SOVO: Usability Questionnaire for Voice-Only User Interfaces

A. L. Iñiguez-Carrillo, A. Venegas-Reynoso, L. S. Gaytán-Lugo and P. C. Santana-Mancilla

***Abstract*— This article presents the creation and validation of the SOVO instrument, which measures perceived usability by users. To achieve this, we conducted a three-stage instrument study. Twenty out of 160 items from nine questionnaires used to assess usability in VUI (Voice User Interfaces) were selected. The researchers established factors and the type of response. Expert judges supported the application of CVC, and after two rounds, the instrument had 15 items, with eight items rated as good and seven as excellent. Next, we administered the SOVO instrument to a sample of 314 users and conducted a statistical analysis, which validated the instrument. We found that the items demonstrate high internal consistency, with an alpha of 0.96 and a lambda-6 of 0.97. Finally, we performed an exploratory factor analysis, identifying three factors: Likeability, Intelligibility, and Usability. It is important to have reliable instruments since it is very common to see VUI in a wide range of the users' activities.**

***Index Terms*— Questionnaire, SOVO, Usability, Voice User Interfaces.**

I. Introducción

Una interfaz de voz de usuario (VUI, por sus siglas en inglés) es una interfaz de un sistema computacional que permite a los usuarios interactuar con el sistema a través del habla, utilizando el reconocimiento automático de voz para interpretar las entradas de voz y la síntesis de voz para proporcionar las salidas de voz. Para interactuar las VUI, el dispositivo utilizado debe contar con un micrófono y una bocina. La implementación de asistentes personales inteligentes (IPA, por sus siglas en inglés) como Alexa, Siri y Google Assistant han popularizado este tipo de interfaces. En 2021, más de 20% de las personas de países occidentales indicaron usar asistentes de voz digitales varias veces al día [1]. Se estima que, en el 2024 habrá 8.4 billones de asistentes de voz digitales [2].

El objetivo principal de utilizar la voz como medio de interacción es ofrecer una forma más natural e intuitiva de interactuar con un sistema computacional, lo que puede resultar en una experiencia de usuario más rápida, fácil y agradable. Estas interacciones deben ser tan intuitivas y naturales como las conversaciones entre dos humanos. Una forma de verificarlas es aplicando evaluaciones de usabilidad, la cual define con qué eficacia, eficiencia y satisfacción puede un usuario interactuar con una interfaz de usuario [3].

May 22, 2023.

A. L. Iñiguez-Carrillo is with Universidad de Guadalajara, Cd. Guzmán, JAL 49000 MEX (e-mail: adriana.iniguez@academicos.udg.mx).

A. Venegas-Reynoso is with Université de Lille, Villeneuve d'Ascq, Nord, Francia. (e-mail: adrian.venegasreynoso.etu@univ-lille.fr)

(e-mail: adrian.venegasreynoso.etu@univ-lille.fr)

L. S. Gaytán-Lugo, is with Universidad de Colima, Coquimatlán, COL 28400 MEX. (e-mail: laura@ucol.mx).*

P. C. Santana-Mancilla, is with Universidad de Colima, Colima, COL 28040 MEX. (e-mail: psantana@ucol.mx).

Diversos cuestionarios han sido utilizados para evaluar la usabilidad de tecnologías que utilizan la voz como medio de interacción, tales como AttrakDiff [4] y UEQ [5], que fueron diseñados para sistemas interactivos de propósito general, por lo que no contienen ítems respecto al uso de la voz; o como MOS-X [6] y SASSI [7] que sí toman en cuenta la voz en la evaluación, pero que se enfocan únicamente en elementos específicos de esta.

En este trabajo se describe el diseño y validación de SOVO (Sólo Voz), un instrumento que tiene el objetivo evaluar la usabilidad de sistemas computacionales que utilizan la voz como principal medio de interacción. SOVO se diseñó en español, puesto que se busca evaluar interfaces que utilizan la voz en este idioma.

A. Interacción por Medio de la Voz

La interacción de voz se basa en el uso del lenguaje natural para apoyar la comunicación. Este tipo de aplicaciones requieren de una secuencia de interacciones (turnos de diálogo) entre el usuario y el sistema informático para lograr la tarea deseada [8]. Su ventaja principal es que permite a los usuarios realizar actividades simultáneas; por ejemplo, conducir y solicitar una llamada telefónica, sin quitar las manos del volante. Además, este tipo de interacción favorece la accesibilidad, ya que los usuarios con discapacidad visual o motriz pueden utilizar su voz para iniciar un proceso automatizado. La comunicación mediante voz también resulta práctica, dado que su naturaleza intuitiva demanda menos capacitación en comparación con otros sistemas [9]. Además, realizar una tarea con la voz suele ser más rápido que teclear [10]. Hoy en día, las interacciones por medio de la voz a través de un IPA muestran algunas cualidades asociadas a la inteligencia, muchas aplicaciones de voz tienen acceso a una gran cantidad de información en un dispositivo o en línea, lo que les permite realizar diversas tareas [11]. Esto se realiza a través de servicios de procesamiento en la nube, por lo que la velocidad de respuesta del sistema es un elemento indispensable para mantener una conversación sin pausas.

