PIXEL-BIT
e-ISSN: 2171-7966
REVISTA DE MEDIOS Y EDUCACIÓN
MEDIA & EDUCATION JOURNAL

Agente conversacional virtual: la inteligencia artificial para el aprendizaje autónomo

Embodied conversational agents: artificial intelligence for autonomous learning

Dr. Josué Artiles Rodríguez
Profesor Contratado Doctor. Facultad de Ciencias de la Educación. Universidad de Las Palmas de Gran Canaria. España

Dra. Mónica Guerra Santana
Profesora Ayudante Doctor. Facultad de Ciencias de la Educación. Universidad de Las Palmas de Gran Canaria. España

Dra. Mª Victoria Aguiar Perera
Profesora Titular de Universidad. Facultad de Ciencias de la Educación. Universidad de Las Palmas de Gran Canaria. España

Dra. Josefa Rodríguez Pulido
Profesora Titular de Universidad. Facultad de Ciencias de la Educación. Universidad de Las Palmas de Gran Canaria. España

Recibido: 2020-11-20 **Revisado:** 2020-12-22 **Aceptado:** 2021-02-17 **Preprint:** 2021-05-20 **Publicado:** 2021/09/01

RESUMEN

El presente artículo de investigación profundiza sobre las posibilidades de los agentes virtuales conversacionales como herramienta para tutorizar trabajos del alumnado universitario. Se utilizó una metodología cuantitativa con diseño descriptivo, correlacional y diferencial, para evaluar su usabilidad del agente conversacional en una muestra de 303 estudiantes universitarios. Para ello, se diseñó y evaluó un agente conversacional virtual para apoyar las tutorías de Trabajos de Fin de Grado con la Escala SUS (System Usability Scale). Los resultados indican que la escala posee una satisfactoria calidad métrica y una buena bondad del modelo, aspectos que se constatan en la estructura empírica y en la consistencia interna favorables del cuestionario. Los datos también muestran que existen diferencias significativas (IC 99,95%) en las variables género, grado, nivel de conocimiento y el grado de uso de chatbots. Se completó con el registro de uso real del agente, en un plazo de seis meses, por 589 estudiantes de tres titulaciones diferentes, contestando a 3025 preguntas en seis meses. En conclusión, los resultados permiten establecer criterios explicativos sobre el uso de chatbots. Es necesario seguir profundizando en este tipo de herramientas para el seguimiento y evaluación del alumnado.

ABSTRACT

This paper delves into the possibilities of conversational virtual agents as a tool to tutor university students' work. A quantitative methodology with a descriptive, correlational and differential design was used to evaluate the usability of the conversational agent in a sample of 303 university students. For this, a virtual conversational agent was designed and evaluated to support the End-of-Degree Project tutorials with the SUS Scale (System Usability Scale). The results indicate that the scale has a satisfactory metric quality and good model goodness, aspects that are verified in the empirical structure and in the favorable internal consistency of the questionnaire. The data also show that there are significant differences (99.95% CI) in the variables gender, grade, level of knowledge, and the grade of chatbot use. It was completed with the record of the actual use of the agent, within a period of six months, by 589 students from three different degrees, answering 3025 questions in six months. In conclusion, the results allow to establish explanatory criteria on the use of chatbots. It is necessary to continue deepening in this type of tools for the monitoring and evaluation of students.

PALABRAS CLAVES · KEYWORDS

Inteligencia artificial; tutoría; educación superior; agente conversacional virtual
Artificial intelligence; tutoring; University education; embodied conversational agents

1.Introducción

Los Agentes Conversacionales (AC), también denominados chatbots, son paquetes de software capaces de establecer interacciones con el ser humano utilizando el lenguaje natural (Dale, 2016). Es decir, son programas de inteligencia artificial capaces de generar una conversación concreta con personas, a través del procesamiento del lenguaje natural (Zeni et al., 2019). Son accesibles a una gran cantidad de personas, gracias a la capacidad de interaccionar mediante voz y texto (Fadhil & Gabrielli, 2017). Para ello, son necesarios múltiples datos para entrenar al agente conversacional (Fast et al., 2018; Sumikawa et al., 2020).

Generan un compromiso en la conversación capaz de centrar la atención de las personas en la interacción (Wargnier et al., 2015). Esto ha resultado de interés en investigación ya que las personas reaccionan positivamente cuando los (AC) muestran pequeñas actitudes sociales relacionadas con expresiones, gestos y señales sociales (Feine et al., 2019).

Este proceso no está exento de dificultad ya que los (AC) deben utilizar composiciones sintácticas y semánticas de gran complejidad, de manera que puedan interactuar con las expresiones de una conversación (Herbert & Kang, 2018). Los (AC) pueden desarrollar una conversación compleja sobre un tema determinado, pero presenta limitaciones cuando las conversaciones cruzan diferentes temáticas, o cuando el usuario utiliza expresiones que tiene varios significados o varias interpretaciones según el contexto (Colace et al., 2018).

La utilización de los (AC) abarca numerosos ámbitos ya que pueden programar citas médicas o reservas en restaurantes, comprar paquetes de viajes, responder a preguntas frecuentes (FAQ), denominadas así por las siglas de Frequently Asked Questions, para mejorar la salud mental de pacientes, en terapias de estrés postraumático o para mejorar la comunicación interna dentro de una organización entre otros (Fast et al., 2018; Ghosh et al., 2018; Sáez et al., 2017; Tielman et al., 2017; Schroeder et al., 2018).

1.1.El agente conversacional en procesos educativos

En el plano educativo hay universidades que brindan un escenario virtual como apoyo a las clases presenciales. El campus virtual es un entorno de enseñanza muy importante, aunque en muchas ocasiones no dispone de herramientas dinámicas para responder al alumnado. El agente conversacional puede solventar este problema, respondiendo a las necesidades de los estudiantes en una interacción directa (Sumikawa et al., 2020).

Kerly et al. (2008) señalan dos ejemplos sobre el uso de los agentes de conversación, uno que proporciona apoyo de aprendizaje, y el otro como apoyo para la autoevaluación. Además, los autores exponen que mejora la motivación, las habilidades metacognitivas y las calificaciones del alumnado. También resulta positivo para el profesorado, proporcionando apoyo a la docencia y ofreciendo datos para evaluar.

Son eficaces en el aprendizaje de otro idioma, basándose en un diálogo productivo en interacción con el alumnado. Así, un (AC) puede interactuar entre dos estudiantes mientras mantienen conversaciones a través de textos (Tegos et al., 2014; Tegos & Demetriadis, 2017). De esta forma, entre estudiantes con el mismo nivel de competencia lingüística obtenían mejores resultados aquellos que habían utilizado un (AC) en la interacción entre pares. Babu et al. (2011), estudiaron los efectos del uso de (AC) en la interacción

multimodal natural para enseñar a los usuarios protocolos culturales verbales y no verbales de conversación en la cultura del sur de India. Los resultados del estudio señalan que los participantes que se entrenaron con los (AC) tuvieron un rendimiento significativamente mayor que aquellos que aprendieron de la guía de estudio.

El apoyo a la tutoría en la universidad es otro de los usos de los (AC). Puede ayudar a solventar dudas sobre exposiciones a un número elevado de estudiantes, a modo de fuente de información y reflexión, (Wellnhammer et al., 2020). Esta opción es muy interesante en escenarios con una presencia masiva de alumnos, como los MOOC. Por ejemplo, como (AC) para aprender JAVA de manera que el profesorado también tiene una retroalimentación, en tiempo real (Catalán et al., 2018). También ha sido utilizado para el apoyo del alumnado universitario en la materia de Fundamentos de Ciencias de la Computación y Redes de Computadores (Colace et al., 2018).

En la actualidad hay una línea de implementación de (AC) en aplicaciones móviles para educación, pero se necesita más investigación sobre estas herramientas (Hobert & Meyer von Wolff, 2019). Por otra parte, es necesario un determinado nivel de conocimiento para poder desarrollar un chatbot en asignaturas universitarias que no estén vinculadas a la informática. En muchas ocasiones se implementa el chatbot en carreras técnicas, siendo menor en carreras humanísticas y de ciencias sociales. Para solventar este problema se puede llevar a cabo un diseño participativo que responda a las necesidades del profesorado que no sepa desarrollarlo (Neumann et al., 2019).

1.2. La valoración sobre el uso de un agente conversacional en procesos educativos

Un aspecto fundamental a tener en cuenta es la valoración del usuario tras la experiencia de utilizar los (AC) de manera eficaz (Ren et al., 2019). Para valorar la usabilidad de los (AC) se puede recurrir a un instrumento ampliamente utilizado y validado denominado Sistem Usability Scale (SUS) (Brooke, 1996; Brooke, 2013) que ha sido objeto de estudio y validación por otros autores (Finstad, 2010; Lewis & Sauro, 2009). Relacionado con la valoración de recursos digitales, entre muchos otros, ha sido utilizado para el estudio de chatbots que permiten mejorar los procesos de comunicación interna de una empresa por parte de 8 sujetos (Sáez et al., 2017), el estudio con 12 usuarios sobre la usabilidad de chatbot con interfaz de voz como Alexa o Siri (Ghosh et al., 2018) o con un agente conversacional para el tratamiento de trastornos mentales (Schroeder et al., 2018). En aspectos relacionados con la educación, el instrumento SUS se ha utilizado para comprobar desde la usabilidad de un agente conversacional para educar a niños y niñas con diabetes con una muestra de 21 sujetos (Sinoo et al., 2018), la formación en patrones culturales mediante chatbots estudiado en 40 sujetos (Babu et al., 2011) hasta la usabilidad de chatbot para recoger dudas sobre una materia por parte de 12 estudiantes universitarios (Neumann et al., 2019), entre otros.

El objetivo de esta investigación es valorar el uso de un agente conversacional diseñado para ofrecer ayuda al alumnado que está desarrollando su trabajo de fin de carrera. Para ello, se diseñó un chatbot, denominado CLOE, como herramienta para tutorizar al alumnado en preguntas frecuentes (FAQ frequently asked questions) relacionadas con el Trabajo Fin de Título (TFT) de diferentes Grados de Educación Social, Infantil y el Grado de Primaria. Se desarrolló en base a la aplicación Dialogflow que es una API, siglas que responden a Application Programming Interface o Interfaz de Programación de Aplicaciones, para tareas de Procesamiento de Lenguaje Natural (PLN) y comprensión del

lenguaje natural (Zeni et al., 2019). La herramienta permite exportar la base de datos para que pueda ser utilizada por otros desarrolladores y docentes en diferentes universidades.

Previamente fue testado por las coordinadoras de TFT de las diferentes titulaciones, ya que anualmente se enfrentan a numerosas cuestiones del alumnado. También fue depurado por tutores/directores de TFT y, finalmente, se estableció un árbol de respuestas para cuestiones que se aglutinaron en las siguientes dimensiones:

1) Aspectos formales del TFT tales como fuente, márgenes, interlineado, etc.

2) Tipos de TFT y sus características.

3) Aspectos relacionados con las citas.

4) Referencias según las normas APA.

5) Documentos a presentar en la administración y firma con certificado digital.

6) Fechas vinculadas con la presentación, defensa y convocatorias.

7) Aspectos relacionados con las calificaciones.

8) Respuestas de tipo variado relativas a tribunales o tutores.

9) Paquete de conversación natural para humanizar la interacción del (AC).

Una vez diseñado el (AC), y antes de la implementación, se pretende conocer, a través de la aplicación de una prueba validada, la usabilidad del chatbot a través de una muestra de estudiantes universitarios (características, aspectos técnicos, aprendizaje y funcionalidad) como herramienta intuitiva y de comunicación. En paralelo, se quiere comprobar si existen diferencias en función del género, el Grado que cursa el alumno, el nivel de conocimiento sobre chatbots y el grado de uso de chatbots.

