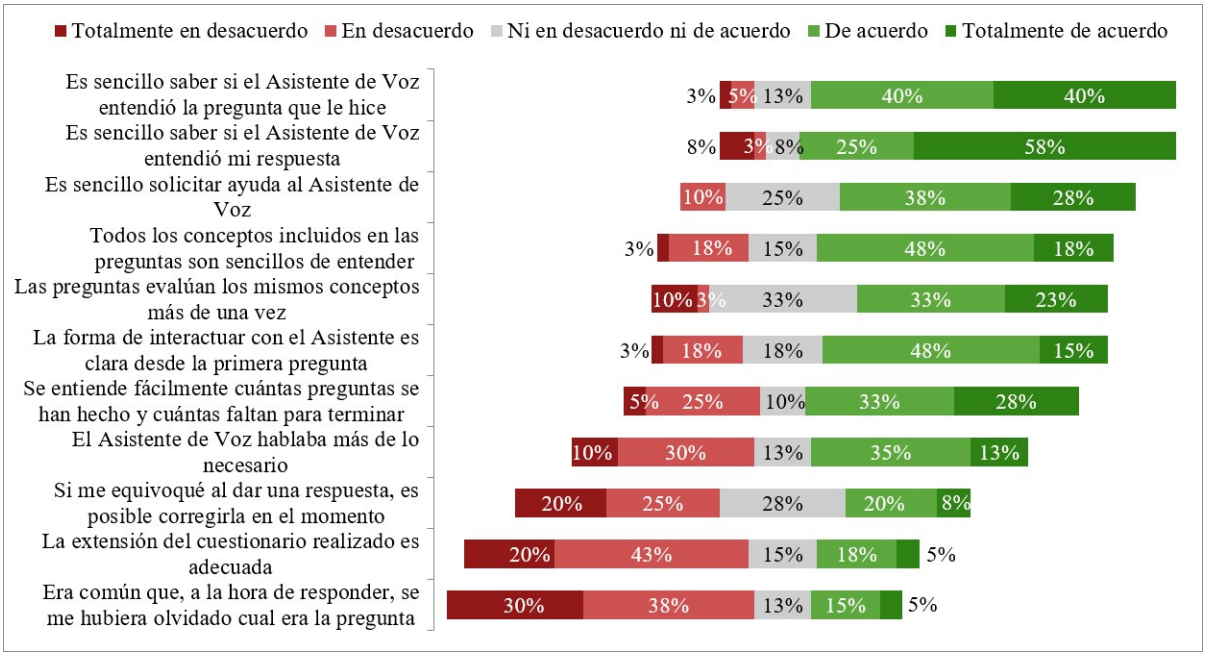


Figura 12. Evaluación de Usabilidad de la interfaz por voz, utilizando patrones conversacionales.

Haciendo una comparación de las evaluaciones de los mecanismos por voz en las preguntas de Usabilidad arriba mencionadas, las Figuras 13 y 14 muestran de manera contrastada el comportamiento de los resultados estas dos evaluaciones de la interfaz por voz, para las implementaciones sin y con patrones conversacionales.

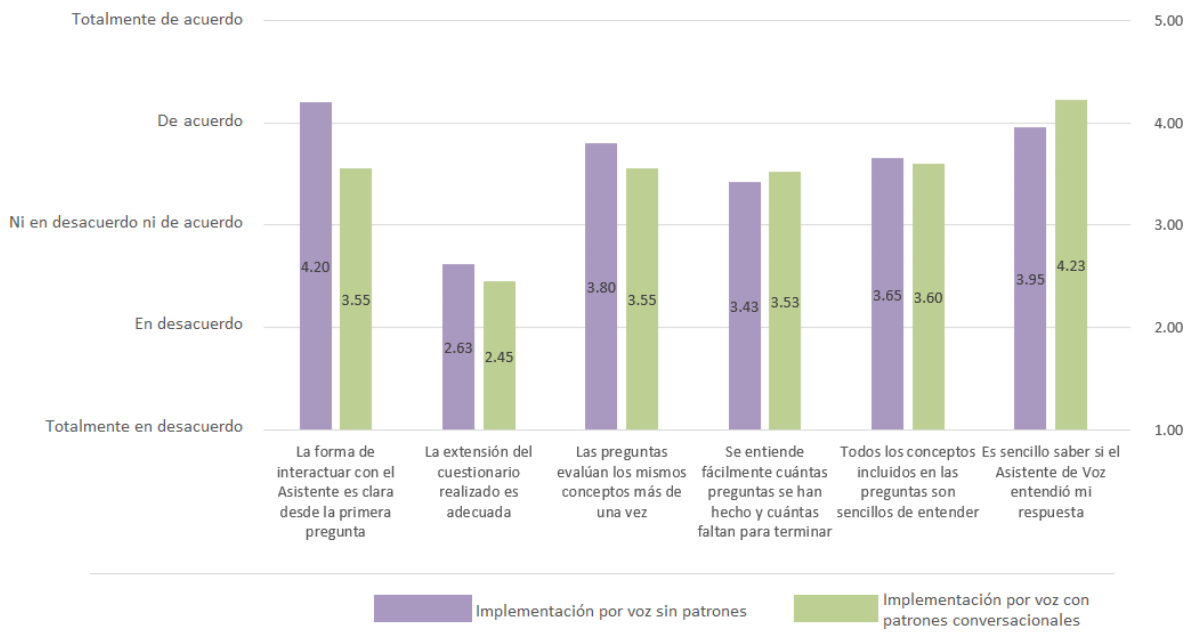


Figura 13. Evaluación de Usabilidad para ambas implementaciones de mecanismos de voz. Parte I.

Al analizar estos contrastes podemos ver en Figura 13 aspectos como: la claridad al interactuar con el asistente disminuye al incorporar patrones conversacionales, hay mejoras en entender con mayor facilidad cuantas preguntas se han hecho y cuantas faltan para terminar, Incorporando patrones conversacionales, es más sencillo saber si el asistente de voz entendió la respuesta.

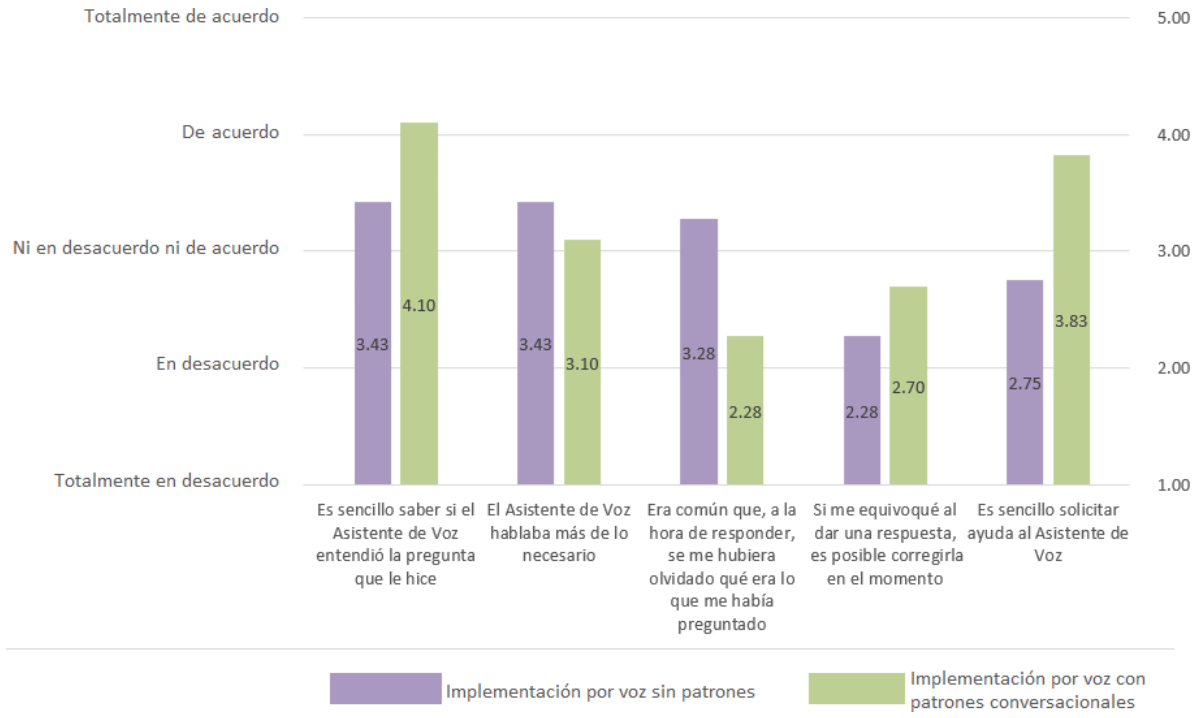


Figura 14. Evaluación de Usabilidad para ambas implementaciones de mecanismos de voz. Parte II.