Las interacciones que utilizan la voz se pueden clasificar en tres tipos: *solo voz*, *primero voz* y *voz añadida* [12]. En las interacciones de *solo voz* no hay retroalimentación visual, toda la interacción se realiza por medio de la voz del usuario y por medio de una voz sintética del sistema computacional. En la interacción de *primero voz*, la voz es el principal medio de interacción, sin embargo, el usuario recibe retroalimentación visual y puede interactuar por medio del tacto con una pantalla.

Las interacciones con *voz añadida*, se refiere a dispositivos como teléfonos inteligentes y pantallas con la adición de control por voz, siendo esta última sólo una opción. Por lo tanto, la experiencia de la interacción humano-computadora es diferente si la voz es el único medio de interacción o si hay varios canales de interacción disponibles.

B. Herramientas para Evaluar la VUI

El campo de las técnicas y medidas de evaluación para sistemas conversacionales que utilicen la voz como medio de interacción se encuentra en una etapa temprana de desarrollo [8]. Se ha utilizado el método de inspección para encontrar problemas de usabilidad en el diseño de las especificaciones de la interfaz de usuario que aún no se han implementado necesariamente, tales como evaluaciones heurísticas y paseos cognitivos [13]. Lo que significa que la inspección se puede llevar a cabo al comienzo del ciclo de vida de un sistema [14]. Este tipo de evaluación se caracteriza por el uso de expertos para emitir informes, juicios y posibles errores en el diseño de un sistema. Otro método es el testing, que consiste en detectar las necesidades del usuario y obtener información específica del sistema [15]. Wei y Landay [16] proponen una lista de heurísticas creadas específicamente para interfaces basadas en voz. Un método de uso frecuente es la observación, con este método es posible evaluar la interacción con un sistema basado en el habla tomando evidencia sistemáticamente de lo que se ve y lo que se escucha dentro de un contexto real [17]. Se ha utilizado pensamiento en voz alta, donde debe diferenciarse los comandos realizados hacia el sistema y los comentarios de los usuarios cuando no se está diciendo un comando durante la interacción [18]. Así también, los cuestionarios son una herramienta muy popular para realizar evaluaciones de usabilidad debido a que son una forma sencilla de recopilar datos, pueden utilizarse durante todo el proceso de desarrollo y es posible abarcar a un grupo de usuarios numeroso.

El método utilizado para evaluar las VUIs debe tener en cuenta sus características inherentes. Las interfaces que utilizan la voz son secuenciales (el habla ocurre una palabra a la vez), dinámicas (el habla cambia constantemente) y transitorias (no dejan un registro permanente) [15].

C. Cuestionarios para Evaluar Interacciones Basadas en Voz

Se han utilizado diversos cuestionarios para evaluar la usabilidad en sistemas que utilizan la voz. Por ejemplo, los cuestionarios MOS-X[6] y SUISQ-R[19] fueron diseñados para evaluar sistemas que utilizan la voz en la interacción, sin embargo, MOS-X solo se centra en la calidad de voz y habla sintética, más no en la usabilidad; mientras que SUISQ-R [19] solo se enfoca en sistemas de respuesta de voz interactiva. PARADISE [20] se enfoca en evaluar únicamente la calidad de la salida de voz y SASSI [7] solo en la calidad de la entrada de voz. Los cuestionarios AttrakDiff, ICF-US, SUS, UEQ y USE fueron diseñados para evaluar la interacción general del sistema con el usuario, en sus ítems no existen preguntas ni oraciones para evaluar la voz [21]. Por otro lado, es una práctica habitual utilizar diferentes cuestionarios para evaluar la usabilidad. Por ejemplo, MOS-X [6] evalúa principalmente la calidad del habla, por lo que se recomienda usarlo junto con otro cuestionario para tener una medida más completa de usabilidad. Según Ghosh et al. [22], SUS es una herramienta de evaluación válida para evaluar sistemas que interactúan por medio de la voz, sin embargo, puede no ser la mejor opción para evaluar VUI ya que su diseño está profundamente arraigado a entornos gráficos [23]. Otra práctica usual, es crear cuestionarios personalizados sin que estos pasen por un proceso de validación [24], por lo que no se garantiza que el instrumento realmente mide lo que pretende medir.

II. MÉTODO

Considerando que no existe un instrumento especializado que permita evaluar este tipo de tecnología de manera concreta, fue necesario desarrollar SOVO, un cuestionario que mide la usabilidad de las interfaces que utilizan únicamente la voz como medio de interacción. El estudio se realizó en tres etapas: 1) diseño del cuestionario, 2) validación de contenido, y 3) un análisis estadístico donde se expone la validación de confiabilidad y el análisis factorial exploratorio.

A. Etapa 1. Diseño del Cuestionario

Para el diseño del instrumento propuesto en esta investigación, se realizó una revisión teórica de los instrumentos aplicados a las VUI. Se analizaron nueve cuestionarios que fueron el resultado de una revisión sistemática en [21] (Ver Tabla I).