2. Metodología

2.1. Objetivo

La presente investigación persigue un doble objetivo. Por un lado, queremos analizar las percepciones sobre la usabilidad y el manejo del chatbot en los procesos de enseñanza-aprendizaje que tienen los estudiantes universitarios de Educación Social, Educación Infantil y Educción Primaria. De forma más específica, también se pretende analizar la relación entre variables predictivas o independientes con la usabilidad de los chatbots (variable dependiente) y, para ello, tratamos de comprobar si existen diferencias significativas en función del género, el Grado que cursan, el grado de conocimiento y de uso de los chatbots. Por otro lado, queremos comprobar si el agente conversacional se usó por el alumnado como tutor virtual de apoyo.

2.2. Hipótesis

En función de los objetivos señalados hemos considerado las siguientes hipótesis:

- Hipótesis 1. La satisfacción por utilizar esta herramienta se relaciona con el grado de conocimiento sobre chatbots.
- Hipótesis 2. Los aprendizajes previos al manejo de esta herramienta se relacionan con el grado de uso de chatbots.
- Hipótesis 3. La valoración de la usabilidad de los chatbots determina la significatividad en función del género.
- Hipótesis 4. La valoración de la usabilidad de los chatbots determina la significatividad en función del Grado que cursan los estudiantes.

2.3. Diseño

Se utilizó un método de carácter cuantitativo, con diseño descriptivo, correlacional y diferencial, en situación natural.

2.4. Participantes

Desde la perspectiva cuantitativa, los supuestos matemáticos para el cálculo de la representatividad indican que, para poblaciones de 100.000 sujetos, siendo el margen de error del 6%, el nivel de confianza del 96% y el nivel de heterogeneidad del 50%, la muestra debe estar conformada por 293 sujetos. En base a ello, la muestra de nuestro estudio, basada en los criterios del muestreo probabilístico aleatorio simple (Cochran & Bouclier, 1980), está comprendida por 303 estudiantes, que fueron seleccionados al azar, de los Grados de Educación Social (13,2%), Educación Infantil (6,3%) y Educación Primaria (80,5%) de la Universidad de Las Palmas de Gran Canaria; el 70,3% son mujeres (N=213) y el 29,7% (N=90) son hombres, cuyas edades se concentran, de forma mayoritaria, entre los 18 y 24 años de edad (95%; N=288). Asimismo, los estudiantes muestran un alto grado de uso de las redes sociales, un 85,8% así lo considera. Sin embargo, al preguntarles sobre su grado de uso de chatbots y su nivel de conocimiento de los mismos, el 76,9% manifiestan que lo utilizan más bien “nada y poco”, y el 52,8% de alumnos califica su nivel de conocimiento en esta área como pobre o escaso.

2.5. Instrumento

Es importante conocer la valoración que el alumnado, como usuario, tiene sobre la utilización de un agente virtual (Ren et al., 2019). Para valorar la usabilidad de este agente, denominado CLOE, se empleó un cuestionario ampliamente utilizado y validado denominado *Sistem Usability Scale* (SUS) (Babu et al., 2011; Brooke, 1996; Brooke, 2013; Finstad, 2010; Lewis & Sauro, 2009; Ghosh et al., 2018; Neumann et al., 2019; Schroeder et al., 2018; Sinoo et al., 2018), recogiendo las aportaciones realizadas por Lewis (2018). El cuestionario aplicado contó con una serie de variables sociodemográficas donde se preguntó al estudiante por el género, la edad, el Grado que cursa. Le siguen cinco preguntas que se califican en una escala donde 1 significa “nada” y 5 significa “mucho”. Las cinco preguntas serían:

- Puntúa tu grado de uso de redes sociales.
- Puntúa tu nivel de conocimiento sobre chatbots.
- Puntúa tu grado de uso de chatbots.
- Puntúa tu uso de tutorías para el TFG.
- Versión CLOE utilizada.

2.6. Procedimiento

Los estudios sobre usabilidad de (AC) se basan en muestras de pocos sujetos entre 8 y 40 personas. Con el objetivo de conseguir una muestra más amplia se contactó con el profesorado que imparte clases al alumnado de las titulaciones vinculadas para poder administrar el cuestionario *Sistem Usability Scale (SUS)* a un colectivo amplio de estudiantes.

Todos los participantes fueron informados del objetivo de la investigación y se les solicitó su participación de forma anónima y voluntaria. Para que los estudiantes pudieran contestar al cuestionario en las mismas condiciones se implementó un protocolo para probar al (AC) donde el administrador del cuestionario señalaba donde se encontraba ubicada la herramienta en el campus virtual. Realizaron una experiencia usando cinco preguntas frecuentes:

1. ¿Cuáles son los tipos de TFT que pueden realizar?
2. ¿Cómo se realiza una cita en forma de paráfrasis?
3. ¿Cómo se referencia un libro?
4. ¿Cómo conseguir el certificado digital para firmar el TFT?
5. ¿Cuál es el periodo de validez del documento de compromiso entre alumnado y profesorado?

Se implementó el cuestionario *Sistem Usability Scale (SUS)* para valorar la usabilidad de la herramienta que habían visto previamente. La escala fue administrada a un total de 303 sujetos, lo que supone una muestra más amplia que las utilizadas, donde se administra la prueba a un rango comprendido entre 8 y 40 estudiantes (Babu et al., 2011; Ghosh et al., 2018; Neumann et al., 2019; Sáez et al., 2017; Schroeder et al., 2018; Sinoo et al., 2018).

Finalmente, una vez testada la usabilidad, se puso en marcha el (AC) CLOE a comienzos del año 2020 para el alumnado de las titulaciones del Grado de Educación Social, Educación Infantil y el Grado de Primaria. En total estuvo disponible para 589 estudiantes. Se inició una campaña de difusión a través de la plataforma virtual y se recogió la interacción del agente virtual durante seis meses, hasta junio de 2020, con un total de 3025 interacciones que también se describen en los resultados de este estudio.

2.7. Análisis de datos

Para los análisis de los datos se utilizaron los paquetes estadísticos IBM SPSS 20.0 y AMOS 21.0. Se analiza la estructura empírica y la validez de la Escala aplicada mediante indicadores y técnicas de análisis multivariantes, aunque no constituya un propósito fundamental de estudio, ya que no pretendemos confirmar la validez de constructo de una Escala ya estandarizada, específica y validada (Babu et al., 2011; Brooke, 1996; Brooke, 2013; Finstad, 2010; Lewis & Sauro, 2009; Ghosh et al., 2018; Neumann et al., 2019; Schroeder et al., 2018; Sinoo et al., 2018). Sin embargo, interesa comprobar si la matriz de correlaciones de los ítems puede ser factorizada y, por tanto, someterse a análisis factorial exploratorio (AFE), utilizando para ello el método de estimación de factores Mínimos Cuadrados Ordinarios (MCO), en su vertiente de Mínimos Cuadrados No Ponderados (USL); se considera este método cuando se trabaja con muestras relativamente pequeñas, el número de factores a retener es pequeño y evita saturaciones mayores que 1 y varianzas de error negativas (Jung, 2013). Asimismo, para verificar el modelo empleado, este se analiza mediante indicadores o índices *ad hoc* de ajuste de la bondad (Hair et al., 1998; 2005).

Se estudia la consistencia interna del cuestionario aplicado a la muestra objeto de estudio para poder considerar este instrumento de evaluación fiable. Para el cálculo de la fiabilidad se ha utilizado el estadígrafo α de Cronbach, siendo un buen índice o coeficiente de fiabilidad lo suficientemente bueno o determinante a partir de un valor de .70 (Nunnally, 1978; Kapplan & Sacuzzo, 2009). En el caso de que se obtenga más de un constructo en el Análisis Factorial, es preciso realizar el cálculo del Índice de la Fiabilidad Compuesta (Fornell & Larcker, 1981); este se interpreta como el α de Cronbach, y tiene en cuenta las interrelaciones de los constructos o factores extraídos. Este estadístico también mide el índice de Varianza Extraída (IVE), que muestra la relación entre la varianza de un factor (j) y la varianza total (Fornell & Larcker, 1981). El IVE debe ser superior a .50, lo que se traduce que más del 50% de la varianza del constructo es debida a sus indicadores.

Se realizan análisis estadísticos univariados o descriptivos que resumen las características del conjunto de la muestra. Además, para estudiar las relaciones que se establecen entre las variables de estudio se emplea el coeficiente "r" de Pearson. Asimismo, se llevan a cabo análisis diferenciales ($p < .05$) mediante la prueba de Anova de contrastes para comparaciones múltiples "post hoc" de Bonferroni y la prueba "t" de Levene para la igualdad de varianzas en muestras independientes.

Se emplea el análisis de regresión lineal múltiple para comprobar si el nivel de conocimiento y el grado de uso de chatbots se pueden considerar variables dependientes y, como variables independientes o predictoras, los ítems que configuran la Escala SUS.

Finalmente, debemos mencionar que se ha realizado un análisis de agrupación de las alternativas de respuesta en las preguntas del cuestionario referidas a "puntúa tu nivel de conocimiento sobre chatbots" y "puntúa tu grado de uso de chatbots" en dos categorías: valores bajos y valores altos. Esta categorización de las opciones de respuesta nos permite disponer de dos grupos para poder efectuar los análisis diferenciales y relacionales de manera eficaz, uno con valor bajo, agrupando las opciones de respuesta "Nada" y "Poco", y otro con valores altos, agrupando las opciones de respuesta "Suficiente", "Bastante" y "Mucho".

3. Análisis y resultado

3.1. Análisis factorial

El análisis factorial realizado originariamente de la Escala SUS reveló 2 factores (Brooke, 1986): un factor, con 8 ítems, relacionado con la dimensión usabilidad y, un segundo factor, conformado por 2 ítems referido al aprendizaje. En la actualidad, investigaciones a este respecto (Bangor et al., 2008; Sauro & Lewis, 2009; Tullis & Stetson, 2004) han asumido que la Escala presenta una validez de constructo unidimensional, ya que evalúa un único factor significativo, el constructo de usabilidad (Lewis & Sauro, 2009, p. 96).

Nuestra investigación revela una serie de indicadores que verifican la pertinencia y validez de realizar el AFE: prueba de Kolmogorov-Smirnov (Sig= p> .05) y prueba de Shapiro-Wilk (Sig= p> .05); coeficiente KMO= .869, prueba de esfericidad de Barlett p= 0,000 y determinante= .043. Además, la matriz de correlaciones anti-imagen (Tabla 1) presenta valores en su diagonal por encima de .5, muy cercanos a 1, por lo que se constata que la matriz de correlaciones puede ser factorizada y el AFE puede ser realizado.

Tabla 1.

Matriz de correlación anti-imagen

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6	ITEM 7	ITEM 8	ITEM 9	ITEM 10
ITEM 1	**.874**[(a)]	-.104	-.108	-.106	-.165	-.205	-.064	.074	-.264	.006
ITEM 2	-.104	**.924**[(a)]	-.173	.048	.024	-.059	-.055	-.100	-.062	-.022
ITEM 3	-.108	-.173	**.913**[(a)]	.086	-.092	-.125	-.173	-.163	.000	.087
ITEM 4	-.106	.048	.086	**.693**[(a)]	-.033	.018	.077	-.016	.090	-.406
ITEM 5	-.165	.024	-.092	-.033	**.877**[(a)]	-.369	-.102	-.062	-.048	-.016
ITEM 6	-.205	-.059	-.125	.018	-.369	**.875**[(a)]	-.168	-.077	-.062	-.029
ITEM 7	-.064	-.055	-.173	.077	-.102	-.168	**.909**[(a)]	-.185	-.095	.184
ITEM 8	.074	-.100	-.163	-.016	-.062	-.077	-.185	**.877**[(a)]	-.301	-.018
ITEM 9	-.264	-.062	.000	.090	-.048	-.062	-.095	-.301	**.875**[(a)]	-.020
ITEM10	.006	-.022	.087	-.406	-.016	-.029	.184	-.018	-.020	**.699**[(a)]

Nota: [(a)] Medida de adecuación muestral

El USL extrae dos factores que explican el 55,73% de la varianza total; este dato coincide con estudios e investigaciones afines cuyos resultados convergen en una solución de dos factores que representan el 55%-58% de la varianza total (Lewis & Sauro, 2009, pp.98; Bangor et al, 2008). Por tanto, la bondad de la solución factorial descansa en 2 factores que explican el 41,46% y el 14,27% respectivamente. La estructura factorial final se refleja en la *Tabla 2*.