En el caso de la Figura 14 vemos una mejora considerable al incorporar patrones en saber si el Asistente de Voz entendió la pregunta que se le hace y en solicitar ayuda al Asistente. Por otro lado, sin patrones conversacionales es más frecuente que a la hora de responder los participantes olvidaran lo que el agente les preguntaba.

Por otro lado, si revisamos los resultados de las evaluaciones UX de las implementaciones por voz, con y sin patrones, la Figura 15 muestra las respuestas obtenidas del cuestionario UEQ+, para las escalas de Comprensibilidad, Comportamiento de Respuesta, Uso Intuitivo, Calidad de Respuesta y Novedad.

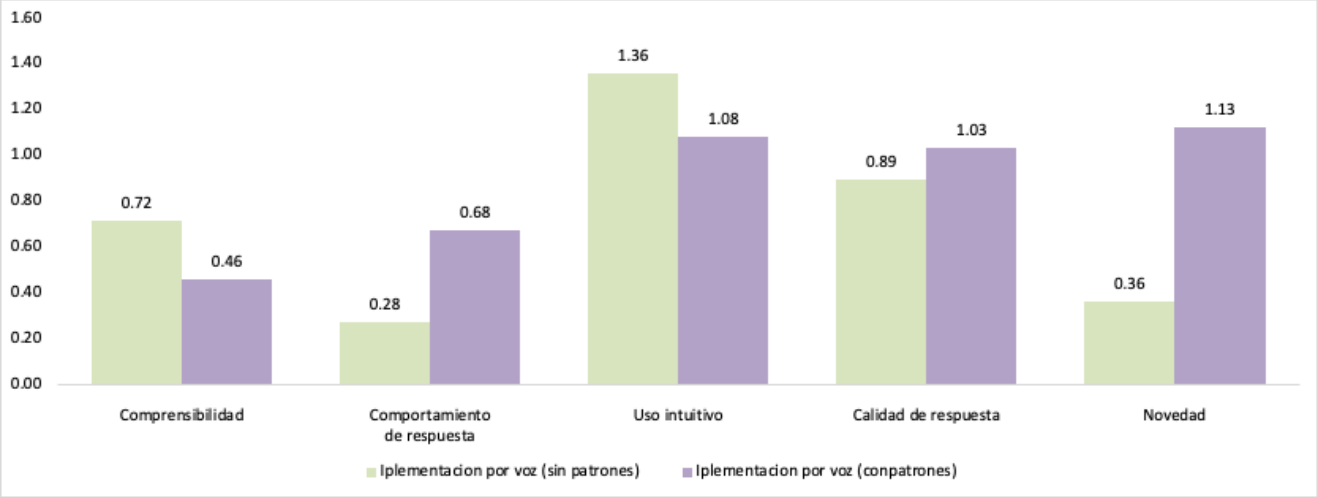


Figura 15. Comportamiento de las medias de las evaluaciones UEQ+ para ambos casos de estudio.

Podemos observar como la incorporación de patrones conversacionales mejoraron las escalas Comportamiento de Respuesta, Calidad de Respuesta y Novedad. Dentro de Comportamiento de Respuesta encontramos evaluaciones referentes a si la experiencia es natural, placentera y entretenida, por lo que es muy valioso el hecho de que el mecanismo por voz con patrones haya superado significativamente a su contraparte sin patrones. En la misma línea, en la escala Novedad se marca significativamente una superioridad por parte del mecanismo con patrones, en conceptos con inventivo y creativo.

Por otro lado, las escalas de Comprensibilidad y Uso Intuitivo se vieron reducidas al incorporar patrones conversacionales. En Uso Intuitivo esta diferencia es esperable, dado que el mecanismo por voz con patrones conversacionales es más complejo y tiene implementado diferentes patrones que no son obvios, mientras que el mecanismo sin patrones es bastante directo y fácil de entender desde la primera vez. De la misma forma, en Comprensibilidad, el mecanismo sin patrones es más sencillo y fácil de comprender que el mecanismo que implementa patrones conversacionales, por lo que no sorprende la mayor calificación del cuestionario por voz sin patrones.

Inconsistencias	Total	Respuestas críticas
VoiceFlow Sin Patrones	60 / 240 (25%)	7 / 40 (18%)
VoiceFlow Con Patrones	18 / 240 (7.5%)	0 / 40 (0%)

Tabla 8. Inconsistencias de los datos por Caso.

Respecto a las inconsistencias presentadas en cada uno de los 2 casos de estudio ejecutados, en la Tabla 8 se presentan el total de estas y además cuantas fueron críticas. Se considera una inconsistencia en cualquiera de las 6 escalas si no coinciden las respuestas de un participante en esa escala. Se considera una respuesta crítica a la respuesta de un participante si entre las 6 escalas hay 3 o más inconsistencias. El número 40, en la columna Respuestas críticas, representa las 40 respuestas, una por cada participante, mientras que 240 representa el producto de 40 participantes por 6 escalas.

Inconsistencias	Atracción	Transparencia	Eficiencia	Contro-labilidad	Estimulación	Novedad
VoiceFlow Sin Patrones	10 / 60 (17%)	10 / 60 (17%)	10 / 60 (17%)	13 / 60 (22%)	8 / 60 (13%)	9 / 60 (15%)
VoiceFlow Con Patrones	3 / 18 (17%)	3 / 18 (17%)	2 / 18 (11%)	9 / 18 (50%)	1 / 18 (5%)	0 / 18 (0%)

Tabla 9. Inconsistencias de los datos por Caso y Escala.

Finalmente, en la Tabla 9 se presentan las inconsistencias por escala para cada uno de las dos implementaciones del mecanismo de captura por voz.

Capítulo 6 - Conclusiones y trabajo futuro

En este trabajo de investigación se propuso implementar el mecanismo de recolección de respuestas del cuestionario User Experience Questionnaire (UEQ), por medio de interfaces de voz. Este cuestionario estandarizado de evaluación de experiencia de usuario (UX) utiliza normalmente el formato escrito para recolectar las respuestas de los participantes, que seleccionan de 26 preguntas presentadas en forma de diferencial semántico, y que corresponden a seis escalas de experiencia de usuario: Atractivo, Transparencia, Eficiencia, Controlabilidad, Estimulación y Novedad.

Se implementaron y evaluaron dos versiones del mecanismo de captura por voz. En una primera versión se diseñó cada diferencial semántico del cuestionario (formado por dos conceptos opuestos, uno presentado a la izquierda y otro a la derecha) como una pregunta que reflejara de forma directa, usando la siguiente forma: “En la escala de 1 a 7 donde 1 significa {concepto izquierdo} y 7 significa {concepto derecho}, ¿cómo calificaría este producto?”. Le participante indica su respuesta por medio de un número.

En una segunda implementación, se transformó el diferencial semántico en dos preguntas: una inicial donde se determina la dirección de la actitud o sentimiento (“¿Usted califica este producto como {concepto izquierdo}, {concepto derecho}, o ni uno ni otro?”), y una pregunta complementaria para definir la intensidad de la actitud (“¿Y qué tan {concepto izquierdo}/{concepto derecho}: un poco, mucho o en extremo?”). El participante entonces debe entonces indicar mediante palabras o frases las respuestas a estas dos preguntas, y no solamente un número, como en la primera implementación. Adicionalmente, en esta segunda versión del mecanismo de captura por voz, también se implementaron varios patrones conversacionales con el fin de que la interacción y el participante fuera más parecida a una conversación.

Cada una de estas implementaciones se probó con 40 participantes (80 en total) que analizaron la experiencia de usuario de dos productos distintos, utilizando tanto el mecanismo usual escrito que provee el UEQ, como el nuevo mecanismo implementado por voz.

A partir de estos casos de estudio, el resultado más importante a resaltar es que no se presentaron diferencias significativas en la evaluación de los dos productos mencionados, tanto si se utiliza el mecanismo de captura tradicional escrito como si el cuestionario es presentado con el mecanismo por voz propuesto. En las seis escalas del cuestionario UEQ, los resultados para ambos mecanismos mostraron resultados similares. Sí vale la pena mencionar, que en la primera implementación del mecanismo por voz (que llamaremos "directa"), este resultado se obtuvo con 40 participantes, pero no con un grupo inicial de 20 participantes. En la segunda implementación del mecanismo, que utiliza patrones conversacionales, la similitud de resultados en la evaluación se obtuvo desde los primeros 20 participantes.

Por otro lado, en la implementación por voz "directa" se detectó un número mayor de inconsistencias (incongruencias en las respuestas de las preguntas de una misma escala) que en la versión por voz con patrones conversacionales, atribuible al hecho de que en la versión directa el participante solamente tiene que responder con números (que puede emitir sin pensar realmente en lo que se le está preguntando), mientras que en la versión con patrones, al tener que responder con palabras específicas que califican la actitud y el sentimiento, el participante se ve obligado a hacer un esfuerzo mayor para entender la pregunta y contestar de forma acorde.