En esta revisión se encontró que el 44% de los trabajos analizados evalúan la usabilidad con cuestionarios ad-hoc. Un problema de esta forma de evaluar es que los cuestionarios autoconstruidos pueden no estar debidamente validados [24], lo que no garantiza que el instrumento mida lo que pretende medir; el resto (56%) utiliza cuestionarios estandarizados que han sido sometidos a una calificación psicométrica, sin embargo, no todos sus ítems aplican a las VUI debido a que han sido diseñados para interfaces gráficas o de propósito general.

TABLA I
CUESTIONARIOS PARA EVALUAR LA USABILIDAD EN INTERFACES DE VOZ [21]

	Mide	Tipo de interfaz
AttrakDiff	Usabilidad, atractividad y estética	Propósito general
ICF-US	Usabilidad	Propósito general
MOS-X	Calidad de voz y habla sintética	Sistemas de voz
SUISQ-R	Usabilidad	Sistemas de respuesta de voz interactiva
SUS	Usabilidad	Propósito general
SASSI	Usabilidad	Sistemas de voz
UEQ	Usabilidad y experiencia de usuario	Propósito general
PARADISE	Satisfacción de usuario	Agentes de diáogo
USE	Usabilidad	Propósito general

1) Diseño de Items

Se categorizaron un total de 160 ítems de los nueve

cuestionarios mencionados. Estos ítems se agruparon según los elementos de usabilidad: eficiencia, eficacia y satisfacción [25]. En una segunda iteración de revisión, el número de ítems se redujo a 26, ya que solo se seleccionaron aquellos que funcionaban para evaluar dispositivos solo de voz teniendo en cuenta la heurística de VUI propuesta por Wei y Landay [16] y las pautas de diseño para la interacción de voz con manos libres [26]. Además, se inició con la revisión del idioma, ya que originalmente los ítems estaban en inglés, por lo que se requirió la traducción al español. En este sentido, Gao y Kortum [27] argumentan que es importante usar instrumentos en el idioma nativo de la cultura destino para evitar resultados menos confiables. Para la tercera ronda de revisión, los elementos que eran similares se fusionaron o eliminaron, resultando en 20 ítems (Ver Tabla II), los cuales fueron categorizados en factores.

TABLA II
PRIMERA VERSIÓN DEL INSTRUMENTO SOVO

ID	Item	CVC
R1	Hace que las cosas que quiero lograr sean más fáciles de hacer	0.77
R2	Me sentí incómodo con la voz y tono del asistente	0.80
R3	Puedo utilizar diferentes frases para ejecutar una tarea	0.81
R4	El sistema tiene cambios de voz o contexto que resultan extraños	0.68
R5	Supe decir las frases correctas para encontrar lo que necesitaba sin ninguna dificultad	0.71
R6	El sistema no entendía lo que le decía	0.81
R7	Entendí lo que me decía el sistema, los mensajes y las respuestas del sistema eran claros	0.78
R8	Necesito un nivel alto de concentración para usar el sistema	0.77
R9	El sistema me contestaba correctamente	0.74
R10	El sistema se tardaba en contestarme	0.85
R11	Pude realizar la tarea que quería hacer	0.84
R12	No puede recuperarme fácilmente de los errores	0.72
R13	Tuve suficiente guía sobre las capacidades, limitaciones y operaciones del sistema	0.69
R14	Cuando necesité ayuda no supe encontrarla fácilmente	0.74
R15	Me siento satisfecho con el uso del asistente	0.88
R16	No me gustaría usar este sistema con frecuencia	0.81
R17	Fue agradable usar el sistema	0.90
R18	Es difícil aprender a usar el sistema	0.88
R19	Recordaré cómo usar el sistema fácilmente	0.80
R20	Durante la interacción no sabía el estado del sistema, no me daba cuenta si me estaba escuchando	0.76

Con este análisis, el instrumento contó con las dimensiones de usabilidad: eficacia, eficiencia y satisfacción; así como las variables: facilidad de uso, naturalidad, flexibilidad, recuperación de errores, inteligibilidad, confiabilidad, desempeño, uso futuro, simpatía, facilidad de aprendizaje, habitabilidad, consistencia y memorabilidad. El tipo de respuesta que se estableció fue una escala tipo Likert de 5 puntos desde 1: Totalmente de acuerdo hasta 5: Totalmente en desacuerdo.

B. Etapa 2. Validación del Contenido

La validez de contenido asegura la calidad y precisión de un instrumento [28]. Se invitó a jueces expertos para validar el contenido utilizando el Coeficiente de Validez de Contenido (CVC) que mide la concordancia y la validez de contenido[29]. Después, se realizó una selección de expertos del área con experiencia. Se buscó diversidad entre academia, industria, rangos de edad y género. Entre las características de los jueces expertos se destaca que seis son hombres y tres mujeres. La edad promedio es de 41 años. Siete de ellos cuentan con nivel doctoral. Entre sus áreas de especialización destacan Interacción Humano-Computadora (IHC), Experiencia de Usuario (UX) y diseño y evaluación de VUI.