Los datos obtenidos señalan que la solución factorial generada es parsimoniosa. Los resultados han confirmado que la estructura empírica del cuestionario aplicado goza de una buena calidad psicométrica. Se han extraído 2 factores responsables del 55,73% de la varianza total y, aunque los dos factores posean saturaciones factoriales altas y sean independientes entre sí, compartimos con Bangor et al. (2008), Finstad (2010), Lewis y

Sauro (2009) y Tabachink y Fidell (1989), que esta Escala de usabilidad de 10 elementos, desde su introducción en 1986, tiene carácter unidimensional.

Tabla 2.

Matriz de componentes rotados[a]

ITEMS	Componente	
	1	2
ITEM 6: Creo que la herramienta es consistente (coherente)	**.757**	
ITEM 5: Creo que las diversas funciones en esta herramienta estaban bien integradas	**.685**	
ITEM 1: Creo que me gustaría usar esta herramienta con frecuencia	**.668**	
ITEM 7: Creo que la mayoría de la gente aprendería a usar esta herramienta muy rápidamente	**.637**	
ITEM 9: Me sentiría muy seguro/ a usando la herramienta	**.629**	
ITEM 3: Creo que la herramienta es fácil de usar	**.619**	
ITEM 8: Encuentro la herramienta muy intuitiva	**.599**	
ITEM 2: Encuentro la herramienta bastante simple	**.457**	
ITEM 10: Necesitaría aprender muchas cosas antes de empezar a utilizar esta herramienta		**.668**
ITEM 4: Creo que necesitaría el apoyo de una persona técnica para poder usar esta herramienta		**.664**

Nota: Método de extracción: Mínimos cuadrados no ponderados. Método de rotación: Normalización Varimax con Kaiser.

[a] La rotación ha convergido en 3 iteraciones.

En cuanto a los índices de ajuste de la bondad del modelo se obtienen los siguientes resultados: En primer lugar, Chi-cuadrado≠ 0, no se rechaza la hipótesis nula, lo que indica ajuste o idoneidad de los datos; índices de ajustes de bondad: GFI= .88 y AGFI= .898; índice de ajuste comparado: CFI= .953; índice de ajuste no normativo: NNFI= 0,913; índice de error de aproximación cuadrático medio: RMSEA=.062; índice de residuo cuadrático medio: RMR= .051. Estos resultados muestran y verifican la bondad del modelo aplicado, indicativo que la Escala presenta validez de constructo.

3.2. Análisis de fiabilidad

Los estudios realizados de la Escala SUS sugieren coeficientes de fiabilidad que oscilan entre α de Cronbach = .70 - .92 (Bangor et al., 2008; Lucey, 1991; Lewis & Sauro, 2009; Kirakowski, 1994). En este sentido, y siguiendo el procedimiento de cálculo de inter-correlación de elementos, la consistencia interna de la Escala SUS de nuestro estudio es α de Cronbach= .808, un coeficiente de fiabilidad bastante aceptable. Por su parte, el coeficiente de fiabilidad para la Dimensión 1 fue de .852, mientras que la Dimensión 2 obtuvo un Alfa de .703. Este último dato coincide con el obtenido por Lewis y Sauro (2009) en el que, a pesar de que el Factor contiene dos elementos, la dimensión ofrecía la suficiente confiabilidad para cumplir con el estándar mínimo de 0.70 para este tipo de medición (Landauer, 1997; Nunnally, 1978).

Los datos muestran que existe más de un constructo y, por tanto, es necesario analizar el Índice de la Fiabilidad Compuesta (IFC) y el de la Varianza Extraída (IVE). Así se obtiene

que para la Dimensión 1: α de Cronbach= .852; IFC= .889; IVE= .714. Para la Dimensión 2: α de Cronbach= .703; IFC= .737; IVE= .497. Podemos ver que la fiabilidad compuesta es algo más alta en ambas dimensiones que en el coeficiente de Cronbach; sin embargo, mientras que el IVE en la Dimensión 1 refleja que el 71% de la varianza de este constructo se debe a sus elementos, el índice de la Dimensión 2 está ligeramente por debajo del valor .5, lo que indica que está en el límite para ser recomendable.

3.3. Análisis descriptivo de la escala SUS

En la *Tabla 3* se presentan los estadísticos de tendencia central y de dispersión: las medias y las desviaciones estándar para cada uno de los ítems del cuestionario. Asimismo, se muestran también los resultados relativos al sumatorio de las opciones de respuesta en dos categorías de acuerdo, en valores porcentuales, para cada uno de los ítems de la escala. Esto es, se organizan las respuestas agrupándolas según las alternativas: valor alto (opciones "de acuerdo" y "totalmente de acuerdo") y valor bajo (opciones "totalmente en desacuerdo", "en desacuerdo", "ni de acuerdo ni en desacuerdo").

Tabla 3.

Relación y distribución de estadísticos descriptivos de los ítems de la Escala SUS

ITEMS	N	Media[(1)]	D. S.	%Valores altos[(2)]	%Valores bajos[(3)]
1.- Creo que me gustaría usar esta herramienta con frecuencia	303	3.96	.984	70.9	29.1
2.- Encuentro la herramienta bastante simple	303	3.96	1.065	71.3	28.7
3.- Creo que la herramienta es fácil de usar	302	4.39	.839	86.5	13.5
4.- Creo que necesitaría el apoyo de una persona técnica para poder usar esta herramienta	303	1.66	.946	5.6	94.4
5.- Creo que las diversas funciones en esta herramienta estaban bien integradas	301	4.05	.837	74.7	25.3
6.- Creo que la herramienta es consistente (coherente)	303	4.22	.778	82.2	17.8
7.- Creo que la mayoría de la gente aprendería a usar esta herramienta muy rápidamente	303	4.44	.702	91	9
8.- Encuentro la herramienta muy intuitiva	303	4.07	.918	75.6	24.4
9.- Me sentiría muy seguro/ a usando la herramienta	303	4.08	.903	75.3	24.7
10.- Necesitaría aprender muchas cosas antes de empezar a utilizar esta herramienta	303	1.93	1.150	12.2	87.8

Nota: [(1)] Escala: "1" Totalmente en desacuerdo; "2" En desacuerdo; "3" Ni en acuerdo ni en desacuerdo; "4" De acuerdo; "5" Totalmente de acuerdo. [(2)] Valores altos: "de acuerdo" y "totalmente de acuerdo" [(3)] Valores bajos: "totalmente en desacuerdo". "en desacuerdo". "ni de acuerdo ni en desacuerdo"

Si atendemos los resultados expuestos, estos señalan que los estudiantes calificaron de forma muy favorable todos los ítems de la Escala. De este modo, vemos, por ejemplo, que el valor más alto se sitúa en el ítem 7: "*Creo que la mayoría de la gente aprendería a usar esta herramienta muy rápidamente*", con $\bar{X}$= 4,44 y un valor porcentual de acuerdo del

91%; le sigue el ítem 3: "*Creo que la herramienta es fácil de usar*", con una puntuación media de 4,39 y un porcentaje satisfactorio del 86,5%. Los datos invitan a pensar que la usabilidad y el manejo de esta herramienta es bastante sencilla y fácil de utilizar.

Las puntuaciones más bajas se alcanzaron en el ítem 4: "*Creo que necesitaría el apoyo de una persona técnica para poder usar esta herramienta*" ($\bar{x}$= 1,66; valor bajo= 94,4%) y en el ítem 10: "*Necesitaría aprender muchas cosas antes de empezar a utilizar esta herramienta*", con una puntuación media de 1,93 y un valor porcentual bajo de 87,8%. No obstante, se debe decir que, a pesar de obtener valores bajos, el enunciado de las preguntas está redactado de forma positiva, aunque el contenido del mismo sea negativo, lo que coincide con los propósitos del enunciado original (Lewis & Sauro, 2009, p. 94).

En líneas generales, los resultados obtenidos muestran un alto grado de satisfacción de los estudiantes. Por tanto, podemos considerar que la experiencia con el chatbot (CLOE) ha sido altamente positiva dadas las puntuaciones notables que han emitido los alumnos tras el uso de esta tecnología.

3.4. Análisis relacional de la Escala SUS

La *Tabla 4* presenta las distintas correlaciones entre los ítems que conforman la Escala SUS. Como se observa, existe una relación lineal positiva entre los ítems 1, 2, 3, 5, 6, 7, 8 y 9, estos es, aquellos ítems que estructuran el primer factor extraído; así, por ejemplo, el ítem 5 ("Creo que las diversas funciones en esta herramienta estaban bien integradas") y el ítem 6 ("Creo que la herramienta es consistente (coherente)") presentan una correlación lineal significativamente positiva (r= .606), lo que indica que estos miden similares características. Del mismo modo, pero a la inversa (r< 0), existen relaciones lineales negativas entre aquellos ítems que componen el segundo factor, ítems 4 y 10, con el resto de los ítems de la Escala. Por ejemplo, la relación entre el ítem 10 ("Necesitaría aprender muchas cosas antes de empezar a utilizar esta herramienta") y el ítem 7 ("Creo que la mayoría de la gente aprendería a usar esta herramienta muy rápidamente") es negativa (r= - .319), indicativo de que miden diferentes características. Por supuesto, la correlación es r>0 entre los ítems 4 y 10.

Tabla 4.

Correlaciones entre los ítems de la Escala SUS

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10
ITEM1		**.305(**)**	**.406(**)**	-.051	**.476(**)**	**.519(**)**	**.402(**)**	**.318(**)**	**.477(**)**	-.102
ITEM2	**.305(**)**		**.374(**)**	**-.141(*)**	**.258(**)**	**.310(**)**	**.318(**)**	**.317(**)**	**.303(**)**	-.110
ITEM3	**.406(**)**	**.374(**)**		**-.241(**)**	**.424(**)**	**.481(**)**	**.507(**)**	**.443(**)**	**.382(**)**	**-.252(**)**
ITEM4	-.051	**-.141(*)**	**-.241(**)**		-.092	**-.139(*)**	**-.269(**)**	**-.157(**)**	**-.182(**)**	**.468(**)**
ITEM5	**.476(**)**	**.258(**)**	**.424(**)**	-.092		**.606(**)**	**.445(**)**	**.377(**)**	**.392(**)**	**-.114(*)**
ITEM6	**.519(**)**	**.310(**)**	**.481(**)**	**-.139(*)**	**.606(**)**		**.505(**)**	**.420(**)**	**.432(**)**	**-.142(*)**
ITEM7	**.402(**)**	**.318(**)**	**.507(**)**	**-.269(**)**	**.445(**)**	**.505(**)**		**.474(**)**	**.442(**)**	**-.319(**)**
ITEM8	**.318(**)**	**.317(**)**	**.443(**)**	**-.157(**)**	**.377(**)**	**.420(**)**	**.474(**)**		**.501(**)**	**-.156(**)**
ITEM9	**.477(**)**	**.303(**)**	**.382(**)**	**-.182(**)**	**.392(**)**	**.432(**)**	**.442(**)**	**.501(**)**		**-.145(*)**
ITEM10	-.102	-.110	**-.252(**)**	**.468(**)**	**-.114(*)**	**-.142(*)**	**-.319(**)**	**-.156(**)**	**-.145(*)**	

Nota: ** La correlación es significativa al nivel 0,01 (bilateral). * La correlación es significante al nivel 0,05 (bilateral).

3.5. Análisis diferencial

Si comparamos las variables medidas en los estudiantes por razón del género, del Grado que cursan, del nivel de conocimiento de chatbots y del grado de uso de chatbots, se han obtenido diferencias de medias estadísticamente significativas en las siguientes variables:

En función del género

Se dan diferencias significativas (IC 99,95%) en el ítem 1: *Creo que me gustaría usar esta herramienta con frecuencia* (F 301= .149; p< .020), en el ítem 4: *Creo que necesitaría el apoyo de un técnico para poder usar esta herramienta* (F 301= .427; p< .004), en el ítem 5: *Creo que las diversas funciones en esta herramienta estaban bien integradas* (F 299= .078; p< .020) y en el ítem 6: *Creo que la herramienta es consistente* (F 301= 2.325; p< .009). Los resultados del análisis diferencial se muestran en la *Tabla 5*.