En cuanto a la experiencia de usuario reportada al utilizar cada una de estas dos implementaciones por voz, los resultados obtenidos con el cuestionario UEQ+ muestran que la versión de mecanismo "directo" obtendría puntajes mayores en las escalas de Comprensibilidad (simple, inequívoco, exacto, explicable) y Uso Intuitivo (fácil, lógico, evidente, coherente) para los participantes, lo que es esperable dado que la implementación es sencilla y se apega más al cuestionario escrito que la versión con patrones conversacionales, que se presenta más compleja y con funcionalidades no tan evidentes al usuario. A su vez, la versión con patrones conversacionales toma ventaja en las escalas de Comportamiento de la Respuesta (natural, agradable, simpático, entretenido), Calidad de la Respuesta (adecuado, útil, provee ayuda, inteligente) y Novedad (creativo, original, novedoso, innovador).

En la misma línea de la evaluación UX de ambos mecanismos por voz, las preguntas respecto de la Usabilidad de ambos mecanismos muestran que la versión “directa” obtiene mayor puntaje en la pregunta “La forma de interactuar con el Asistente es clara desde la primera vez”, que coincide con los resultados antes mencionados de la escala Uso Intuitivo del UEQ+.

Otra pregunta donde la implementación “directa” obtiene calificación mayor es en el ítem: “Era común que, a la hora de responder, se me hubiera olvidado qué era lo que el Asistente me había preguntado”. En este caso, la puntuación mayor refleja una deficiencia del mecanismo “directo”. Con el mecanismo con patrones conversacionales, que parte cada diferencial semántico en dos preguntas (dirección de la actitud e intensidad), los participantes indican que la pregunta se les olvida significativamente menos, lo que es positivo a la hora de llenar el cuestionario.

Otras preguntas en las que la implementación con patrones obtiene mejores puntajes de Usabilidad son: “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”, “Si me equivoqué al dar una respuesta, es posible corregirla en el momento” y “Es sencillo solicitar ayuda al Asistente de voz”.

Podemos concluir entonces, que una implementación por voz del mecanismo de captura de un cuestionario estandarizado de evaluación UX UEQ es posible, dado que los resultados obtenidos en las evaluaciones UX de productos no presentan diferencias significativas respecto de la versión del cuestionario UEQ con recolección de respuestas por escrito. En este sentido, la versión del mecanismo de voz con patrones conversacionales presenta ventajas dado que refleja una evaluación equivalente al mecanismo escrito utilizando una muestra menor de participantes. Se respondería entonces la pregunta de investigación planteada al principio de la investigación, a la que respondemos que es posible implementar mecanismos de captura por voz y contar con alternativas válidas para aplicar cuestionarios estandarizados de evaluación UX.

En cuanto a la Usabilidad y experiencia de usuario, es claro que la implementación por voz “directa” es más intuitiva y sencilla que la que presenta el diferencial semántico en dos preguntas junto y patrones conversacionales, mientras que esta última es vista como más

agradable, natural y creativa. Dado también el hecho de que la versión directa presenta más inconsistencias entre preguntas de una misma escala y requiere más participantes para coincidir con los resultados de la versión escrita tradicional, consideramos que las futuras versiones del mecanismo por voz deben seguir el formato de la segunda implementación por voz, utilizando los patrones conversacionales y la transformación del diferencial semántico en dos preguntas complementarias.

Como trabajo futuro, queda pendiente comparar la experiencia de usuario (UX) y Usabilidad del cuestionario que utiliza un mecanismo de captura por voz con el cuestionario tradicional escrito. En esta investigación solamente se compararon las dos versiones del mecanismo por voz, pero es importante también determinar cómo califican los participantes la experiencia de usuario de llenar el cuestionario estandarizado escrito (sobre lo que no se ha encontrado documentación) y cómo se compara esta experiencia a llenar un cuestionario cuyo mecanismo de entrada es implementado por voz.

Adicionalmente, y también como trabajo futuro, es posible identificar nuevos patrones conversacionales que puedan ser implementados y evaluados con el objetivo de seguir mejorando la Usabilidad y UX de un mecanismo por voz. Por ejemplo, la capacidad de explicar los conceptos podría ampliarse para que existan respuestas sucesivas que extiendan la explicación dada según el participante solicite mayores explicaciones.

Otro caso de estudio a realizar sería implementar el mecanismo por voz a la versión reducida del UEQ, que tiene solamente ocho preguntas, dado que a extensión del cuestionario es una de las características que los participantes consideran como negativa. Si, al igual que en este trabajo de investigación, los resultados de usar un mecanismo por voz y uno escrito en un cuestionario no muestran diferencias significativas, se tendría entonces también una alternativa para el cuestionario en su versión reducida. Es claro que, como lo indican los propios creadores de UEQ (en sus versiones completa y reducida) la aplicación de la versión completa del cuestionar es siempre preferible.

Referencias

1. ISO, ISO DIS 9241-210:2010 Ergonomics of Human-System Interaction - Part 210: Human- centred design for interactive systems, International Standardization Organization (ISO), Geneva, Switzerland, 2010.
2. Vermeeren, A.P.O.S.; Law, E.L.-C.; Roto, V.; Obrist, M.; Hoonhout, J.; Väänänen-Vainio- Mattila, K. User experience evaluation methods. InProceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10, Reykjavik, Iceland, 16–20 October 2010.
3. Schrepp, M.; Hinderks, A.; Thomaschewski, J. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In Lecture Notesin Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); SpringerInternational Publishing: Crete, Greece, 2014.
4. Wallach, D.; Conrad, J.; Steimle, T. The UX metrics table: A missing artifact. In Lecture Notes in Computer Science (Including Subseries LectureNotes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer International Publishing: Vancouver, BC, Canada, 2018.
5. Lallemand, C.; Gronier, G. Méthodes de design UX. 30 méthodes fondamentales pour concevoir des expériences optimales. 2eme édition; ÉditionsEyrolles: Paris, France, 2019.
6. Maia, C. L. B., & Furtado, E. S. A Systematic Review About User Experience Evaluation. Lecture Notes in Computer Science, 2016.
7. Ten, A.C.; Paz, F. A systematic review of user experience evaluation methods in information driven websites. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer International Publishing: Vancouver, BC, Canada, 2018.
8. Gronier, G.; Lallemand, C.; Chauvet, A. Mesurer la formation de la première impression d'une interface à l'aide du test des 5 secondes; HuitièmeColloque de Psychologie Ergonomique (EPIQUE), Aix-en-Provence, France, July 8-10. 2015.
9. Kortum, P. HCI beyond GUI: Design for haptic, speech, olfactory and other nontraditional interfaces, 1st ed.; Elsevier/Morgan Kaufmann PublishersInc.: California, United States, 2008.
10. Karray, F.; Alemzadeh, M.; Abou Saleh, J.; Nours Arab, M. Human-Computer Interaction: Overview on State of the Art. International Journal onSmart Sensing and Intelligent Systems. 2008.
11. López G., Quesada L., Guerrero L.A. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In: Nunes I. (eds) Advances in HumanFactors and Systems Interaction. AHFE 2018. Advances in Intelligent Systems and Computing.Springer, Cham. 2019.
12. Van Beurden, M. H., Ijsselsteijn, W. A., & de Kort, Y. A. User experience of gesture-based interfaces: a comparison with traditional interaction methods on pragmatic and hedonic qualities. In International Gesture Workshop. Springer, Berlin, Heidelberg. 2011.