Una vez que los jueces expertos confirmaron su participación, se les envió un correo electrónico con un documento de presentación e información sobre la actividad a realizar. Se adjuntó el instrumento SOVO; así como el formato para evaluar el instrumento en cuestión. Los indicadores que se declararon para revisión específica de los jueces expertos fueron coherencia, claridad, escala y relevancia. Además, se agregó un espacio para retroalimentar de forma cualitativa cada ítem. Por último, al final del formato se anexó un espacio para que el juez experto dé una crítica constructiva sobre el instrumento en general.

Una vez recolectada la información, se inició con el cálculo del CVC, de acuerdo con los indicadores evaluados. El ítem 4 y el 13 resultaron deficientes con 0.68 y 0.69 respectivamente, así como el ítem 5 resultó en un mínimo aceptable 0.71. Además, se eliminaron los reactivos 9 y 19, puesto que, aunque resultaron aceptables, de acuerdo con los comentarios de los jueces eran similares a otros ítems, sonaban confusos, o bien, no era sencillo responder a lo que cuestionaban.

Dado lo anterior, en esta etapa de la validación se terminó por desechar cinco de los veinte ítems. Asimismo, se realizaron cambios para la segunda versión del instrumento siguiendo las observaciones de los expertos, entre dichos cambios destacan: 1) la edición de ítems que aparecían con una connotación negativa [30] y el reacomodo en la escala de respuesta (desde 1: totalmente en desacuerdo hasta 5: totalmente de acuerdo).

Para la segunda ronda de evaluación por jueces expertos se tuvo el apoyo de cinco de los nueve jueces expertos. Se realizó el mismo procedimiento narrado anteriormente, pero enviando la nueva versión del instrumento. Los resultados del CVC se aprecian en la Tabla III que, como se observa, tienen resultados buenos (entre 0.81 y 0.90) o excelentes (> 0.90).

C. Etapa 3. Análisis Estadístico

Para esta etapa se diseñó una actividad donde participaron 314 usuarios, 49.7% de género masculino, 50% de género femenino y 0.3% prefiere no decirlo. Las edades de los

participantes oscilaron entre 17 y 43 años, siendo la edad promedio 20 años. Para la actividad utilizaron un sistema informático de tutorías universitarias desarrollado en la plataforma Amazon Lex, en el cual el principal medio de interacción es la voz. Donde realizaron una serie de tareas para interactuar con el sistema, como preguntar sobre el reglamento de alumnos, así como información de profesores y de contenidos de los cursos. Al finalizar la actividad se les aplicó el instrumento SOVO a todos los usuarios. Con los resultados del instrumento se obtuvo una base de dato.

TABLA III
SEGUNDA VERSIÓN DEL INSTRUMENTO SOVO

ID	Item	CVC
R1	El asistente permite que las tareas que quiero realizar sean fáciles de hacer	0.86
R2	Me siento cómodo con la voz del asistente	0.86
R3	Puedo utilizar diferentes frases para realizar una tarea	0.89
R4	El asistente permite que las tareas que quiero realizar sean fáciles de hacer	0.86
R5	El asistente entiende lo que le digo	0.86
R6	Entiendo lo que me dice el asistente, sus mensajes son claros	0.89
R7	El asistente me da información fiable	0.89
R8	El asistente me responde rápido	0.98
R9	Puedo realizar la tarea que quiero hacer de forma efectiva	0.85
R10	Cuando se presenta un error, el asistente me sugiere soluciones que funcionan	0.99
R11	Cuando necesito ayuda, el asistente me la proporciona.	0.98
R12	Me siento satisfecho con el uso del asistente	1.00
R13	Usaría este asistente con frecuencia	0.99
R14	Es agradable conversar con el asistente	0.92
R15	Es fácil aprender a usar el asistente	0.99

El análisis estadístico fue realizado utilizando el lenguaje de programación R [31], junto con las librerías polychor y psych [32] para el cálculo de pruebas diagnósticas, matrices de correlación policóricas y análisis factorial. Después de la eliminación de 12 datos atípicos, el tamaño de la muestra fue de 302 estudiantes.

Antes de comenzar con las pruebas diagnósticas se calculó la matriz de correlación. Dado que los datos obtenidos por medio de una escala Likert son ordinales, se utilizó la correlación policórica en lugar de la correlación de Pearson [33]. Posteriormente, se utilizaron dos criterios para determinar si los datos eran adecuados para el análisis factorial: la prueba de Kaiser-Meyer-Olkin (KMO) de adecuación del muestreo y la prueba de esfericidad de Bartlett para la adecuabilidad de la correlación. La prueba KMO indica la cantidad de varianza común en los datos que puede ser causada por factores subyacentes. Esta prueba produce valores entre 0 y 1, donde valores menores a 0.5 se consideran inaceptables, entre 0.50 y 0.59 pobres, entre 0.60 y 0.69 bajos, entre 0.70 y 0.79 modestos, entre 0.80 y 0.89 buenos y entre 0.90 y 1.0 excelentes [34].