Tabla 5.

Prueba de muestras independientes "T" de Levene

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
								Superior	Inferior
ITEM1	.149	.700	-2.334	301	**.020**	-.287	.123	-.528	-.045
ITEM2	3.514	.062	1.094	301	.275	.146	.134	-.117	.410
ITEM3	.007	.933	-1.181	300	.238	-.125	.105	-.332	.083
ITEM4	.427	.514	2.906	301	**.004**	.341	.117	.110	.572
ITEM5	.078	.781	-2.346	299	**.020**	-.245	.105	-.451	-.040
ITEM6	2.325	.128	-2.644	301	**.009**	-.256	.097	-.447	-.065
ITEM7	.566	.452	-1.347	301	.179	-.119	.088	-.292	.055
ITEM8	1.244	.266	-.265	301	.791	-.031	.116	-.258	.197
ITEM9	1.378	.241	-1.836	301	.067	-.208	.113	-.430	.015
ITEM10	2.524	.113	1.339	301	.181	.193	.144	-.091	.478

Nota:* La Diferencia de Medias es significativa al nivel p< .05

Como se puede ver, se dan diferencias de medias (p (t) < .05) en función del género; por tanto, en la Hipótesis 3 se rechaza la hipótesis nula y se acepta la hipótesis alternativa, H1: *Existen diferencias estadísticamente significativas en función del género respecto a la valoración de la usabilidad de los chatbots*.

En función del Grado

Se han encontrado diferencias significativas (IC 99,95%) en todos los ítems. Para una mejor lectura del análisis realizado, la *Tabla 6* refleja las puntuaciones medias y la desviación típica obtenidas por los distintos grupos, así como los índices de significación en cada uno de los ítems.

Tabla 6.

Pruebas post hoc. Comparaciones múltiples

			ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10
GRADO	PRIMARIA	Media	**4.01**	**3.99**	**4.50**	**1.62**	**4.10**	**4.27**	**4.49**	**4.51**	**4.09**	**1.91**
		N	244	244	243	244	242	244	244	244	244	244
		D. t.	.927	1.058	.729	.902	.819	.737	.657	.875	.882	1.124
	ED. SOCIAL	Media	**3.23**	**3.43**	**3.75**	**2.05**	**3.65**	**3.75**	**4.05**	**3.60**	**3.68**	**2.43**
		N	40	40	40	40	40	40	40	40	40	40
		D. t.	1.074	1.059	1.032	1.108	.975	.927	.904	1.105	.997	1.318
	INFANTIL	Media	**4.79**	**4.74**	**4.26**	**1.37**	**4.26**	**4.63**	**4.58**	**4.74**	**4.74**	**1.11**
		N	19	19	19	19	19	19	19	19	19	19
		D. t.	.419	.452	1.098	.955	.452	.496	.507	.452	.452	.315
ANOVA		F	20.622	10.874	15.375	4.647	5.778	11.176	7.531	10.986	9.603	9.072
		gl	300	300	299	300	298	300	300	300	300	300
		Sig	**000**	**.000**	**.000**	**.010**	**.003**	**.000**	**.001**	**.000**	**.000**	**.000**

Nota: * *Bonferroni. La diferencia de medias es significativa al nivel p< .05*

Se aprecia que existen diferencias de medias (p (t) < 0,05) en función del Grado; por tanto, en la Hipótesis 4 se rechaza la hipótesis nula y se acepta la hipótesis alternativa, H1: *Existen diferencias estadísticamente significativas en función del Grado respecto a la valoración de la usabilidad de los chatbots*.

En función del nivel de conocimiento de chatbots

Existen diferencias significativas (IC 99,95%) en el ítem 1: *Creo que me gustaría usar esta herramienta con frecuencia* (F 299= 4.558; p<.015), en el ítem 3: *Creo que la herramienta es fácil de usar* (F 298= 5.594; p<.018) y en el ítem 10: *Necesitaría aprender muchas cosas antes de empezar a utilizar esta herramienta* (F 299= 11.260; p<.000). La *Tabla 7* muestra los resultados del análisis diferencial.

Tabla 7.

Prueba de muestras independientes "T" de Levene

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
								Superior	Inferior
ITEM1	4.558	.034	-2.436	299	**.015**	-.275	.113	-.497	-.053
ITEM2	2.637	.105	-1.681	299	.094	-.204	.121	-.443	.035
ITEM3	5.594	.019	-2.370	298	**.018**	-.228	.096	-.417	-.039
ITEM4	.801	.372	1.154	299	.249	.125	.108	-.088	.337
ITEM5	.383	.537	-1.039	297	.300	-.101	.097	-.292	.090
ITEM6	.488	.485	-.317	299	.751	-.029	.090	-.206	.149
ITEM7	3.342	.069	-1.764	299	.079	-.143	.081	-.302	.016
ITEM8	.080	.778	-.397	299	.692	-.042	.106	-.251	.167
ITEM9	.228	.633	-1.188	299	.236	-.124	.104	-.329	.081
ITEM10	11.260	.001	3.965	299	**.000**	.515	.130	.260	.771

**La Diferencia de Medias es significativa al nivel p<.05*

En función del grado de uso de chatbots

Se dan diferencias significativas (IC 99,95%) en el ítem 1: *Creo que me gustaría usar esta herramienta con frecuencia* (F 295= 4.012; p<.011) y en el ítem 2: *Encuentro la herramienta bastante simple* (F 295= .616; p<.016). En la *Tabla 8* figuran los datos que hemos alcanzado en este análisis de contraste.

3.6. análisis de regresión

Se realiza un primer análisis de regresión donde los ítems de la Escala SUS son las variables explicativas-predictoras de la variable "Nivel de conocimiento sobre chatbots" (Tabla 9). Un segundo análisis recae en la variable "Grado de uso de chatbots" (Tabla 10). En el caso del primer análisis, de las diez variables, dos de ellas presentan índices de regresión estandarizados significativos: el ítem 1: *Creo que me gustaría usar esta herramienta con frecuencia* y el ítem 10: *Necesitaría aprender muchas cosas antes de empezar a utilizar esta herramienta*.

Tabla 8.

Prueba de muestras independientes "T" de Levene

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
								Superior	Inferior
ITEM1	4.012	.046	-2.560	295	**.011**	-.355	.139	-.627	-.082
ITEM2	.616	.433	-2.422	295	**.016**	-.357	.148	-.648	-.067
ITEM3	3.915	.049	-1.576	294	.116	-.186	.118	-.419	.046
ITEM4	.362	.548	.691	295	.490	.091	.132	-.169	.352
ITEM5	.779	.378	-1.187	293	.236	-.141	.119	-.376	.093
ITEM6	.459	.499	-1.545	295	.123	-.171	.110	-.388	.047
ITEM7	1.668	.197	-.354	295	.724	-.035	.100	-.231	.161
ITEM8	2.608	.107	-.991	295	.322	-.129	.130	-.385	.127
ITEM9	.060	.807	-.906	295	.366	-.116	.128	-.368	.136
ITEM10	.034	.855	1.305	295	.193	.213	.163	-.108	.534

Nota: *La Diferencia de Medias es significativa al nivel p<.05

Tabla 9.

Coeficientes (a) del Análisis de regresión lineal de la Variable Nivel de conocimiento sobre chatbots con la Escala SUS

Modelo	B	t	p	R^2	F
2 (Constante)	0.22	29.946	0.000	0.56	1.765
10.- Necesitaría aprender muchas cosas antes de empezar a utilizar esta herramienta	0.21	10.861	0.000	0.63	1.671
1.- Creo que me gustaría usar esta herramienta con frecuencia	0.11	2.046	0.042	0.49	0.115

Nota: (a)Variable dependiente: Nivel de conocimiento sobre chatbots

El criterio o indicador para comprobar el modelo explicativo es el valor del coeficiente de determinación (R^2), que estima la proporción de la variable en la variable dependiente. En nuestro caso, R^2= .56 y, por tanto, el grado de explicación de nuestro modelo= 56% (R^2= 56% F (1.765) = 20,94, p< .000); es decir, estos resultados sugieren que el 56% de la variación de la varianza en la variable dependiente "Nivel de conocimiento" puede ser explicada por la combinación lineal de estos dos ítems incluidos en el modelo. Se verifica la Hipótesis alternativa (H1): *Existe relación entre la satisfacción por el uso de esta herramienta con el nivel de conocimiento de los chatbots*.

Por último, en el segundo análisis de regresión, solo el ítem 2 queda incluido en el modelo de regresión, con lo cual, es la única variable relacional explicativa de la variable

dependiente “Grado de uso de chatbots”. El grado de explicación de este modelo indica que el 27% de la variación en la variable dependiente puede ser explicada por este ítem.

Tabla 10.

Coeficientes (b) del Análisis de regresión lineal de la Variable Grado de uso de chatbots con la Escala SUS

Modelo	B	t	p	R^2	F
1 (Constante)	0.959	10.252	0.000	0.27	7.988
2.- Encuentro la herramienta bastante simple	0.64	2.826	0.005		

Nota: (b) Variable dependiente: Grado de uso de chatbots

No se rechaza, sino que se acepta la hipótesis nula (H_0), no existe relación entre la adquisición de aprendizajes previos para utilizar la herramienta con el grado de uso de chatbots.

3.7. Uso real del agente conversacional CLOE

Entre los meses de enero y junio del año 2020 se puso en marcha el agente virtual para 589 estudiantes de las titulaciones del Grado de Educación Social, Infantil y el Grado de Primaria. El (AC) contestó un total de 3025 cuestiones repartidas entre los 6 meses, teniendo un pico de 912 preguntas contestadas en febrero, al coincidir con la presentación del compromiso entre profesorado y alumnado, y una media de 504 respuestas por mes (Tabla 11).

En cuanto al tipo de respuestas, destacó por contestar a cuestiones relativas a la citación de autores en los trabajos (852) y a la correcta referencia en formato APA (808). El alumnado también hizo uso de la conversación ligera del agente virtual (510), que retornaba con sus respuestas al tema para el que fue creado. Otras preguntas recurrentes que resolvió hacían referencia a los tipos de TFG y sus características (208), clarificando cuestiones específicas sobre aspectos concretos de las distintas modalidades de trabajo. Algo que se complementó con respuestas a los aspectos formales del TFG (222), relacionadas con el formato, la sangría, la extensión o el número de páginas entre otros.

4. Discusión y conclusiones

Los resultados nos permiten conocer mejor las características que pueden llegar a configuran el manejo y la utilidad de esta tecnología en los entornos académicos. Para ello, ha sido necesaria una revisión exhaustiva del agente conversacional por parte de las coordinadoras de la titulación, debido a la gran cantidad de información que se necesita para que la herramienta para responder a todas las preguntas que se le pueden realizar (Fast et al., 2018; Sumikawa et al., 2020). Los datos revelan una alta satisfacción de los estudiantes respecto al empleo de esta herramienta, por tanto, podemos afirmar que la experiencia ha sido notable para los estudiantes. En cierta medida, la interacción, y la respuesta inmediata a la hora de resolver dudas, pueden facilitar una visión positiva hacia la comunicación con el agente virtual (Wargnier et al., 2015).

Tabla 11.