13. Pradhan, A., Mehta, K. & Findlater, L. Accessibility Came by Accident: Use of Voice-Controlled Intelligent Personal Assistants by People withDisabilities. In Proceedings of the 219 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery.2019.
14. Baumgartner, J.; Sonderegger, A.; Sauer, J. No need to read: Developing a pictorial single- item scale for measuring perceived usability. Int. J. Hum. Comput. Stud. 2020.
15. Bradley, M.; Lang, P. Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry. 1994.
16. Diaz-Oreiro, I.; López, G.; Quesada, L.; Guerrero, L.A. UX Evaluation with Standardized Questionnaires in Ubiquitous Computing and Ambient Intelligence: A Systematic Literature Review. Advances in Human-Computer Interaction, 2022.
17. Moore, R.J., Arar, R. Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework. Association for Computing Machinery, New York, NY, USA. 2020.
18. Chien, Y.-H., Chang, W.-T., Chuang, C.-C., & Chen, S.-H. A Taiwanese User Experience Questionnaire. Advances in Intelligent Systems and Computing, 347–355. doi:10.1007 978-3-319-41685-4_31. 2016.
19. Lallemand, C., Koenig, V., Gronier, G., Martin, R. Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmesinteractifs. Revue Européenne de Psychologie Appliquée /European Review of Applied Psychology, 65(5), 239-252. doi: 10.1016/j.erap.2015.08.002. 2015.
20. Vallerand, R. J. Vers une méthodologie de validation transculturelle de questionnairespsychologiques : Implications pour la recherche en langue française. Psychologie Canadienne,30(4), 662–689. 1989.
21. Bernhaupt, R., & Pirker, M. Evaluating User Experience for Interactive Television: Towards the Development of a Domain-Specific User Experience Questionnaire. Lecture Notes in Computer Science, 642–659. doi:10.1007/978-3-642-4290-1_45. 2013.
22. Tomlinson, B., Noah, B. and Walker, B. BUZZ: An Auditory Interface User Experience Scale. CHI Conference on Human Factors in Computing Systems. Paper No. LBW096. doi: 10.1145/317497.3188659. 2018.
23. Sauro, J. and Zarolia, P. SUPR-Qm: A Questionnaire to Measure the Mobile App UserExperience. Journal of Usability Studies, 13(1), 17-37. 2017.
24. Petrov, C.; 49 Voice Search Stats to Help You Rethink Your Strategy in 2020. August, 2020. Taken from: https://techjury.net/blog/voice-search-stats/#gref
25. Andersen, D. 26 Voice Search Stats Marketers Need to Know in 2020. DialogTech. October, 2019. Taken from: https://www.dialogtech.com/blog/voice-search-statistics/
26. Lin, Y. 10 Voice Search Statistics You Need to Know in 2020. Oberlo. August, 2020. Taken from: https://www.oberlo.com/blog/voice-search-statistics
27. Borgia, E. The Internet of Things vision: Key features, applications, and open issues. Computer Communications, 2014.

28. Yang, X., Aurisicchio, M., & Baxter, W. Understanding affective experiences withconversational agents. In proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-12). 2019, May.
29. Berdasco, A., López, G., Diaz, I., Quesada, L., & Guerrero, L. A. 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). 2019.
30. López, G., Quesada, L., & Guerrero, L. A. 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. 2017, July.
31. Pradhan, A., Mehta, K., & Findlater, L. " 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). 2018, April.
32. Quamar, A., Özcan F., Miller, D., Moore, R., Niehus, R., Kreulen, J. Conversational BI: an ontology-driven conversation system for business intelligence applications. Proc. VLDB Endow.13, 12, 3369–3381. 2020.
33. Liao Q.V., Geyer W., Muller M., Khazaen Y. Conversational Interfaces for Information Search. Understanding and Improving Information Search. Springer, Cham. 2020.
34. Vepsäläinen, H., Salovaara, A. & Paakki, H. What Would Be the Principles for Successful Trollbot Design? ACM CHI Conference workshop. 2021.
35. Schrepp, M., & Thomaschewski, J. Design and validation of a framework for the creation of user experience questionnaires. International Journal of Interactive Multimedia andArtificial Intelligence. 2019.
36. Chaudhri, V. K., Cheyer, A., Guilii, R., Jarrold, W., Myers, K., & Niekarsz, J. A Case Study in Engineering a Knowledge Base for an Intelligent Personal Assistant. In SemDesk. 2006, November.
37. Roto, V., Obrist, M. and Väänänen-Vainio-Mattila. K. (2009). User Experience Evaluation Methods in Academic and Industrial Contexts. Workshop in Interact'09 conference, Uppsala, Sweden, August 25th, 2009.
38. Hassenzahl, M. The thing and I: understanding the relationship between user and product, in Funology : From Usability to Enjoyment, M. A. Blyth, Ed., pp. 1–12, Kluwer Academic Publishers, New York, NY, USA, 2003.
39. Laugwitz, B.; Held, T.; Schrepp, M. Construction and evaluation of a user experience questionnaire, Computer Science, vol. 5298, pp. 63–76, 2008.
40. Minge, M.; Riedel, L. meCUE – Ein modularer Fragebogen zur Erfassung des Nutzungserlebens [meCUE - A modular questionnaire for capturing the user experience]. Mensch und Comput. 2013, 9, 89–98.
41. Hassenzahl, M. The effect of perceived hedonic quality on product appealingness, InternationalJournal of Human-Computer Interaction, vol. 13, pp. 481–499, 2001.

42. Klammer, J.; van den Anker, F.W.G. Design, User Experience, and Usability: Users, Contextsand Case Studies, Springer International Publishing, Berlin, Heidelberg, Germany, 2018.
43. Roto, V.; Law, E.; Vermeeren, A.; Hoonhout, J. User experience white paper: bringing clarity tothe concept of user experience, in Proceedings of the Result from Dagstuhl Seminar on Demarcating User Experience, Helsinki, Finland, September 2011.
44. Thüring, M. Mahlke, S. Usability, aesthetics and emotions in human- technologyinteraction, International Journal of Psychology, vol. 42, no. 4, pp. 253–264, 2007.
45. Osgood, E.C. The Nature and measurement of meaning. Psychol. Bull. 49, 197–237. 1952.
46. Likert, R. A technique for measurement of attitudes. Archives of Psychology, 140, 5-55. 1932.
47. Klein, A., Hinderks, A., Schrepp, M., & Thomaschewski, Jörg. Construction of UEQ+ Scales forVoice Quality. 2020.
48. Chaudhri, V. K., Cheyer, A., Guilii, R., Jarrold, W., Myers, K., & Niekarsz, J. A Case Study in Engineering a Knowledge Base for an Intelligent Personal Assistant. In SemDesk. 2006, November.
49. Kaushik, D., & Jain, R. (2014). Natural user interfaces: Trend in virtual interaction.
50. Russell, S. & Norvig, P. (2009). Artificial intelligence: a modern approach.
51. Sacks, H., Schegloff, E., Jefferson, G. A Simplest Systematics for the Organization of Turn- Taking for Conversation. Language, 50 (4), 696-735. 1974.
52. Vaishnavi, V.K. and Kuechler Jr., W., Design science research methods and patterns: Innovating Information and Communication Technology 2nd ed., Florida, UAS: CRC Press. 2015.
53. Wieringa, R. J. Design Science Methodology for Information Systems and Software Engineering. the 32nd ACM/IEEE International Conference. Springer. 2014.
54. Mark Saunders and Paul Tosey. 2013. The layers of research design. Rapport 30 2013, 58- 59.
55. Chacón Chaves, M. Diseño de un cuestionario de evaluación de experiencia de usuario por medio de un asistente inteligente por voz. Trabajo Final de Investigación Aplicada. Posgrado en Computación e Informática. Universidad de Costa Rica. 2021
56. Díaz-Oreiro, I., López, G., Quesada, L., Guerrero, L.A. Conversational Design Patterns for a UX Evaluation Instrument Implemented by Voice. In: Rocha, Á., Ferrás, C., Méndez Porras, A., Jimenez Delgado, E. (eds) Information Technology and Systems. ICITS 2022. Lecture Notes in Networks and Systems, vol 414. Springer, Cham. 2022.
57. Yánez, G; Behar, R. Interpretaciones erradas del nivel de confianza en los intervalos de confianza y algunas explicaciones plausibles, 2001.

Anexos

Anexo 1. Artículo publicado

Este anexo incluye el texto completo del artículo publicado en la conferencia ICITS` 2022 – *The 2022 International Conference on Information Technology & Systems*, como parte de los libros de la serie *Lecture Notes in Networks and Systems de Springer. Comparing Written and Voice Captured Responses of the User Experience Questionnaire (UEQ)".* Mata-Serrano, Díaz-Oreiro, López y Guerrero. DOI 10.1007/978-3-030-96293-7_43 2.

Comparing Written and Voice Captured Responses of the User Experience Questionnaire (UEQ)

Jean Carlo Mata-Serrano(✉), Ignacio Díaz-Oreiro, Gustavo López, and Luis A. Guerrero

University of Costa Rica, San José 11501-2060, Costa Rica
{jean.mata,ignacio.diazoreiro,gustavo.lopezherrera,
luis.guerreroblanco}@ucr.ac.cr

Abstract. Standardized questionnaires are widely used instruments to evaluate UX and their capture mechanism has been implemented in written form, either on paper or in digital format. This study aims to determine if the UX evaluations obtained in the standardized UEQ questionnaire (User Experience Questionnaire) are equivalent if the response capture mechanism is implemented using the traditional written form (digitally) or if a conversational voice interface is used. Having a UX evaluation questionnaire whose capture mechanism is implemented by voice could provide an alternative to collect user responses, preserving the advantages present in standardized questionnaires (quantitative results, statistically validated, self-reported by users) and adding the ease of use and growing adoption of conversational voice interfaces. The results of the case study described in this paper show that, with an adequate number of participants, there are no significant differences in the results of the six scales that make up UEQ when using either of the two response capture mechanisms.

Keywords: Human-computer interaction · HCI · User experience · UX evaluation · Conversational interfaces · Voice interfaces · Standardized questionnaires · UEQ

1 Introduction

ISO defines user experience (UX) as a person's perceptions and responses resulting from the use and/or anticipated use of a product, system, or service [1]. This definition of UX includes users' emotions, beliefs, physical and psychological responses and is also influenced by brand image, product presentation, system performance, user's prior experiences, attitudes, skills, and personality, among others.