Por otra parte, la prueba de esfericidad de Bartlett, tiene por hipótesis nula que la matriz de correlación de los datos es una matriz identidad, implicando que las variables o ítems se encuentran totalmente no correlacionados, y por tanto, no son aptos para un análisis factorial. La validación del instrumento fue calculada a partir del alfa de Cronbach y la lambda 6 de Guttman [27], [28]. Estos estadísticos se encuentran entre 0 y 1, donde valores superiores a 0.70 indican alta consistencia interna. Posteriormente, se utilizó análisis factorial exploratorio para el desarrollo del modelo. El número *n* de factores a extraer fue obtenido usando el método de análisis paralelo, el cual indica el número óptimo de factores al comparar la magnitud de los eigenvalores extraídos de la matriz de correlación contra eigenvalores de datos obtenidos aleatoriamente [35]. Posteriormente, se extrajeron los modelos con el número de factores indicados por el análisis factorial, junto con los modelos con $n \pm 1$ factores utilizando el método de residuos mínimo. A continuación, se utilizó la rotación oblicua *oblimin* para obtener factores fácilmente interpretables. Finalmente se evaluaron los modelos utilizando cuatro parámetros de bondad del ajuste: el índice de Tucker-Lewis (TLI), el error cuadrático medio de la aproximación (RMSEA), el residuo cuadrático medio (RMSR) y el criterio de información Bayesiana. Los criterios de estas pruebas se encuentran en la Tabla IV.

TABLA IV
ESTADÍSTICOS DE BONDAD DEL AJUSTE Y SUS CRITERIOS [43]

Estadístico	Excelente	Aceptable	Pobre
TLI	> 0.95	> 0.90	< 0.90
RMSEA	< 0.06	< 0.09	> 0.10
RMSR	< 0.06	< 0.09	> 0.10
BIC	Menor es mejor		

III. RESULTADOS

A. Pruebas Diagnósticas

Se encontró que los ítems tienen valores de KMO entre 0.99 y 0.90, siendo R7 el ítem con el coeficiente más alto, mientras R13, el más bajo. Por otra parte, el KMO global es de 0.94, indicando una excelente adecuabilidad de la muestra.

Por otra parte, la prueba de esfericidad de Bartlett produjo una chi cuadrada de 4604.6 (gl=105), con un valor p de 0, por tanto se rechaza la hipótesis nula y se concluye que la matriz de correlación de los datos no es una matriz identidad. Respecto a la validez del instrumento, se encontró que los ítems presentan una alta consistencia interna, al presentar una alfa de 0.96 y una lambda-6 de 0.97, lo que indica que los ítems del instrumento están altamente correlacionados entre sí y miden lo mismo.

B. Análisis Factorial Exploratorio

El análisis paralelo sugirió la extracción de 3 factores, sin embargo, modelos de 2 y 4 factores también fueron calculados. Al analizar el modelo de 4 factores se encontró que uno de los factores sólo estaba compuesto por un ítem, por lo que fue descartado. Al calcular la bondad del ajuste se encontró que el modelo de 3 factores tiene un índice de Tucker-Lewis de 0.92, encontrándose en el rango aceptable. El error cuadrático medio de la aproximación (RMSEA) es de 0.09 y el residuo cuadrático medio (RMSR) de 0.03, por lo que se encuentran en el rango aceptable y excelente, respectivamente. Finalmente, el criterio de información Bayesiana (BIC) es de -70. Estos estadísticos expresan un mejor ajuste que los obtenidos por el modelo de 2 factores (TLI=0.87, RMSEA=0.11, RMSR=0.04, BIC=64). El modelo de 3 factores se presenta en la Tabla V.

TABLA V
CARGAS FACTORIALES DESPUÉS DE ROTACIÓN OBLIMIN

	FACTOR1	FACTOR 2	FACTOR 3
R1	**0.88**	-0.05	0.06
R2	0.07	0.24	**0.53**
R3	**0.78**	0.07	-0.05
R4	**0.91**	-0.07	0.04
R5	0.23	**0.51**	0.11
R6	0.22	**0.67**	0.04
R7	**0.65**	0.25	-0.05
R8	**0.86**	0.02	0.05
R9	**0.75**	0.08	-0.04
R10	**0.60**	0.42	-0.10
R11	**0.78**	-0.01	0.19
R12	0.41	0.00	**0.52**
R13	0.03	0.10	**0.91**
R14	-0.07	**0.80**	0.15
R15	0.11	**0.65**	0.20
S2 (%)	54.8	31.5	13.6

IV. DISCUSIÓN

Uno de los puntos interesantes respecto al proceso de diseño y evaluación del instrumento fue durante la etapa de retroalimentación de los jueces expertos donde se encontró que el ítem 4 y el 13 resultaron deficientes (Tabla II). Respecto al ítem 4, uno de los comentarios señalados por los evaluadores fue el siguiente: "No estoy tan seguro de que abone a la experiencia de usuario. Extraño pudiera ser inusual o atípico, aunque tiene una connotación negativa..."