Tipos de respuestas contestadas por el agente virtual implementado

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Total
1. Aspectos formales	2	61	20	24	53	62	222
2. Tipos y características de TFG	14	78	35	27	24	30	208
3. Citación	25	232	136	12	167	150	852
4. Referencias APA	16	104	210	132	180	166	808
5. Documentación necesaria	17	76	24	3	14	14	148
6. Fechas de entrega y defensa	14	39	8	6	8	7	82
7. Calificaciones	28	74	3	7	6	10	128
8. Docentes: tribunales y tutores	24	38	2	1	0	2	67
9. Conversación natural	69	210	89	51	52	39	510
Total	209	912	527	393	504	480	3025

Asimismo, se confirma que existen diferencias significativas en las variables analizadas: género, Grado, Nivel de conocimiento y grado de uso de los chatbots. Dichas diferencias han sido más sobresalientes en función del Grado que cursan los estudiantes. En este sentido, se ha podido comprobar que los estudiantes de Educación Primaria y Educación Infantil comparten criterios valorativos similares respecto a la usabilidad de la herramienta. No ocurre lo mismo con los estudiantes de Educación Social. Estos últimos difieren en la perspectiva que tienen respecto a los chatbots y su manejo; aunque su visión y creencias también sean positivas, sus puntuaciones medias obtenidas en los diferentes ítems que componen la Escala revelan que se posicionan entre las opciones de respuesta “ni de acuerdo ni en desacuerdo” y “de acuerdo”, mientras que las de los otros estudiantes se ubican en las alternativas más altas. Ejemplo de ello lo encontramos en el ítem 8: “*Encuentro la herramienta muy intuitiva*”; los alumnos de Ed. Primaria y Ed. Infantil obtienen una $\bar{x} = 4.51$ y $\bar{x} = 4.74$ respectivamente, frente a una $\bar{x} = 3.60$ que alcanzan los estudiantes de Educación Social ($F\ 300 = 10.986$; $p<.000$).

Respecto al análisis factorial efectuado, la estructura empírica y la consistencia interna de la Escala es similar a la que se han estudiado y evaluado en otras investigaciones análogas. Los coeficientes estadísticos verifican la pertinencia de realizar el AFE, la bondad de la matriz de correlaciones para ser factorizada, los índices de ajuste del modelo y que la solución factorial sea parsimoniosa, requisitos metodológicos y psicométricos condicionantes para aceptar la validez de constructo. La solución generada descansa en dos factores que explican el 55,73% de la varianza que presentan cargas factoriales altas. A este respecto, las correlaciones realizadas inter-elementos señalan relaciones positivas entre los ítems que componen el factor 1 y los ítems del factor 2, siendo lineal y significativamente negativas entre los ítems de cada factor, indicativo de que son independientes entre sí. No obstante, lo indicado, compartimos que la Escala tiene un carácter unidimensional (Bangor et al, 2008; Finstad, 2010; Lewis & Sauro, 2009), lo que facilita su aplicabilidad.

El análisis de regresión múltiple efectuado sobre las diez variables de la Escala sobre el nivel de conocimiento y grado de uso de los chatbots muestra que tres de esas diez variables se relacionan con las variables dependientes analizadas. Estas variables

explicativas-predictivas tienen que ver con la *satisfacción* por usar la herramienta con más frecuencia, la *sencillez* del manejo de esta tecnología y la *necesidad de aprendizajes previos* antes de comenzar a usar el chatbot. Por tanto, se pueden considerar elementos predictores, capaces de estar en la base que sustenta la usabilidad de los chatbots. El uso del agente conversacional está determinado por la temática para la que fue creado. Aunque el diseño está condicionado por la complejidad semántica del tópico (Herbert & Kang, 2018) puede ser una ventaja que el tema sea cerrado, ya que los conceptos relativos al trabajo de fin de grado eran expresiones que el alumnado conocía y compartía. Compartimos con autores como Wellnhammer et al. (2020) o Colace et al. (2018) en que los agentes virtuales pueden ser una buena herramienta de apoyo para trabajar con el alumnado universitario.

Es importante tener en presente el objetivo del agente virtual para poder hacer una valoración, ya que la elaboración de un agente conversacional es compleja (Sumikawa et al., 2020). Es algo que fue básico para plantear el diseño de este chatbot, ya que debía responder a numerosas cuestiones del alumnado. En este caso, el agente tenía unas categorías iniciales que fueron matizándose y ampliándose después de ser probados por los coordinadores de las diferentes titulaciones de Grado. Una vez testado, y después de una entrevista, se desarrolló una base de datos para dar respuesta a 174 tipos de preguntas diferentes.

Sin embargo, el número de cuestiones fue superior puesto que tenían que diseñarse diferentes formas de reconocer una misma pregunta, es decir, de formularlas de manera distinta. Los (AC) deben utilizar composiciones sintácticas y semánticas de gran complejidad, de manera que puedan interactuar con las expresiones de las personas con las que interactúan en el proceso de construcción de la conversación (Herbert & Kang, 2018). Los (AC) presentan limitaciones en la comunicación, tal y como la desarrollan las personas, pudiendo simular conversaciones en un tema determinado sin que puedan desarrollar diálogos complejos en diferentes temáticas o ámbitos. En ocasiones, la herramienta puede dar una respuesta errónea cuando el alumnado habla sobre un argumento que tiene varios significados (Colace et al., 2018). Aunque los usuarios pudieron comportarse de manera inadecuada con el agente (Park et al., 2019), fue algo que se observó de manera residual en las interacciones que realizó durante los seis meses de implementación.

Relacionado con las limitaciones que presenta esta investigación en cuanto a la usabilidad, es preciso señalar que la muestra no permite llegar a generalizar y extrapolar los resultados debido a que se necesitaría aplicar la escala a una población universitaria más amplia y diversa mediante muestreo aleatorio estratificado. Por tanto, sería conveniente replicar el estudio con titulaciones, grados, etc., diferentes. Esta limitación, junto a realizar un análisis de segmentación donde se evalúe la usabilidad del chatbot (variable dependiente) con variables categóricas “grupo experimental”-“grupo control” (variables independientes) determinaría futuras investigaciones.

En la literatura hay pocos experimentos que evalúen la usabilidad de los chatbots, por lo que se necesita más esfuerzo en esta línea de investigación (Ren et al., 2019). La base de datos del (AC) puede exportarse para otras titulaciones o universidades, y es algo que podría mejorar el apoyo en tutorías al profesorado. También es interesante para asignaturas que tienen un gran número de alumnos, como las prácticas universitarias en titulaciones similares a las que asistió el agente.

Embodied conversational agents: artificial intelligence for autonomous learning

1.Introduction

Conversational Agents (CA), also known as chatbots, are software packages that can interact with humans using natural language (Dale, 2016). In other words, they are artificial intelligence programmes that can have a specific conversation with people by means of the processing of natural language (Zeni et al., 2019). They are available to a large number of people due to their capacity of interacting with voice and text (Fadhil & Gabrielli, 2017). In order to do this, multiple data are required to train the conversational agent (Fast et al., 2018; Sumikawa et al., 2020).

CAs create a commitment in the conversation that is capable of focusing the attention of those involved in the interaction (Wargnier et al., 2015). This is of interest for research since people tend to react positively when a CA shows slight social attitudes related to expressions, gestures and social signals (Feine et al., 2019).

This process is not exempt from difficulty, since CAs must use highly intricate syntactic and semantic constructions so that they can interact with expressions within a conversation (Herbert & Kang, 2018). CAs can maintain complex conversations on a specific topic, yet they have certain limitations when conversations englobe different topics, or where the user uses expressions that may have various meanings or interpretations depending on the context (Colace et al., 2018).

CAs can be used in many fields since they can do anything from programming medical appointments or restaurant bookings, purchasing travel packages, answering frequently asked questions (FAQs), to improving the mental health of patients, in post-traumatic stress therapies, or improving the internal communication within an organization, among others (Fast et al., 2018; Ghosh et al., 2018; Sáez et al., 2017; Tielman et al., 2017; Schroeder et al., 2018).

1.1.Conversational agents for education processes

In the world of education, there are universities that provide a virtual scenario as additional support to face-to-face lessons. These virtual platforms are part of an extremely important teaching environment, although they usually lack dynamic tools capable of replying to students. Conversational agents can solve this problem by answering the needs of students in a direct interaction (Sumikawa et al., 2020).

Kerly et al. (2008) mention two instances on the use of conversational agents, one that provides learning support, and the other that provides support for self-assessment. Furthermore, the authors argue that it also improves motivation, metacognitive capabilities and the grades achieved by students. Moreover, it has a positive outcome for lecturers since it provides support to teaching and offers data for evaluation.

Furthermore, they are effective when learning another language, as they are based on a productive interactive dialogue with students. Thus, a CA may interact between two students while following conversations through texts (Tegos et al., 2014; Tegos & Demetriadis, 2017). To this regard, students with the same language

level who had used a CA in the interaction between peers achieved much better results. Babu et al. (2011) studied the effects of using CAs in a natural multimodal interaction to teach its users cultural verbal and non-verbal conversation protocols in the culture of southern India. The outcome of this research demonstrated that the participants who trained using CAs had a substantially higher performance than those who learnt from the study guide.

University tutoring support is another potential that CAs have. They could help to solve doubts on presentations to a high number of students, as a source of information and reflexion (Wellnhammer et al., 2020). This option is considered of interest in those scenarios where there is a mass presence of students, for instance, in MOOC courses. In these cases, using CAs to learn JAVA in such a manner that the lecturer also receives feedback in real time (Catalan et al., 2018). They have also been used as a support tool for university students studying the course Computer Science and Computing (Colace et al., 2018).

Conversational Agents are currently being implemented in mobile phone applications for education, however, more research is required on these tools (Hobert & Meyer von Wolff, 2019). On the other hand, a certain level of knowledge is required in order to develop a chatbot for university courses that are not linked to computer sciences. Chatbots have been implemented on many occasions in technical degrees, yet not so frequently in humanistic and social sciences degrees. In order to solve this issue, a participative design could be carried out that would support the needs of lecturers who do not know how to develop them (Neumann et al., 2019).

1.2. Assessment of the use of a conversational agent in educational processes

One key aspect that should be taken into account is the user's assessment after their experience using the CAs efficiently (Ren et al., 2019). In order to analyse to usability of CAs, a widely used and validated instrument called System Usability Scale (SUS) could be used (Brooke, 1996; Brooke, 3013), which has also been studied and validated by other authors (Finstad, 2010; Lewis & Sauro, 2009). With regard to the assessment of digital resources, among many others, this scale has been used to analyse chatbots that allow the improvement of internal communication processes in a company with 8 subjects (Sáez et al., 2017), a study with 12 users on the usability of the chatbot with a voice interface such as Alexa or Siri (Ghosh et al., 2018), or a conversational agent for the treatment of mental conditions (Schroeder et al., 2018). For aspects related to education, the SUS tool has been used to test aspects regarding the usability of a conversational agent to educate children with diabetes with a sample of 21 subjects (Sinoo et al., 2018), teach cultural patterns by means of a chatbot studied on 40 subjects (Babu et al., 2011), to the usability of a chatbot to collect doubts on a subject by 12 university students (Neumann et al., 2019), among others.

The aim of this paper is to evaluate the use of a conversational agent designed to offer support to students who are drafting their dissertation. To this extent, a chatbot was designed, called CLOE, as a tool to tutor students and answer frequently asked questions (FAQ) regarding their dissertation from degrees in Social, Early Years, and Primary Education. It was developed based on the application Dialogflow, which is an API, or Application Programming Interface, for Natural Language Processing (NLP) and natural language comprehension tasks (Zeni et al., 2019). This tool exports the database so that it may be used by other developers and lecturers at different universities.

It was previously tested by the Dissertation coordinators from the different degrees mentioned, since they answer a myriad of questions every year from students. It was also depurated by Dissertation tutors/directors and, finally, an answer tree was established for matters that could be brought together under the following dimensions:

1) Formal aspects of the dissertation such as the source, margins, spacing, etc.
2) Types of dissertation and their features.
3) Aspects related to referencing.
4) Referencing following APA rules.
5) Documents to be submitted to the administration, signed electronically.
6) Dates linked to the presentation, defence, and announcements.
7) Aspects related to qualifications.
8) Varied answers regarding tribunals or tutors.
9) Natural conversation package to humanise the CA's interaction.

Once the CA has been designed, and before its implementation, the aim would be to analyse, through the application of a validated test, the usability of the chatbot, by means of a sample of university students (features, technical aspects, learning and functionality), as an intuitive and communication tool. Moreover, there is also an interest in verifying whether there are any differences based on *gender*, the student's *Degree*, the *level of knowledge on chatbots* and the amount of *use of the chatbots*.