To evaluate UX, researchers rely on different methods and instruments such as expert evaluation, ethnographic studies, interviews, tailored or standardized questionnaires, to name a few of the most widely applied [2, 3]. The use of standardized questionnaires is extensive as they provide quantitative scores which can be compared with benchmarks or scores of other evaluations.

Á. Rocha et al. (Eds.): ICITS 2022, LNNS 414, pp. 519–529, 2022.
https://doi.org/10.1007/978-3-030-96293-7_43

Moreover, fewer resources are needed to administer the questionnaires and collect the data since no effort should be invested in designing them, as the set of questions is invariable, and the experience is reported by themselves [3]. Finally, the questionnaires have been statistically validated, so they are considered reliable and valid for measuring UX [4].

Standardized questionnaires are sometimes considered boring to fill out since all the questions follow the same pattern (semantic differential or Likert scales). Each questionnaire is composed of a considerable number of items, which raises the question of whether there are alternative ways of capturing the responses.

In recent years various non-traditional interfaces have been developed to exploit alternative ways of human-computer interaction, such as gaze or gesture detection and tracking, haptic and tangible interfaces, and voice or conversational interfaces [5]. Voice or conversational interfaces use natural language that resembles a conversation [6], they have been implemented with great success in different environments, such as intelligent voice assistants. This success is due both to the ease of use, as well as the continuous substantial improvements they experience in speech pattern recognition or the expansion into languages other than English [7].

We are interested in exploring the use of conversational voice interfaces as capture mechanisms of a standardized UX evaluation questionnaire, seeking more natural user interactions while filling the questionnaire and taking advantage of an automated self-reported process.

In this paper, we decided to work with User Experience Questionnaire (UEQ), which is based on the UX model proposed by Hassenzahl [8] and consists of 26 items belonging to subscales attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty.

Therefore, we propose an implementation of the UEQ in which the capture of user responses is carried out by a voice mechanism, and we present the results obtained from this implementation compared to those taken from the same UEQ questionnaire whose capture method is a traditional written interface.

It is important to mention that different efforts have been made to create other standardized evaluation mechanisms that modify the presentation format of questions, but they mainly focus on pictorial alternatives [9, 10]. To our knowledge, there are no efforts to implement standardized UX evaluation questionnaires through conversational interfaces. This conclusion was obtained as part of a systematic literature review [11] developed by the researchers on the use of standardized questionnaires in primary studies, in which, in a first initial filter, variants to the questionnaires or other evaluation mechanisms were identified.

2 Related Work

Researchers have proposed modifications to the standardized questionnaires, such as adapting the questionnaire to a different culture [12] or language [13], or specific questionnaires for specific domains, such as measuring UX of interactive television [14] or auditory interfaces [15]. Other researchers have presented UX evaluation instruments whose format for presenting the questions is pictorial [9, 10].

As for intelligent voice assistants, there are studies on comparing different virtual assistants performing the same tasks, as well as reviewing security and privacy issues [16]. Other studies focus on evaluating UX from an emotional point of view when using these devices [17], while others on how correct or good is the answer given by an intelligent voice assistant [18], or on assessing the naturalness of these assistants' answers [6]. For its part, [7] shows how intelligent voice assistants help reduce the technological gap that may exist between users with some type of disability or physical impairment, which restricts or prevents them from using products or devices.

Currently, survey can be programmed in virtual assistants such as Amazon Alexa through skills, which are additional functionalities that could be installed in the smart assistant. There is also a private web service [19], which offers the ability to create and administer surveys using the smart voice assistants Alexa and Google Assistant. These surveys are freely defined with the questions that the user deems appropriate.

However, to our best understanding, there are not UX evaluation instruments implemented through conversational voice interfaces, nor studies on evaluating the equivalence of capturing the UX evaluation using different formats, as no references to studies of this type were found in the academic literature, and the available commercial solutions do not present evidence to support that capturing user information conversationally is equivalent to capturing the same information in written form.

3 Methodology

To design an instrument that uses a conversational voice interface, it was decided to take advantage of the statistical support of one of the 3 best-known standardized questionnaires, and only replace the response capture mechanism. It was decided to use UEQ, given that it is the one with the most uses in recent years [11], as well as the fact that the tools for analyzing the results are freely distributed.

The UEQ questionnaire was implemented in two different tools with two different capture formats: a written format using Google Forms, and a voice interface implemented in VoiceFlow [20], a tool to develop interactive conversations that can be exported to smart voice assistants or accessed through a web interface.

The conversational version that simulates the UEQ questions was designed in a simple way aiming to emulate the semantic differential in the most similar way possible but using a spoken question. Each UEQ question was asked as follows: "on a scale of 1 to 7, where 1 means {left concept} and 7 means {right concept}, how would you rate this product?", using the left and right concepts that comprise each UEQ question. For example: on a scale of 1 to 7, where 1 means "annoying" and 7 means "enjoyable", how would you rate this product?

The Voice Assistant was instructed to recognize numbers "one" to "seven" as valid answers, and the equivalent utterances "first" to "seventh". This allows the Assistant to recognize utterances like "first option" as an equivalent to the answer "one".

The assistant was also programmed so that, if it does not recognize what the participant said, asks the participant to repeat the chosen option. As a counterpart, the participant can request the Assistant to repeat a question. Below is an example of a conversation between the Assistant and the participant:

Assistant: Question 1. On a scale of 1 to 7, where 1 means "annoying" and 7 means "enjoyable", how would you rate this product?
User: Four
Assistant: Question 2. On a scale of 1 to 7, where 1 means "not understandable" and 7 means "understandable", how would you rate this product?
User: Eight
Assistant: Excuse me, I didn't understand you. You must enter a number between 1 and 7. What number did you say?
User: Seven
Assistant: Question 3. On a scale of 1 to 7, where 1 means "creative" and 7 means "dull", how would you rate this product?
User: Repeat question
Assistant: On a scale of 1 to 7, where 1 means "creative" and 7 means "dull", how would you rate this product?
User: Two

Two products were selected in the form of videos, so participants could evaluate the anticipated UX before using them. The two products were chosen as they feature innovative characteristics: a pair of sneakers that project images and transmit haptic impulses, and a wallet that charge electronic devices via USB and has an integrated GPS.

A total of 40 participants took part in the study, all university students with basic knowledge on HCI, with ages ranging from 19 to 29 years, with a median of 21 years. Six of them were women (15%) and 34 men (85%).

Table 1. Participant's distribution for product UX evaluation

	Group 1	Group 2	Group 3	Group 4
First evaluation Product/Data capture method	Wallet/Google Forms	Wallet/VoiceFlow	Sneakers/Google Forms	Sneakers/VoiceFlow
Second evaluation Product/Data capture method	Sneakers/VoiceFlow	Sneakers/Google Forms	Wallet/VoiceFlow	Wallet/Google Forms

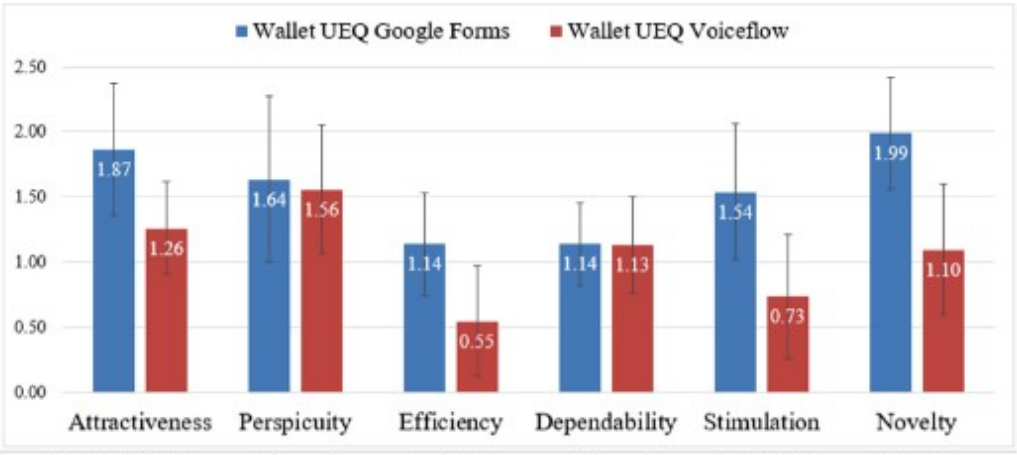
Evaluations were conducted in two sessions, with 20 participants organized into four groups of five people each, using the configuration shown in Table 1. The goal of setting up this configuration was to ensure that the same mechanism was not always used first, since the experience of the second would be influenced by the performance of the first mechanism.