La mayoría de los evaluadores realizó una crítica respecto al uso de la palabra extraño. Por lo que se decidió eliminar dicho ítem. Respecto al ítem 13, la mayoría de los evaluadores resultaron confundidos respecto a la redacción del reactivo, como se muestra a continuación: "No me queda claro la guía que puede dar el asistente, creo que un primer paso es informar las capacidades, limitaciones y operaciones del sistema básicas y un segundo paso es guiar sobre su uso."

Por lo anterior, aunado a su baja valoración, el ítem 13 se eliminó. Dos ítems más fueron señalados por preguntar lo mismo con diferentes palabras. Mientras que uno más, se prestaba a confusión al usar el nivel concentración como un atributo negativo. Por ello, se realizaron ajustes en los ítems, y se determinó eliminar cinco de estos; para después enviar a una segunda ronda de evaluación, pasando de un resultado general de 0.79, a un 0.93.

Por otro lado, durante el análisis estadístico se realizó el EFA donde se obtuvieron modelos de dos, tres y cuatro factores. Se analizó el modelo de 4 factores, sin embargo, se encontró que uno de los factores sólo estaba compuesto por un ítem por lo que se descartó, ya que no es recomendable [36]. Al comparar el modelo de 3 factores vs el modelo de 2 factores, se eligió el primero, ya que se tuvo una mejor bondad del ajuste.

Antes de aplicar el EFA, se hipotetizó sobre la clasificación de los ítems respecto a las dimensiones de usabilidad de la norma ISO 9241-11, además de sugerir una dimensión de voz como un cuarto factor del modelo. Sin embargo, dado que con el EFA se identificaron 3 factores, se realizó la agrupación de los ítems de acuerdo con dicho análisis, y se clasificaron en función de sus características, por lo que los factores que se encontraron fueron: Agradable, Inteligible y Usable.

Los ítems R2, R12 y R13 se relacionan al factor Agradable (Ver Tabla III). Este factor refiere a que el usuario se siente cómodo usando el sistema, la voz del sistema le resulta agradable y le gustaría utilizar de nuevo dicho sistema. Este factor es crítico para la adopción de una tecnología [37].

Coppens et al. [38] definen la inteligibilidad del habla como la claridad con la que una persona habla de modo que su discurso sea comprensible para un oyente. En el caso de las VUI se puede definir como la claridad con la que la VUI transmite un mensaje de modo que su discurso sea comprensible para el usuario y viceversa. Los ítems R5, R6, R14 y R15 se relacionan al factor Inteligible. El usuario entiende lo que le dice el sistema y, a la vez, el asistente interpreta correctamente las frases del usuario. Respecto al R14, Nielsen [14] explica que la capacidad de aprendizaje es, en cierto sentido, el atributo de usabilidad más fundamental. En este trabajo, dicha capacidad se refiere a que es fácil "aprender" a usar el asistente de voz. La inteligibilidad también se refiere a aquello que puede ser entendido sin problema. Por otro lado, el uso del lenguaje humano implica la necesidad de contar con un sistema de comunicación efectivo que permita la transmisión de información precisa y confiable, sin que esta sufra alteraciones o distorsiones en el proceso [39]. Por ello, para el R6, se busca que la información que se solicita al asistente llegue de manera íntegra. En general, si una VUI no es inteligible, la usabilidad general puede verse afectada por la frustración.

El factor usable, en este trabajo, se refiere al grado en que un

sistema puede ser utilizado por usuarios específicos para lograr objetivos específicos con eficacia, eficiencia y satisfacción en un contexto de uso específico [40]. Los ítems que se relacionan con el factor usable son R1, R3, R4, R7, R8, R9, R10 y R11. Como se observa, es el factor que cuenta con la mayor cantidad de ítems; esto se debe a que tal como su definición lo indica, y debido al modelo estadístico encontrado, los ítems relacionados con los atributos de satisfacción, eficacia y eficiencia se encontraron dentro del mismo factor.

Si bien este instrumento busca evaluar la usabilidad de las interfaces de sólo voz como medio de interacción, es importante destacar que, 1) otros autores también han extendido el concepto de usabilidad incluyendo otros atributos que contribuyen en general a una mejor la experiencia de usuario [41]; y 2) no todas las tecnologías pueden evaluarse de la misma forma al tener características propias, lo cual es el caso de las VUI [42]. Finalmente, este resultado que se obtuvo no deja ser un modelo con un error del ajuste, por lo que se debe evaluar otros asistentes de voz para confirmar los resultados obtenidos.

V. Conclusión

Las VUIs son cada vez más comunes, lo que plantea nuevos desafíos en cuanto a usabilidad de estas interfaces. Tener métodos de evaluación confiables aumentará la presencia de esta tecnología en una gama más amplia de actividades, apoyando la interacción natural con los humanos.

Aunque el español es el segundo idioma más hablado en el mundo, los instrumentos reportados en la literatura se han centrado principalmente en la población angloparlantes. Por ello, es importante contar también con instrumentos en el idioma debidamente validados.

Estas tecnologías se están implementando en diversos contextos, a medida que se expande su uso es necesario contar con instrumentos de evaluación que ayuden a identificar procesos, conductas o acciones que reducen la usabilidad del asistente, y con ello poder identificar mejoras en su diseño, uso y aplicación.