2.Methodology

2.1. Objective

This paper has a double objective. On the one hand, one of the purposes is to analyse perceptions on the usability and use of a chatbot in the teaching-learning processes of university students from different degrees, such as Social Education, Early Years Education and Primary Education. More specifically, it aims at analysing the relation between predictive or independent variables and the usability of the chatbots (dependent variable). Hence, to this extent, this paper analyses whether there are any significant differences based on gender, the student's degree, the amount of knowledge and the use of chatbots. On the other hand, another aspect analysed is whether the conversational agent was used by students as a virtual support tutor.

2.2. Hypothesis

Based on the mentioned objectives, the following hypothesis have been taken into account:

- Hypothesis 1. The satisfaction of using this tool is linked to the amount of knowledge on chatbots.
- Hypothesis 2. Prior learning before using this tool is related to the usage of chatbots.
- Hypothesis 3. The valuation of the usability of chatbots determines the significance based on gender.
- Hypothesis 4. The valuation of the usability of chatbots determines the significance based on the Degree each student is studying.

2.3. Design

A quantitative method, with a descriptive, correlational, and differential design was used in a natural situation.

2.4. Participants

From a quantitative point of view, the mathematical assumptions used to calculate representativity show that, for populations of 100,000 subjects, where the margin of error is 6%, the reliability level is 96% and the level of heterogeneity is 50%, the sample must be composed of 293 subjects. Based on this, the sample of this study, based on criteria of simple random sampling (Cochran & Bouclier, 1980), includes 303 students, chosen at random, from degrees in Social Education (13.2%), Early Years Education (6.3%) and Primary Education (80.5%), all from the University of Las Palmas de Gran Canaria; 70.3% of the participants are women (N=213) and 29.7% are men (N=90), most aged between 18 and 24 years (95%; N=288). In addition, 85.8% consider that students tend to use social networks with frequency. However, when asked about the frequency with which they use chatbots and their knowledge thereof, 76.9% state that they "barely or never" use them, whilst 52.8% of students classify their knowledge in this area as "poor" or "scarce".

2.5. Instrument

It is of importance to know the expectations that students, as users, have on the use of a virtual agent (Ren et al., 2019). In order to assess the usability of this agent, called CLOE, a widely used and validated questionnaire was used called *System Usability Scale (*SUS) (Babu et al., 2011; Brooke, 1996; Brooke, 2013; Finstad, 2010; Lewis & Sauro, 2009; Ghosh et al., 2018; Neumann et al., 2019; Schroeder et al., 2018; Sinoo et al., 2018), including the contributions provided by Lewis (2018). The questionnaire applied had a series of socio-demographical variables where students were asked about their gender, age, and the degree they are studying. This was followed by five questions that are qualified on a scale where 1 means "nothing" and 5 means "a lot". The five questions were:

- Score how much you use social networks.
- Score your knowledge on chatbots.
- Score how much you use chatbots.
- Score your use of tutorials for the dissertation.

- Version of CLOE used.

2.6. Procedure

Studies performed on the use of CAs are based on a limited sample of subjects, between 8 and 40. In order to achieve a broader sample, the lecturers teaching the students of the chosen degrees were contacted in order to distribute the *System Usability Scale* (SUS) questionnaire to a higher number of students.

All participants were informed of the objective of the research and were asked to participate anonymously and voluntarily. To guarantee that all students would be able to answer the questionnaire in the same conditions, a protocol was implemented to test the CA where the questionnaire administrator pointed out where the tool was located on the online platform. They carried out a test run using five frequently asked questions:

1. What types of dissertation can be performed?
2. How do I reference using paraphrasis?
3. How can I reference a book?
4. How do I get the digital certificate to sign my dissertation?
5. What is the validity period of the commitment document between students and lecturers?

The *System Usability Scale* (SUS) questionnaire was introduced to assess the usability of the tool that they had previously been introduced to. The scale was distributed to a total of 303 subjects, which is a wider sample than others used previously, where the test was provided to an average of 8 to 40 students (Babu et al., 2011; Ghosh et al., 2018; Neumann et al., 2019; Sáez et al., 2017; Schroeder et al., 2018; Sinoo et al., 2018).

Finally, once the usability had been tested, the CLOE CA was implemented at the beginning of 2020 for students enrolled in the following degrees: Social Education, Early Years Education, and Primary Education. In total, it was available to 589 students. A dissemination campaign was introduced on the online platform and interaction with the virtual agent was collected over a six-month period, until June 2020, with a total of 3025 interactions, which are also described in the results of this paper.

2.7. Data analysis

In order to analyse data, the following statistic packages were used: IBM SPSS 20.0 and AMOS 21.0. The empirical structure and validity of the scale applied were analysed by means of multivariant indicators and techniques, although this is not one of the main aspects of the paper since the aim is not to confirm the construct validity of a Scale that has already been standardised, specific and validated (Babu et al., 2011; Brooke, 1996; Brooke, 2013; Finstad, 2010; Lewis & Sauro, 2009; Ghosh et al., 2018; Neumann et al., 2019; Schroeder et al., 2018; Sinoo et al., 2018). However, it is interesting to verify whether the correlation matrix of the items could be factorised and, therefore, submitted to an

exploratory factorial analysis (EFA), using for this purpose the Ordinary Least Squared (OLS) factor estimation method, in its Unweighted Least Squares (ULS) aspect. This method is considered when working with relatively small samples, the number of factors to retain is small and it avoids saturations above 1 and negative error variances (Jung, 2013). Similarly, in order to verify the model used, this is analysed using *ad hoc* goodness of fit indicators or indexes (Hair et al., 1998; 2005).

The internal consistency of the questionnaire applied to the sample under study is analysed in order to consider the reliability of this evaluation tool. To calculate reliability, Cronbach's α stratigraphy was used, since this provides a positive reliability index or coefficient which is sufficiently adequate and determining for values above .70 (Nunnally, 1978; Kapplan & Sacuzzo, 2009). In the event of achieving more than one Factorial Analysis construct, the Complex Reliability Index must be calculated (Fornell & Larcker, 1981). This is interpreted in the same manner as Cronbach's α, and takes into account the interrelationships of the extracted constructs or factors. This statistician also measures the Average Variance Extracted Index (AVE), which shows the relationship between the variance of a factor (j) and the total variance (Fornell & Larcker, 1981). AVE must be higher than .50, which means that over 50% of the variance of the construct is because of its indicators.

Univariate or descriptive statistical analysis are performed which summarise the features of the overall sample. In addition, in order to study the relationships that are established between the variables of the study, Pearson's "r" variant is used. Likewise, differential analysis ($p<.05$) are carried out by means of the Anova contrast test for Bonferroni's multiple "post hoc" comparisons and the Levene "t" test for equality of variances in independent samples.

Moreover, a multiple linear regression analysis is used to confirm whether the level of knowledge and the usage of chatbots may be considered as dependent variables and the items on the SUS Scale, as independent or predictive variables.

Last, it is worth mentioning that an analysis has been performed on the aggrupation of the alternative answers given to the questions included in the questionnaire regarding "score your level of knowledge on chatbots" and "score how much you use chatbots" into two categories: low values and high values. This classification of the possible response options provides two groups in order to carry out the differential and relational analysis effectively, one with a low value, grouping the response options "Nothing" and "Little", and another with high values, grouping the response options "Enough", "Quite a lot" and "A lot".

3. Analysis and results

3.1. Factorial analysis

The factorial analysis carried out originally on the SUS Scale revealed 2 factors (Brooke, 1986): one factor, with 8 items, related to the usability dimension and, a second factor, made up of 2 items, on learning. Currently, research on this matter (Bangor et al, 2008; Sauro & Lewis, 2009; Tullis & Stetson, 2004) has concluded that the Scale presents a unidimensional construct validity, since it evaluates a single significant factor, the usability construct (Lewis & Sauro, 2009, p. 96).

This research reveals a series of indicators that verify the pertinence and validity of performing the EFA: Kolmogorov-Smirnov test (Sig= p> .05) and Shapiro-Wilk test (Sig= p> .05); KMO coefficient = .869, Barlett's sphericity test p=0.000 and determiner = .043. In addition, the anti-image correlation matrix (Table 1) contains values in its diagonal well over .5, very close to 1, hence, this confirms that the correlation matrix can be factorised and the EFA performed.

Table 1.

Anti-image Correlation Matrix

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6	ITEM 7	ITEM 8	ITEM 9	ITEM 10
ITEM 1	**.874**[(a)]	-.104	-.108	-.106	-.165	-.205	-.064	.074	-.264	.006
ITEM 2	-.104	**.924**[(a)]	-.173	.048	.024	-.059	-.055	-.100	-.062	-.022
ITEM 3	-.108	-.173	**.913**[(a)]	.086	-.092	-.125	-.173	-.163	.000	.087
ITEM 4	-.106	.048	.086	**.693**[(a)]	-.033	.018	.077	-.016	.090	-.406
ITEM 5	-.165	.024	-.092	-.033	**.877**[(a)]	-.369	-.102	-.062	-.048	-.016
ITEM 6	-.205	-.059	-.125	.018	-.369	**.875**[(a)]	-.168	-.077	-.062	-.029
ITEM 7	-.064	-.055	-.173	.077	-.102	-.168	**.909**[(a)]	-.185	-.095	.184
ITEM 8	.074	-.100	-.163	-.016	-.062	-.077	-.185	**.877**[(a)]	-.301	-.018
ITEM 9	-.264	-.062	.000	.090	-.048	-.062	-.095	-.301	**.875**[(a)]	-.020
ITEM10	.006	-.022	.087	-.406	-.016	-.029	.184	-.018	-.020	**.699**[(a)]

Nota: [(a)] Measure of sample adequation

The ULS extracts two factors that explain 55.73% of the total variance; this data coincides with peer studies and research where the results converge in a two-factor solution that represents 55-58% of the total variance (Lewis & Sauto, 2009, pp.98; Bangor et al, 2008). Therefore, the goodness of the factorial solution lies in 2 factors that explain 41.46% and 14.27% respectively. The final factorial structure is detailed in *Table 2.*

The data achieved shows that the factorial solution created is parsimonious. The results confirm that the empirical structure of the questionnaire applied has an adequate psychometric quality. Two factors responsible for 55.73% of the total variance have been extracted and, albeit the two factors have high factorial saturations and are independent between each other, it is agreed with Bangor et al (2008), Finstad (2010), Lewis & Sauro (2009) and Tabachink & Fidell (1989), that this usability Scale of 10 elements has a unidimensional character.

With regard to the goodness-of-fit indexes of the model, the following results were achieved: First of all, Chi-square ≠ 0, null hypothesis is not rejected, which means adjustment or idealness of data; goodness-of-fit indexes: GFI= .88 and AGFI= .898; comparative fit index: CFI= .953; non-normed fit index: NNFI= 0.913; Root mean square error of approximation: RMSEA= .062; root mean square residue index: RMR= .051. These results show and confirm the goodness of the model applied, an indicator that the Scale has construct validity.

Table 2.

Rotated Components Matrix

ITEMS	Component	
	1	2
ITEM 6: I think the tool is consistent (coherent)	**.757**	
ITEM 5: I think the different functions of this tool are well integrated	**.685**	
ITEM 1: I think I would enjoy using this tool frequently	**.668**	
ITEM 7: I think most people would learn to use this tool very quickly	**.637**	
ITEM 9: I would feel very secure using this tool	**.629**	
ITEM 3: I think the tool is easy to use	**.619**	
ITEM 8: I find the tool very intuitive	**.599**	
ITEM 2: I find the tool quite simple	**.457**	
ITEM 10: There is a lot I would need to learn before using this tool		**.668**
ITEM 4: I think I would need help from an expert to start using this tool		**.664**

Nota: Extraction Method: Unweighted least squares. Rotation method: Varimax normalisation with Kaiser. a Rotation has converged for 3 iterations.

3.2. Reliability analysis

The studies carried out using the SUS scale suggest reliability coefficients that oscillate between Cronbach's α = .70 - .92 (Bangor et al., 2008; Lucey, 1991; Lewis & Sauro, 2009; Kirakowski, 1994). To this regard and following the calculation procedure of the inter-correlation of elements, the internal consistency of the SUS Scale for the hereby research is Chronbach's α = .808, a fairly acceptable reliability coefficient. On the other hand, the reliability coefficient for Dimension 1 was .852, whilst Dimension 2 obtained an alpha of .703. The latter coincides with the data achieved by Lewis and Sauro (2009) where, despite the factor having two elements, the dimension offered sufficient reliability to meet the minimum standard of 0.70 for this type of measures (Landauer, 1997; Nunnally, 1978).