Data collected through the two interfaces (written and voice) were evaluated using the comparison tool provided by UEQ. Additionally, an evaluation of the conversational Assistant was conducted using the UEQ + questionnaire, formed with the three voice quality scales and the Intuitive Use and Novelty scales. A set of 10 Usability questions was also employed, presented as five-point Likert scale items.

4 Results

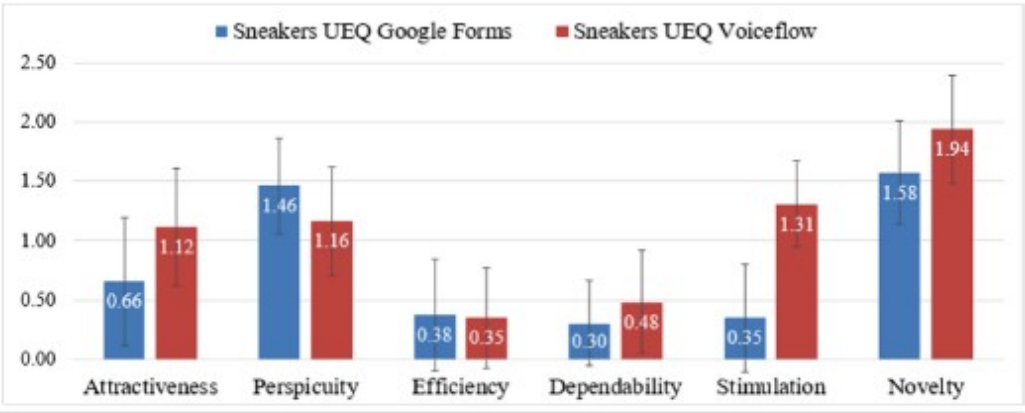
Initially the data was classified in four groups combining product and capture mechanism. These groups made up of 20 participants each correspond to Wallet/Google Forms, Wallet/Voiceflow, Sneakers/Google Forms, and Sneakers/Voiceflow.

Figure 1 shows the means, standard deviation and corresponding 5.0% confidence for each UEQ scale obtained from the Google Forms and Voiceflow capture formats for the Wallet product.



	Attractiveness		Perspicuity		Efficiency		Dependability		Stimulation		Novelty	
Mean	1.87	1.26	1.64	1.56	1.14	0.55	1.14	1.13	1.54	0.73	1.99	1.10
Std. Dev.	1.15	0.82	1.46	1.13	0.91	0.97	0.73	0.85	1.19	1.09	0.97	1.14
Confidence	0.51	0.36	0.64	0.49	0.40	0.42	0.32	0.37	0.52	0.48	0.43	0.50

Fig. 1. Wallet UX evaluation compared by capture method (N = 20)



	Attractiveness		Perspicuity		Efficiency		Dependability		Stimulation		Novelty	
Mean	0.66	1.12	1.46	1.16	0.38	0.35	0.30	0.48	0.35	1.31	1.58	1.94
Std. Dev.	1.23	1.12	0.92	1.04	1.07	0.97	0.82	0.98	1.04	0.83	0.99	1.04
Confidence	0.54	0.49	0.40	0.46	0.47	0.42	0.36	0.43	0.46	0.36	0.44	0.46

Fig. 2. Sneakers UX evaluation compared by capture method (N = 20)

Similarly, Fig. 2 presents the means, standard deviation, and confidence results for the Sneakers product.

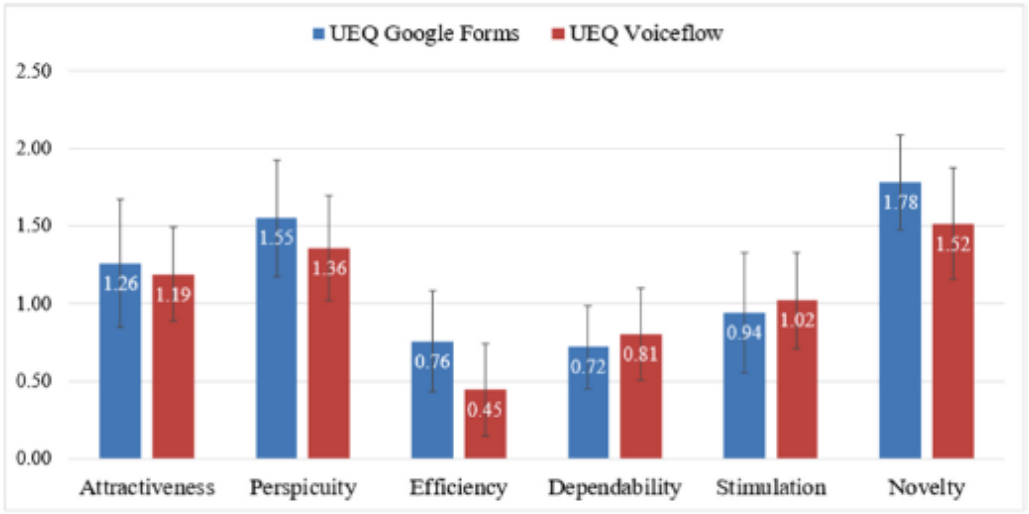
A Student's t-test was performed on each scale comparing capture methods, aiming to find whether existed significant differences between written and voice mechanisms. As it can be seen in Table 2, for the Wallet product, in four of the six UEQ scales the capture method does not present a significant difference. As for the Sneakers, five out of six scales do not present significant differences either.

Table 2. Student's t-test for data capture method for Wallet and Sneakers products (N = 20)

UEQ Scale	Wallet t-value	SignificantDifference	Sneakers t-value	Significant Difference
Attractiveness	0.0635	No	0.2257	No
Perspicuity	0.8490	No	0.3397	No
Efficiency	0.0536	No	0.9386	No
Dependability	0.9736	No	0.5264	No
Stimulation	0.0319	Yes	0.0027	Yes
Novelty	0.0114	Yes	0.2679	No

Alpha level: 0.05.

Finally joining the data for both products and comparing only the capture method the means, standard deviation and confidence results are displayed in Fig. 3.



	Attractiveness		Perspicuity		Efficiency		Dependability		Stimulation		Novelty	
Mean	1.26	1.19	1.55	1.36	0.76	0.45	0.72	0.81	0.94	1.02	1.78	1.52
Std. Dev.	1.33	0.97	1.21	1.09	1.05	0.96	0.88	0.96	1.26	1.00	0.99	1.16
Confidence	0.41	0.30	0.37	0.34	0.33	0.30	0.27	0.30	0.39	0.31	0.31	0.36

Fig. 3. UX evaluation compared by capture method (N = 40)

In this case, having a large enough sample (N = 40), Table 3 presents the Student's t-test shows that there are no significant differences in the results of none the six UEQ scales, when comparing by capture method only.

Table 3. Student's t-test for data capture method on each UEQ scale (N = 40)

UEQ Scale	t-value	Significant Difference
Attractiveness	0.7763	No
Perspicuity	0.4632	No
Efficiency	0.1751	No
Dependability	0.6722	No
Stimulation	0.7623	No
Novelty	0.2766	No

Alpha level: 0.05

An important element to mention is the number of inconsistencies found using each capture format. UEQ processing spreadsheet warns that it may happen that not all participants answer all items seriously and, to detect such random or not serious answers, the UEQ spreadsheet uses a simple heuristic. All items in a scale should measure a similar UX quality aspect, so the spreadsheet checks how much the best and worst evaluation of an item in a scale differs. If there is a big difference (greater than three) this is seen as an indicator for a problematic data pattern. UEQ considers answers from participants with three or more inconsistencies as Critical and suggests removing them from the analysis.

Table 4 shows the total number of inconsistencies detected by capture format as well as the number of critical responses which are those responses from a participant with three or more scales marked as inconsistent. It can be observed that there are more inconsistencies in the questionnaires using the conversational format implemented in VoiceFlow either for total inconsistencies or for responses considered critical.

Table 4. Inconsistencies detected by capture method (N = 40)

Format	Inconsistencies	Critical answers
Google Forms	45/240 (18.75%)	4/40 (10.00%)
VoiceFlow	60/240 (25.00%)	7/40 (17.50%)

As indicated in the methodology section, UEQ questionnaire implemented with the conversational interface was evaluated with the UEQ + questionnaire, formed with the scales Comprehensibility, Response Behavior, and Response Quality developed for voice interfaces along with Intuitive Use, and Novelty.

Table 5 shows that the means for each scale, which vary between -3 and 3, that this UEQ implementation is not very good in terms of UX evaluation. The best rated scale

is Intuitive Use, followed by Response Quality. The individual means of each scale item indicate that participants consider the questionnaire as suitable, useful, easy, logical, conclusive, explainable, and plausible. At the same time, the worst rated items reveal that participants find this conversational implementation as not helpful, boring, artificial, conventional, usual, and conservative.