Referencias

[1] B. Thormundsson, "Voice technology - statistics & facts | Statista." https://www.statista.com (accessed May 16, 2023).

[2] F. Laricchia, "Number of voice assistants in use worldwide 2019-2024 | Statista." https://www.statista.com/ (accessed May 16, 2023).

[3] F. Paz and J. A. Pow-Sang, "Usability Evaluation Methods for Software Development: A Systematic Mapping Review," *ASEA 2015*, vol. 10, no. 1, pp. 1–4, 2016, doi: 10.1109/ASEA.2015.8.

[4] M. Hassenzahl, M. Burmester, and F. Koller, "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität," 2003, doi: 10.1007/978-3-322-80058-9_19.

[5] M. Hernández-Campos, J. Thomaschewski, and Y. C. Law, "Results of a Study to Improve the Spanish Version of the User Experience Questionnaire (UEQ)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. InPress, no. InPress, 2022, doi: 10.9781/ijimai.2022.11.003.

[6] M. D. Polkosky and J. R. Lewis, "Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X," *Int J Speech Technol*, vol. 6, no. 2, 2003, doi: 10.1023/A:1022390615396.

[7] K. Hone, "Usability measurement for speech systems: SASSI revisited," *Proceedings of CHI*, no. 1, 2014.

[8] J. F. Quesada Moreno, Z. Callejas Carrión, and D. Griol Barres, "Informe sobre sistemas conversacionales multimodales multilingues," 2019. [Online]. Available: https://www.plantl.gob.es/

[9] A. Mhaidli, M. K. Venkatesh, Y. Zou, F. Schaub, and M. Kandadai, "Listen Only When Spoken To: Interpersonal Communication Cues as Smart Speaker Privacy Controls," *Proceedings on Privacy Enhancing Technologies*, 2020, doi: 10.2478/popets-2020-0026.

[10] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng, and J. Landay, "Speech Is 3x Faster than Typing for English and Mandarin Text Entry on Mobile Devices," 2016.

[11] D. Ramos, "Voice Assistants" *Smartsheet.com*, 2018. https://www.smartsheet.com/voice-assistants-artificial-intelligence (accessed Oct. 31, 2020).

[12] D. A. Coates, *Voice Aplications for Alexa y Google Assistant*. Editorial Manning, 2019.

[13] C. Wei and J. Finkelstein, "Comparison of Alexa Voice and Audio Video Interfaces for Home-Based Physical Telerehabilitation", *AMIA,* pp.496, 2022.

[14] J. Nielsen, "Usability inspection methods," *CHI '94*, pp. 413–414, 1994, doi: 10.1145/259963.260531.

[15] S. L. Hura, "Usability Testing of Spoken Conversational Systems," *Journal of Usability Studies* vol. 12, pp. 155–163, 2017.

[16] Z. Wei and J. A. Landay, "Evaluating Speech-Based Smart Devices Using New Usability Heuristics," *IEEE Pervasive Computing*, vol. 17, no. June, pp. 84–96, 2018, doi: 10.1109/MPRV.2018.022511249.

[17] S. Atreja *et al.*, "How Do People Interact in Conversational Speech-Only Search Tasks : A Preliminary Analysis," *Univers Access Inf Soc*, vol. 1, no. 1, pp. 1–12, 2018, doi: 10.1145/2160601.2160619.

[18] A. Teixeira *et al.*, "Design and development of Medication Assistant: older adults centred design to go beyond simple medication reminders," *Univers Access Inf Soc*, vol. 16, no. 3, pp. 545–560, 2017, doi: 10.1007/s10209-016-0487-7.

[19] J. R. Lewis and M. L. Hardzinski, "Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire," *Int J Speech Technol*, vol. 18, no. 3, pp. 479–487, 2015, doi: 10.1007/s10772-015-9289-1.

[20] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with PARADISE," *Nat Lang Eng*, vol. 6, 2000, doi: https://doi.org/10.1017/s1351324900002503.

[21] A. L. Iniguez-Carrillo, L. S. Gaytan-Lugo, M. A. Garcia-Ruiz, and R. Maciel-Arellano, "Usability Questionnaires to Evaluate Voice User Interfaces," *IEEE Latin America Transactions*, vol. 19, no. 9, pp. 1468–1477, 2021, doi: 10.1109/TLA.2021.9468439.

[22] D. Ghosh, P. S. Foong, S. Zhang, and S. Zhao, "Assessing the utility of the system usability scale for evaluating voice-based user interfaces," *ACM International Conference Proceeding Series*, pp. 11–15, 2018, doi: 10.1145/3202667.3204844.

[23] D. S. Zwakman, D. Pal, and C. Arpnikanondt, "Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa," *SN Comput Sci*, vol. 2, no. 1, p. 28, 2021, doi: 10.1007/s42979-020-00424-4.