Data show that there are more than one construct and, therefore, there is a need to analyse the Complex Reliability Index (CRI) and the Average Variance Extracted Index (AVE). In order to achieve for Dimension 1: Cronbach's α= .852; CRI= .889; AVE= .714. For Dimension 2: Cronbach's α= .703; CRI=.737; AVE= .497. It can be appreciated how complex reliability is slightly higher in both dimensions than in the Cronbach coefficient; however, whilst the AVE in Dimension 1 shows that 71% of the variance of this construct is due to its elements, the index of Dimension 2 was slightly below the value .5, which demonstrates that it is at the limit of being recommendable.

3.3. Descriptive analysis of the SUS Scale

Table 3 presents the central tendency and dispersion statistics: the means and standard deviations for each of the items of the questionnaire. Likewise, the results for the sum of the answer options in two categories are shown, in percentages, for each of the items on the scale. In other words, the answers are organised by classifying them according to the

alternatives: high value (options “agree” and “totally agree”) and low value (options “totally disagree”, “disagree” and “neither agree nor disagree”).

Table 3.

Relation and distribution of descriptive statistics of the item on the SUS Scale

ITEMS	N	Mean(1)	D. S.	%High value(2)	%Low value
1.- I think I would enjoy using this tool frequently.	303	3.96	.984	70.9	29.1
2.- I find the tool quite simple	303	3.96	1.065	71.3	28.7
3.- I think the tool is easy to use	302	4.39	.839	86.5	13.5
4.- I think I would need support from an expert to use this tool	303	1.66	.946	5.6	94.4
5.- I think the different functions were well integrated in this tool	301	4.05	.837	74.7	25.3
6.-I think the tool is consistent (coherent)	303	4.22	.778	82.2	17.8
7.- I think most people would learn to use this tool very quickly	303	4.44	.702	91	9
8.- I find the tool highly intuitive	303	4.07	.918	75.6	24.4
9.- I would feel sure using this tool	303	4.08	.903	75.3	24.7
10.- There is a lot I would need to learn before using this tool	303	1.93	1.150	12.2	87.8

Note: (1) Scale: “1” Totally disagree; “2” Disagree; “3” Neither agree nor disagree; “4” Agree; “5” Totally agree. (2) High values: “agree” and “totally agree” (3) Low values: “totally disagree”. “disagree”. “neither agree nor disagree”

Based on these results, it is demonstrated that students gave a highly favourable value to each of the items on the Scale. Thus, it can be appreciated, for instance, that the highest value is given for item 7: “*I think most people would learn to use this tool very quickly*”, where $\bar{X}$ = 4.44 and a percentual value of 91%; this is followed by item 3: “*I think the tool is easy to use”*, with an average punctuation of 4.39 and a satisfactory percentage of 86.5%. The data could conclude, therefore, that this tool is quite simple and easy to use.

The lowest punctuations were given in item 4: “*I think I would need support from an expert to use this tool”* ($\bar{X}$ = 1.66; low value= 94.4%), and item 10: *“There is a lot I would need to learn before using this tool*”, with an average punctuation of 1.93 and a low percentual value of 87.8%. Nonetheless, it may be argued that, despite these low values, these two questions are written in positive, albeit the contents thereof being negative, which coincides with the purposes of the original question (Lewis & Sauro, 2009, p.94).

Overall, the results achieved show that students had a high degree of satisfaction. Therefore, it may be considered that the experience with the chatbot (CLOE) was highly positive, based on the high punctuations that students have given after having used this technology.

3.4. Relational Analysis of the SUS Scale

Table 4 presents the different correlations between the items that make up the SUS Scale. As can be appreciated, there is a positive lineal relation between items 1, 2, 3, 5, 6, 7, 8 and 9, this is, those items that structure the first extracted factor; hence, for instance,

item 5 (“I think the diffe rent functions were well integrated in this tool”) and item 6 (“I think the tool is consistent (coherent)”) have a significantly positive lineal correlation (r= .606), which shows that they measure similar characteristics. Likewise, yet conversely (r< 0), there are negative lineal relationships between the items that make up the second factor, items 4 and 10, with the remaining items of the Scale. For instance, the relation between item 10 (“There is a lot I would need to learn before using this tool”) and item 7 (“I think most people would learn to use this tool very quickly”) is negative (r= .319), an indicator that they measure different characteristics. Of course, the correlation between items 4 and 10 is r>0.

Table 4.

Correlation between the ítems of the SUS Scale

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10
ITEM1		**.305(**)**	**.406(**)**	-.051	**.476(**)**	**.519(**)**	**.402(**)**	**.318(**)**	**.477(**)**	-.102
ITEM2	**.305(**)**		**.374(**)**	**-.141(*)**	**.258(**)**	**.310(**)**	**.318(**)**	**.317(**)**	**.303(**)**	-.110
ITEM3	**.406(**)**	**.374(**)**		**-.241(**)**	**.424(**)**	**.481(**)**	**.507(**)**	**.443(**)**	**.382(**)**	**-.252(**)**
ITEM4	-.051	**-.141(*)**	**-.241(**)**		-.092	**-.139(*)**	**-.269(**)**	**-.157(**)**	**-.182(**)**	**.468(**)**
ITEM5	**.476(**)**	**.258(**)**	**.424(**)**	-.092		**.606(**)**	**.445(**)**	**.377(**)**	**.392(**)**	**-.114(*)**
ITEM6	**.519(**)**	**.310(**)**	**.481(**)**	**-.139(*)**	**.606(**)**		**.505(**)**	**.420(**)**	**.432(**)**	**-.142(*)**
ITEM7	**.402(**)**	**.318(**)**	**.507(**)**	**-.269(**)**	**.445(**)**	**.505(**)**		**.474(**)**	**.442(**)**	**-.319(**)**
ITEM8	**.318(**)**	**.317(**)**	**.443(**)**	**-.157(**)**	**.377(**)**	**.420(**)**	**.474(**)**		**.501(**)**	**-.156(**)**
ITEM9	**.477(**)**	**.303(**)**	**.382(**)**	**-.182(**)**	**.392(**)**	**.432(**)**	**.442(**)**	**.501(**)**		**-.145(*)**
ITEM10	-.102	-.110	**-.252(**)**	**.468(**)**	**-.114(*)**	**-.142(*)**	**-.319(**)**	**-.156(**)**	**-.145(*)**	

Note: ** Correlation is significant at level 0.01 (bilateral). * Correlation is significant at level 0.05 (bilateral)

3.5. Differential Analysis

When comparing the variables that have been measured in students based on gender, the degree they are enrolled in, their level of knowledge on chatbots, and the usage of chatbots, significant statistic mean differences have been obtained for the following variables:

Based on gender

There are significant differences (CI 99.95%) for item 1: *I think I would enjoy using this tool frequently* (F 301= .149; p< .020), for item 4: *I think I would need support from an expert to use this tool* (F 301= .427; p< .004), for item 5: *I think the different functions were well integrated in this tool* (F 299= .078; p< .020) and for item 6: *I think the tool is consistent* (F 301= 2.325; p< .009). The results of the differential analysis are detailed in *Table 5*.

Table 5.

Levene`s "T" test of independent simples

	Levene test for equality of variances		T test for the equality of averages						
	F	Sig.	t	gl	Sig. (bilateral)	Mean difference	Typ. Error of the difference	95% Interval of reliability for the difference	
								Higher	Lower
ITEM1	.149	.700	-2.334	301	**.020**	-.287	.123	-.528	-.045
ITEM2	3.514	.062	1.094	301	.275	.146	.134	-.117	.410
ITEM3	.007	.933	-1.181	300	.238	-.125	.105	-.332	.083
ITEM4	.427	.514	2.906	301	**.004**	.341	.117	.110	.572
ITEM5	.078	.781	-2.346	299	**.020**	-.245	.105	-.451	-.040
ITEM6	2.325	.128	-2.644	301	**.009**	-.256	.097	-.447	-.065
ITEM7	.566	.452	-1.347	301	.179	-.119	.088	-.292	.055
ITEM8	1.244	.266	-.265	301	.791	-.031	.116	-.258	.197
ITEM9	1.378	.241	-1.836	301	.067	-.208	.113	-.430	.015
ITEM10	2.524	.113	1.339	301	.181	.193	.144	-.091	.478

Note: * The Mean Difference is significant at level p< .05

As can be appreciated, there are mean differences (a (t) < .05) based on gender; thus, for Hypothesis 3, the null hypothesis is rejected and the alternative hypothesis accepted, H1; *There are statistically significant differences based on gender with regard to the valuation of the usability of the chatbots.*

Based on the degree

Significant differences have been found (CI 99.95%) for all items. In order to better read the analysis performed, *Table 6* shows the mean punctuation and the typical deviation obtained for the different groups, as well as the significance index for each of the items.

Table 6.

Post hoc test. Multiple comparatives

			ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10
DEGRE	PRIMARY	Media	**4.01**	**3.99**	**4.50**	**1.62**	**4.10**	**4.27**	**4.49**	**4.51**	**4.09**	**1.91**
		N	244	244	243	244	242	244	244	244	244	244
		D. t.	.927	1.058	.729	.902	.819	.737	.657	.875	.882	1.124
	SOCIAL ED.	Media	**3.23**	**3.43**	**3.75**	**2.05**	**3.65**	**3.75**	**4.05**	**3.60**	**3.68**	**2.43**
		N	40	40	40	40	40	40	40	40	40	40
		D. t.	1.074	1.059	1.032	1.108	.975	.927	.904	1.105	.997	1.318
	EARLY YEARS	Media	**4.79**	**4.74**	**4.26**	**1.37**	**4.26**	**4.63**	**4.58**	**4.74**	**4.74**	**1.11**
		N	19	19	19	19	19	19	19	19	19	19
		D. t.	.419	.452	1.098	.955	.452	.496	.507	.452	.452	.315
ANOVA		F	20.622	10.874	15.375	4.647	5.778	11.176	7.531	10.986	9.603	9.072
		gl	300	300	299	300	298	300	300	300	300	300
		Sig	**000**	**.000**	**.000**	**.010**	**.003**	**.000**	**.001**	**.000**	**.000**	**.000**

Note: * Bonferroni. The difference of averages is significant at level p< .05

It is appreciated that there are mean differences (p (t) < 0.05) based on the degree; therefore, for Hypothesis 4, the null hypothesis is rejected and the alternative hypothesis is accepted, H1: *There are statistically significant differences based on the Degree with regard to the valuation of the usability of chatbots.*

Based on the level of knowledge of chatbots

There are significant differences (CI 99.95%) for item 1: *I think the tool is easy to use* (F 298= 5,594; p<.018) and for item 10: *There is a lot I would need to learn before using this tool* (F 299= 11,260; p<.000). *Table 7* shows the results of the differential analysis.

Table 7.

Levene "T" test of independent simples

	Levene test for equality of variances		T test for equality of means						
	F	Sig.	t	gl	Sig. (bilateral)	Mean Differences	Typ. Error of the difference	95% Interval of reability for the difference	
								Higher	Lower
ITEM1	4.558	.034	-2.436	299	**.015**	-.275	.113	-.497	-.053
ITEM2	2.637	.105	-1.681	299	.094	-.204	.121	-.443	.035
ITEM3	5.594	.019	-2.370	298	**.018**	-.228	.096	-.417	-.039
ITEM4	.801	.372	1.154	299	.249	.125	.108	-.088	.337
ITEM5	.383	.537	-1.039	297	.300	-.101	.097	-.292	.090
ITEM6	.488	.485	-.317	299	.751	-.029	.090	-.206	.149
ITEM7	3.342	.069	-1.764	299	.079	-.143	.081	-.302	.016
ITEM8	.080	.778	-.397	299	.692	-.042	.106	-.251	.167
ITEM9	.228	.633	-1.188	299	.236	-.124	.104	-.329	.081
ITEM10	11.260	.001	3.965	299	**.000**	.515	.130	.260	.771

Note: *The Mean Difference is significant at level p<.05

Based don the usage of chatbots

There are significant differences (CI 99.95%) for item 1: *I this I would enjoy using this tool frequently* (F 295= 4,012; p<.11) and for item 2: *I find this tool quite simple* (F 295= .616; p<.016). *Table 8* includes the data obtained from this contrast analysis.