Table 5. Mean values for scales and items evaluated with UEQ+

UEQ + Scale	Scale mean	Left item	Right item	Item mean
Comprehensibility	0.72	complicated	simple	0.63
		ambiguous	unambiguous	0.45
		inaccurate	accurate	0.53
		enigmatic	explainable	1.28
Response behavior	0.28	artificial	natural	−0.25
		unpleasant	pleasant	1.03
		unlikeable	likeable	0.78
		boring	entertaining	−0.45
Intuitive Use	1.36	difficult	easy	1.68
		illogical	logical	1.38
		not plausible	plausible	1.10
		inconclusive	conclusive	1.28
Response quality	0.89	inappropriate	suitable	2.13
		useless	useful	1.80
		not helpful	helpful	−1.18
		unintelligent	intelligent	0.83
Novelty	0.36	dull	creative	0.88
		conventional	inventing	0.00
		usual	leading edge	0.15
		conservative	innovative	0.43

In addition to the UEQ+, Usability aspects for the conversational UEQ version were assessed by means of a set of 10 questions in Likert scale format. The results obtained, displayed in Fig. 4, show some positive aspects of the Assistant: the interaction is clear from the first question, it is easy to know if the Assistant understood my answer, as well as knowing how many questions have been asked and how many remain to be asked. Among the negative aspects: it is not possible to correct a given answer, it is not easy to ask the Assistant for help, the Assistant speaks more than necessary and if it was common to forget what the Assistant had asked. Some other negative characteristics identified could be attributed to the implementation of the voice data capture mechanism but also

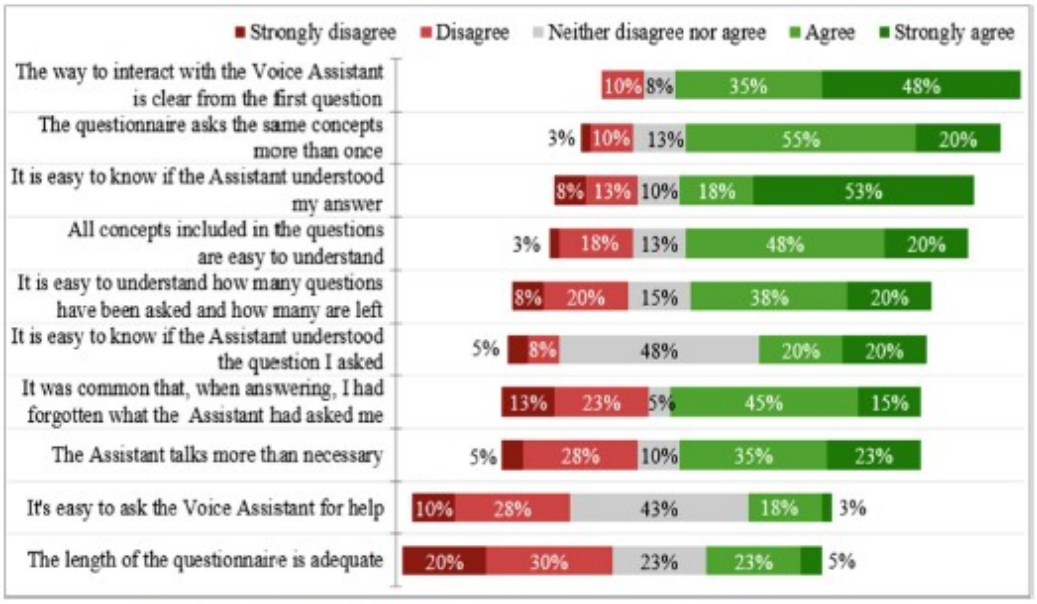


Fig. 4. Usability evaluation for the questionnaire with voice interface

to UEQ inherent characteristics, such as asking for the same concept more than once and having too many questions.

5 Conclusion and Future Work

This article describes the results of a UX evaluation using a standardized questionnaire in two capture formats: written and conversational by voice, allowing to identify the impact of using one or the other.

Results show that there would be no significant difference when using either of the two capture formats to evaluate the anticipated UX of two products, for the six scales of the UEQ questionnaire.

However, there are many opportunities to improve the Usability and UX of the voice Assistant that implements UEQ, which could be addressed in future work. One opportunity would be to apply a different format to formulate the questions in the Assistant so that the response received by the semantic differential can be captured but presented in a more natural conversational format.

Other areas for improvement are to create more friendly and interactive conversation flows where the participant can be aware of the questionnaire progress at any point, have a greater flow control and freedom and prevent errors which could improve the UX and Usability of the conversational interface.

The extension of the questionnaire and quantity of items per scale are factors that could influence the results of the evaluations. These are inherent to UEQ, since the questionnaire is invariably composed of 26 items, but could be perceived as more annoying in a conversational format than in a written one.

Improving the Usability and UX of the questionnaire could also reduce the number of inconsistencies reported by UEQ attributable to the conversational format and maintain the equivalence of the results compared to those obtained by the written method, providing an alternative of conversational capture for the UEQ questionnaire.

Acknowledgments. This research was partially funded by CITIC at the University of Costa Rica, grant number 834-C1-013 and 834-B7-766.

References

1. ISO, ISO DIS 9241-210:2010 Ergonomics of Human-System Interaction, International Standardization Organization (ISO), Geneva, Switzerland (2010)
2. Vermeeren, A., Law, E., Roto, V., Obrist, M., Hoonhout, J., Väänänen, K.: User experience evaluation methods. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries – NordiCHI'10, Reykjavik, Iceland, 16–20 (2010)
3. Lallemand, C., Gronier, G.: Méthodes de design UX. 30 méthodes fondamentales pour concevoir des expériences optimales. 2eme édition; Éditions Eyrolles. Paris, France (2018)
4. Gronier, G., Lallemand, C., Chauvet, A.: Mesurer la formation de la première impression d'une interface à l'aide du test des 5 secondes. In: Huitième Colloque de Psychologie Ergonomique (EPIQUE), Aix-en-Provence, France, pp. 8–10 July (2015)
5. Kortum, P.: HCI beyond GUI: Design for haptic, speech, olfactory and other nontraditional interfaces, 1st edn. Morgan Kaufmann Publishers Inc., California, USA (2008)
6. López, G., Quesada, L., Guerrero, L.A.: Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces. In: Nunes, I. (eds.) Advances in Human Factors and Systems Interaction, AHFE 2017. Springer, Cham (2018)
7. Pradhan, A., Mehta, K., Findlater, L.: 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, ACM (2018)
8. Hassenzahl, M.: The effect of perceived hedonic quality on product appealingness. Int. J. Hum.-Comput. Int. **13**, 481–499 (2001)
9. Baumgartner, J., Sonderegger, A., Sauer, J.: No need to read: developing a pictorial single-item scale for measuring perceived usability. Int. J. Hum. Comput. Stud. **122**, 78–79 (2019)
10. Bradley, M., Lang, P.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)
11. Diaz-Oreiro, I., López, G., Quesada, L., Guerrero, L.A.: UX evaluation with standardized questionnaires in ubiquitous computing and ambient intelligence: a systematic literature review. Adv.Hum.-Comput. Interact. **2021**, 1–22 (2021)
12. Chien, Y.-H., Chang, W.-T., Chuang, C.-C., Chen, S.-H.: A Taiwanese user experience questionnaire. Adv. Intell. Syst. Comput. pp. 347–355 (2016)
13. Lallemand, C., Koenig, V., Gronier, G., Martin, R.: Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs. Eur. Rev. Appl. Psychol. **65**(5), 239–252 (2015)
14. Bernhaupt, R., Pirker, M.: Evaluating user experience for interactive television: towards the development of a domain-specific user experience questionnaire. Lecture Notes in Computer Science, pp. 642–659 (2013)
15. Tomlinson, B., Noah, B., Walker, B.: BUZZ: an auditory interface user experience scale. In: CHI Conference on Human Factors in Computing Systems. Paper No. LBW096 (2018)

16. Hoy, M.B.: Alexa, siri, cortana, and more: an introduction to voice assistants. Med. Ref. Serv. Q. **37**(1), 81–88 (2018)
17. Yang, X., Aurisicchio, M., Baxter, W.: Understanding Affective Experiences with Conversational Agents (2019)
18. Berdasco, A., López, G., Diaz, I., Quesada, L., Guerrero, L.A.: User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana. In: MDPI Proceedings, vol. 31, p. 51 (2019)
19. SurveyLine homepage. https://www.surveybyvoice.com (2021). Accessed 01 Aug 2021
20. VoiceFlow homepage. https://www.voiceflow.com (2021). Accessed 01 Aug 2021

Anexo 2. Cuestionario de evaluación de Experiencia de Usuario (UEQ)