[24] L. B. Larsen, "Assessment of spoken dialogue system usability - What are we really measuring?," *EUROSPEECH 2003*, pp. 1945–1948, 2003.
[25] ISO, "Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts, ISO 9241-11:2018(en)," 2018. https://www.iso.org/
[26] C. Murad, C. Munteanu, L. Clark, and B. R. Cowan, "Design guidelines for hands-free speech interaction," *MobileHCI 2018,* pp. 269–276, 2018, doi: 10.1145/3236112.3236149.
[27] M. Gao, P. Kortum, and F. L. Oswald, "Multi-Language Toolkit for the System Usability Scale," *Int J Hum Comput Interact*, vol. 36, no. 20, pp. 1883–1901, 2020, doi: 10.1080/10447318.2020.1801173.
[28] L. G. Juárez-Hernández and S. Tobón, "Análisis de los elementos implícitos en la validación de contenido de un instrumento de investigación," *Revista Espacios*, vol. 39, no. 53, pp. 1–23, 2018.
[29] R. Hernández-Nieto, *Instrumentos de Recolección de Datos en Ciencias Sociales y Ciencias Biomédicas*. Universidad de los Andes, 2011.
[30] J. Sauro and J. R. Lewis, "When designing usability questionnaires, does it hurt to be positive?," in *CHI '11*, NY, USA, ACM Press, 2011, p. 2215. doi: 10.1145/1978942.1979266.
[31] R Core Team, "R: A language and environment for statistical computing," 2020. https://www.r-project.org/ (accessed May 16, 2023).
[32] W. Revelle, "Procedures for Psychological, Psychometric, and Personality Research [R package psych version 2.3.3]," 2023, Accessed: May 16, 2023. [Online]. Available: https://CRAN.R-project.org/
[33] F. P. Holgado–Tello, S. Chacón–Moscoso, I. Barbero–García, and E. Vila–Abad, "Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables," *Qual Quant*, vol. 44, no. 1, pp.153–166, 2010, doi:10.1007/s11135-008-9190-y.
[34] H. F. Kaiser, "A Second Generation Little Jiffy", *Psychometrika*, vol. 35, pp. 401–415, 1970. doi: 10.1007/BF02291817
[35] J. L. Horn, "A Rationale and Test for the Number of Factors in Factor Analysis," 1965. doi: 10.1007/BF02289447.
[36] L. Fabrigar, D. Wegener, R. MacCallum, and E. Strahan, "Evaluating the use of exploratory factor analysis in psychological research.," *Psychol Methods*, vol. 4, 1999, doi: 10.1037/1082-989X.4.3.272.
[37] B. R. Gaines, M. L. Shaw, and L. L. Chen, "Utility, Usability and Likeability: Dimensions of the Net and Web," 1996. https://algo.informatik.uni-freiburg.de/ (accessed May 16, 2023).
[38] M. Coppens, H. Terband, A. Snik, and B. Maassen, "Speech Characteristics and Intelligibility in Adults with Mild and Moderate Intellectual Disabilities," *Folia Phoniatrica et Logopaedica*, vol. 68, no. 4, pp. 175–182, 2017, doi: 10.1159/000450548.
[39] F. Miyara, "El ruido y la inteligibilidad de la palabra", 2004.https://www.fceia.unr.edu.ar/acustica/biblio/inteligibilidad.pdf (accessed May 10, 2023).
[40] M. P. F. Orlando, C. A. E. Andrea, and F. I. D. Marcela, "Tools evaluation for speech recognition based on domain ontologies over the android platform," in *COLCOM 2012*, doi: 10.1109/ColComCon.2012.6233653.
[41] N. Bevan, "Classifying and selecting UX and usability measures," *VUUM,* pp. 13-18, 2008.
[42] L. Fulfagar, A. Gupta, A. Mathur, and A. Shrivastava, "Development and Evaluation of Usability Heuristics for Voice User Interfaces," in *Design for Tomorrow*, A. Chakrabarti, R. Poovaiah, P. Bokil, and V. Kant, Eds., Singapore: Springer, 2021, pp. 375–385.
[43] W. H. Finch, "Using Fit Statistic Differences to Determine the Optimal Number of Factors to Retain in an Exploratory Factor Analysis," *Educ Psychol Meas*, vol. 80, no. 2, pp. 217–241, Apr. 2020, doi: 10.1177/0013164419865769.

Adriana L. Iñiguez-Carrillo is a research professor at the University of Guadalajara. Her research interests lie in HCI, UX, voice interactions, and AI. She is a member of the National System of Researchers in Mexico and the Mexican Academy of Computing.

Adrián Venegas Reynoso is a Ph.D. student at the University of Lille, specializing in theoretical, physical, and analytical chemistry. His work focuses on the utilization of statistical methods and machine learning for the prediction of chemical properties.

Laura S. Gaytán-Lugo is a research professor at the Universidad de Colima. She focuses on HCI. She is a member of the National System of Researchers in Mexico and the Mexican Academy of Computing. She is member of SIGCHI Latin America Committee (SLAC).

Pedro C. Santana-Mancilla is a research professor at the Universidad de Colima. His research interests focus on HCI, IA, IoT, and technologies for areas that impact social well-being. He coordinates the HCI section of the AmexCOMP. He serves as the president of the AMexIHC.