¿Usted calificaría el producto como: "desagradable", "agradable", o ni uno ni otro?
¿Usted calificaría el producto como: "no entendible", "entendible", o ni uno ni otro?
¿Usted calificaría el producto como: "creativo", "sin imaginación", o ni uno ni otro?
¿Usted calificaría el producto como: "fácil de aprender", "difícil de aprender", o ni uno ni otro?
¿Usted calificaría el producto como: "valioso", "de poco valor", o ni uno ni otro?
¿Usted calificaría el producto como: "aburrido", "emocionante", o ni uno ni otro?
¿Usted calificaría el producto como: "no interesante", "interesante", o ni uno ni otro?
¿Usted calificaría el producto como: "impredecible", "predecible", o ni uno ni otro?
¿Usted calificaría el producto como: "rápido", "lento", o ni uno ni otro?
¿Usted calificaría el producto como: "original", "usual", o ni uno ni otro?
¿Usted calificaría el producto como: "obstructivo", "impulsor de apoyo", o ni uno ni otro?
¿Usted calificaría el producto como: "bueno", "malo", o ni uno ni otro?
¿Usted calificaría el producto como: "complicado", "sencillo", o ni uno ni otro?
¿Califica a el producto como que: "repele", "atrae", o ni uno ni otro?
¿Califica a el producto como: "convencional", "novedoso", o ni uno ni otro?
¿Califica a el producto como: "incómodo", "cómodo", o ni uno ni otro?
¿Califica a el producto como: "seguro", "inseguro", o ni uno ni otro?
¿Califica a el producto como: "activante", "adormecedor", o ni uno ni otro?
¿Califica a el producto como que: "cubre expectativas", " no cubre expectativas", o ni uno ni otro?
¿Califica a el producto como: "ineficiente", "eficiente", o ni uno ni otro?
¿Califica a el producto como: "claro", "confuso", o ni uno ni otro?
¿Califica a el producto como: "no pragmático", "pragmático", o ni uno ni otro?
¿Califica a el producto como: "ordenado", "sobrecargado", o ni uno ni otro?
¿Califica a el producto como: "atractivo", "feo", o ni uno ni otro?
¿Califica a el producto como: "simpático", "antipático", o ni uno ni otro?
¿Califica a el producto como: "conservador", "innovador", o ni uno ni otro?

Anexo 3. Cuestionario de evaluación de Usabilidad de la implementación por voz (UEQ+)

En mi opinión, el Asistente de Voz ha entendido mis comandos de voz de forma: *

	1	2	3	4	5	6	7	
Complicada	○	○	○	○	○	○	○	Simple

En mi opinión, el Asistente de Voz ha entendido mis comandos de voz de forma: *

	1	2	3	4	5	6	7	
Ambigua	○	○	○	○	○	○	○	Inequívoca

En mi opinión, el Asistente de Voz ha entendido mis comandos de voz de forma: *

	1	2	3	4	5	6	7	
Imprecisa	○	○	○	○	○	○	○	Precisa

En mi opinión, el Asistente de Voz ha entendido mis comandos de voz de forma: *

	1	2	3	4	5	6	7	
Enigmática	○	○	○	○	○	○	○	Explicable

Considero que, para el asistente de voz, la propiedad de "Comprensibilidad" es: *

	1	2	3	4	5	6	7	
Completamente irrelevante	○	○	○	○	○	○	○	Muy importante

Comportamiento de la respuesta: El Asistente por Voz se comunica como un humano conversador, respondiendo de forma respetuosa, paciente, educada y confiable.

Descripción (opcional)

En mi opinión, el comportamiento de respuesta del Asistente de Voz es: *

	1	2	3	4	5	6	7	
Artificial	○	○	○	○	○	○	○	Natural

En mi opinión, el comportamiento de respuesta del Asistente de Voz es: *

	1	2	3	4	5	6	7	
Desagradable	○	○	○	○	○	○	○	Agradable

En mi opinión, el comportamiento de respuesta del Asistente de Voz es: *

	1	2	3	4	5	6	7	
Antipático	○	○	○	○	○	○	○	Simpático

En mi opinión, el comportamiento de respuesta del Asistente de Voz es: *

	1	2	3	4	5	6	7	
Aburrido	○	○	○	○	○	○	○	Entretenido

Considero que, para el Asistente de Voz, la propiedad "Comportamiento de la respuesta" es: *

	1	2	3	4	5	6	7	
Completamente irrelevante	○	○	○	○	○	○	○	Muy importante

Uso intuitivo: El uso del Asistente de Voz no requiere mayor explicación o capacitacion para su uso.

Descripción (opcional)

En mi opinión, utilizar el Asistente de Voz es: *

	1	2	3	4	5	6	7	
Difícil	○	○	○	○	○	○	○	Fácil

En mi opinión, utilizar el Asistente de Voz es: *

	1	2	3	4	5	6	7	
Ilógico	○	○	○	○	○	○	○	Lógico

En mi opinión, utilizar el Asistente de Voz es: *

	1	2	3	4	5	6	7	
Equívoco	○	○	○	○	○	○	○	Evidente

En mi opinión, utilizar el Asistente de Voz es: *

	1	2	3	4	5	6	7	
Incoherente	○	○	○	○	○	○	○	Coherente

Considero que, para el Asistente de Voz, la propiedad "Uso intuitivo" es: *

	1	2	3	4	5	6	7	
Completamente irrelevante	○	○	○	○	○	○	○	Muy importante

Calidad de las respuestas: Las respuestas del Asistente de Voz cubren las necesidades de información del usuario.

Descripción (opcional)

Las preguntas y respuestas del Asistente de Voz son: *

	1	2	3	4	5	6	7	
Inapropiadas	○	○	○	○	○	○	○	Adecuadas

Las preguntas y respuestas del Asistente de Voz son: *

	1	2	3	4	5	6	7	
Inútiles	○	○	○	○	○	○	○	Útiles

Las preguntas y respuestas del Asistente de Voz son: *

	1	2	3	4	5	6	7	
Proveen ayuda	○	○	○	○	○	○	○	No proveen ayuda

Las preguntas y respuestas del Asistente de Voz son: *

	1	2	3	4	5	6	7	
Ignorantes	○	○	○	○	○	○	○	Inteligentes

Considero que, para el Asistente de Voz, la propiedad "Calidad de las respuestas" es: *

	1	2	3	4	5	6	7	
Completamente irrelevante	○	○	○	○	○	○	○	Muy importante

Originalidad: El Asistente por Voz representa un enfoque nuevo y original.

Descripción (opcional)

En mi opinión, la idea detrás del Asistente de Voz y su diseño son: *

	1	2	3	4	5	6	7	
Faltos de imaginación	○	○	○	○	○	○	○	Creativos

En mi opinión, la idea detrás del Asistente de Voz y su diseño son: *

	1	2	3	4	5	6	7	
Convencionales	○	○	○	○	○	○	○	Originales

En mi opinión, la idea detrás del Asistente de Voz y su diseño son: *

	1	2	3	4	5	6	7	
Tradicionales	○	○	○	○	○	○	○	Novedosos

En mi opinión, la idea detrás del Asistente de Voz y su diseño son: *

	1	2	3	4	5	6	7	
Conservadores	○	○	○	○	○	○	○	Innovadores

Considero que, para el Asistente de Voz, la propiedad "Originalidad" es: *

	1	2	3	4	5	6	7	
Completamente irrelevante	○	○	○	○	○	○	○	Muy importante

Preguntas generales

Seguidamente se le muestran afirmaciones sobre el Asistente de Voz que debe responder según usted considere está de acuerdo o en desacuerdo.

Pregunta *

	Totalmente en ...	En desacuerdo	Ni en desacuer...	De acuerdo	Totalmente de ...
La forma de int...	○	○	○	○	○
La extensión d...	○	○	○	○	○
Las preguntas ...	○	○	○	○	○

Pregunta *

	Totalmente en ...	En desacuerdo	Ni en desacuer...	De acuerdo	Totalmente de ...
Se entiende fá...	○	○	○	○	○
Todos los conc...	○	○	○	○	○
Es sencillo sab...	○	○	○	○	○

Pregunta *

	Totalmente en ...	En desacuerdo	Ni en desacuer...	De acuerdo	Totalmente de ...
Es sencillo sab...	○	○	○	○	○
El Asistente de ...	○	○	○	○	○
Era común que,...	○	○	○	○	○

Pregunta *

	Totalmente en ...	En desacuerdo	Ni en desacuer...	De acuerdo	Totalmente de ...
Si me equivoqu...	○	○	○	○	○
Es sencillo soli...	○	○	○	○	○

Si tuviera que describir el Asistente de Voz usando solo 3 palabras o conceptos ¿Cuáles serían? *

Texto de respuesta corta

Si desea resaltar aspectos positivos o negativos del Asistente de Voz, puede hacerlo en este espacio

Texto de respuesta larga

Preguntas adicionales

Descripción (opcional)

¿Cuántos años tiene usted? *

Texto de respuesta corta

¿Usted es...? *

- ○ Mujer
- ○ Hombre
- ○ Otro / Prefiero no decirlo