3.6. Regression Analysis

A first regression analysis is performed where the items of the SUS Scale are the explanatory-predictive variable of the variable "Level of knowledge on chatbots" (Table 9). A second analysis is performed on the variable "amount of use of chatbots" (Table 10). With regard to the first analysis, from the ten variables, two of them present significant

standardised regression indexes: item 1: *I think I would enjoy using this tool frequently*, and item 10: *There is a lot I would need to learn before using this tool*.

Table 8.

Levene "T" test of independent

	Levene test for equality of variances		T test for the equality of means						
	F	Sig.	t	gl	Sig. (bilateral)	Mean differences	Typ. Error of difference	95% Interval of relianility for the difference	
								Higher	Lower
ITEM1	4.012	.046	-2.560	295	**.011**	-.355	.139	-.627	-.082
ITEM2	.616	.433	-2.422	295	**.016**	-.357	.148	-.648	-.067
ITEM3	3.915	.049	-1.576	294	.116	-.186	.118	-.419	.046
ITEM4	.362	.548	.691	295	.490	.091	.132	-.169	.352
ITEM5	.779	.378	-1.187	293	.236	-.141	.119	-.376	.093
ITEM6	.459	.499	-1.545	295	.123	-.171	.110	-.388	.047
ITEM7	1.668	.197	-.354	295	.724	-.035	.100	-.231	.161
ITEM8	2.608	.107	-.991	295	.322	-.129	.130	-.385	.127
ITEM9	.060	.807	-.906	295	.366	-.116	.128	-.368	.136
ITEM10	.034	.855	1.305	295	.193	.213	.163	-.108	.534

Note: *Mean Differences are significant at leven p<.05

Table 9.

Coefficients (a) of the Lineal regression analysis of the variable Level of knowledge on chatbots using the SUS Scale

Model	B	t	p	R^2	F
2 (Constant)	0.22	29.946	0.000	0.56	1.765
10.- There is a lot I would need to learn before using this tool	0.21	10.861	0.000	0.63	1.671
1.- I think I would enjoy using this tool frequently	0.11	2.046	0.042	0.49	0.115

Note: (a) Dependent Variable: Leve lof knowledge on chatbots

The criterion or indicator to verify the explanatory model is the value of the determination coefficient (R^2) that calculates the percentage of the variable in the dependent variable. In the hereby case, R^2= .56 and, therefore, the degree of explanation of the model herein= 56% (R^2= 56% F (1,765) = 20.94, p< .000); this means that these results suggest that 56% of the variation of the variance in the dependent variable "Level of knowledge" may be explained by the lineal combination of these two items included in the model. The Alternative hypotheseis is versified (H1): *There is a relation between the satisfaction on the use of this tool and the level of knowledge of chatbots.*

Last, in the second regression analysis, only item 2 included in the regression model, thus, it is the only explanatory relational variable of the dependent variable "Degree of use

of chatbots". The explanatory degree of this model demonstrates that 27% of the variation in the dependent variable can be explained with this item.

Table 10.

Coefficients (b) of the Lineal regression analysis of the Variable Degree of use of chatbots with the SUS Scale

Model	B	t	p	R^2	F
1 (Constant)	0.959	10.252	0.000	0.27	7.988
2.- I find the tool quite simple	0.64	2.826	0.005		

Note: (b) Dependent variable: Degree of use of chatbots

The null hypothesis (H_0) is not rejected, but accepted, there is no relation between the acquisition of previous knowledge for the use of the tool with the degree of use of chatbots.

3.7. Real use of the CLOE conversational agent

Between January and June 2020, the virtual agent was implemented for 589 students enrolled in the degrees of Social, Early Years and Primary Education. The CA replied to a total of 3025 questions over the 6-month period, with a peak of 912 questions replied in February, coinciding with the presentation of the commitment between lecturers and students, and an average of 504 questions per month (Table 11).

With regard to the type of replies, it can be highlighted that it mostly replied to questions on how to cite authors in academic papers (852) and the correct referencing following APA rules (808). Students also used the virtual agent's small talk (510), who returned with replies to the topic for which it was created. Other recurring questions that it answered referred to the types of dissertations and their features (208), clarifying specific matters on certain aspects of the different types of papers. This was complemented with answers on formal aspects of the dissertation (222), regarding the format, indentation, length, or number of pages, among others.

Table 11.

Types of answers given by the implemented virtual agent

	January	February	March	April	May	June	Total
1. Formal aspects	2	61	20	24	53	62	222
2. Types and features of dissertation	14	78	35	27	24	30	208
3. Citing	25	232	136	12	167	150	852
4. APA referencing	16	104	210	132	180	166	808
5. Documents needed	17	76	24	3	14	14	148
6. Submission and defence dates	14	39	8	6	8	7	82
7. Qualifications	28	74	3	7	6	10	128
8. Lecturers: Tribunals and tutors	24	38	2	1	0	2	67
9. Natural conversation	69	210	89	51	52	39	510
Total	209	912	527	393	504	480	3025

4. Discussion and conclusions

The results improve knowledge on the characteristics that could involve the use and usefulness of this technology in academic environments. To this extent, an exhaustive review of the conversational agent was required on behalf of the degree coordinators, due to the large amount of information that was required in order for the tool to reply to all the questions that could have been asked (Fast et al., 2018; Sumikawa et al., 2020). Data reveals a high satisfaction of students with regard to the use of this tool, hence, it can be confirmed that the experience was notable for students. In a certain manner, the interaction, and immediate response when solving doubts, may facilitate a positive view towards communication with the virtual agent (Wargnier et al., 2015).

Likewise, it can be confirmed that there are significant differences in the variables that have been analysed: gender, degree, level of knowledge and amount of use of the chatbots. These differences were highlighted based on the degree each student was enrolled in. To this regard, it has been found that students from the degrees in Primary Education and Early Years Education share similar valuation criteria with regard to the usability of the tool. However, this is not true to the students enrolled in Social Education. The latter differed in their perspective regarding chatbots and their use; although their view and beliefs are also positive, the mean punctuations obtained in the different items that make up the Scale reveal that they position themselves between "neither agree or disagree" and "agree", whilst those of other students are located in higher alternatives. An instance of this can be found in item 8: "*I find the tool very intuitive*", where students from the degree in Primary Education and Early Years Education achieve an $\bar{X}$ = 4.51 and $\bar{X}$ = 4.74, respectively, against the $\bar{X}$ = 3.60 obtained by the Social Education students (F 300= 10,986; p<.000).

With regard to the factorial analysis carried out, the empirical structure and internal consistency of the Scale is similar to those studied and evaluated in sundry analogue research. The statistic coefficients verify the relevance of performing the EFA, the goodness of the matrix of correlations to be factorised, the goodness-of-fit model index, and that the factorial solution is parsimonious, methodological requirements and psychometric conditionings in order to accept the validity of the construct. The solution created lies on two factors that explain 55.73% of the variance that present high factorial charges. To this regard, the correlations performed inter-elements show positive relationships between the items that make up factor 1 and the items of factor 2, which is lineal and significantly negative between the items of each factor, an indication that they are independent between each other. Notwithstanding the above, it is agreed that the Scale has a unidimensional character (Bangor et al., 2008; Finstad, 2010; Lewis & Sauro, 2009), which facilitates its applicability.

The multiple regression analysis performed on the ten variables of the Scale on the level of knowledge and amount of use of the chatbots shows that three out of the ten variables relate to the dependent variables analysed. These explanatory-predictive variables are related to the satisfaction of using the tool more frequently, the easiness to use this technology and the need of previous knowledge before starting to use the chatbot. Therefore, these could be considered predictive elements, capable of being at the base that supports the usability of the chatbots. The use of the conversational agent is determined by the purpose for which it was created. Although the design is conditioned by the sematic complexity of the topic (Herbert & Kang, 2018), it could be an advantage that the topic is limited, since the concepts regarding the dissertation paper were expressions that the students already knew and shared. It is agreed with authors such

as Wellnhammer et al. (2020) or Colace et al. (2018) that virtual agents may be a positive tool to support the work of university students.

It is important to bear in mind the purpose of the virtual agent in order to evaluate it, since the creation of a conversational agent is complex (Sumikawa et al., 2020). This was one of the base factors while designing this chatbot, since it had to be capable of responding to numerous questions asked by students. In this case, the agent was equipped with initial categories that were then nuanced and broadened after having been tested by the coordinators of the different university degrees. Once tested, and after an interview, a database was developed to be able to answer to 174 different types of questions.

However, the number of matters was higher since different ways to recognise a same question had to be developed, this is, different ways of formulating them. The CA must use extremely complex syntactic and semantic constructions, in order for them to interact in the construction process of a conversation (Herbert & Kang, 2018). CAs have communication limitations, in the manner in which people develop it, and are able to simulate conversations on a specific topic without being able to develop complex dialogues in different topics or fields. There are occasions where the tool may provide a wrong answer when the student talks about an argument that has different meanings (Colace et al., 2018). Although the users could behave inadequately with the agent (Part et al., 2019), this was a matter that was observed residually in the interactions it fulfilled throughout the six months of its implementation.

Regarding the limitations that this research presents in terms of usability, it is key to highlight that the sample does not allow for the generalisation and extrapolation of the results since this would require the application of the scale to a broader and more diverse university population by means of stratified random sampling. Therefore, it would be of interest to repeat this research with different degrees, etc. This limitation, and the performance of a segmentation analysis where the usability of the chatbot is evaluated (dependent variable) with categorical variables “experimental group”-“control group” (independent variables) would determine future research.

There are scarce experiments in literature that evaluate the usability of chatbots, hence, more effort would be required in this line of research (Ren et al., 2019). The CA’s database can be exported to other degrees or universities, and could improve support to lecturers with regard to tutorials. It could also be of interest for subjects with a large number of students enrolled, such as university internship courses in similar degrees to those where the agent has assisted.

References

Babu, S. V., Suma, E., Hodges, L. F., & Barnes, T. (2011). Learning cultural conversational protocols with immersive interactive virtual humans. *International Journal of Virtual Reality*, *10*(4), 25-35. http://dx.doi.org/10.20870/IJVR.2011.10.4.2826

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction, 24, 574-594.* http://dx.doi.org/10.1080/10447310802205776

Brooke, J. (1986). SUS: A “quick and dirty” usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., McClelland (Ed.) *Usability Evaluation in Industry* (pp. 189-194). Taylor & Francis.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7.

Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, *8*(2), 29-40.

Bunge, M. (2010). *Las pseudociencias*. Laetoli.

Catalán, C., Delgado, C., Alario-Hoyos, C., & Muñoz-Merino, P. J. (2018). Supporting a MOOC through a conversational agent. Design of a first prototype. In *2018 International Symposium on Computers in Education (SIIE)* (pp. 1-6). IEEE. https://doi.org/10.1109/siie.2018.8586694

Cochran, W. G., & Bouclier, A. S. (1980). *Técnicas de muestreo*. Continental.

Colace, F., De Santo, M., Lombardi, M., Pascale, F., Pietrosanto, A., & Lemma, S. (2018). Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research*, *7*(5), 528-533. https://doi.org/10.18178/ijmerr.7.5.528-533

Coperich, K., Cudney, E., & Nembhard, H. (2017). Continuous improvement study of chatbot technologies using a human factors methodology. In *Proceedings of the 2017 Industrial and Systems Engineering Conference*.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811-817. https://doi.org/10.1017/s1351324916000243

Fadhil, A., & Gabrielli, S. (2017, May). Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare* (pp. 261-265). https://doi.org/10.1145/3154862.3154914

Fast, E., Chen, B., Mendelsohn, J., Bassen, J., & Bernstein, M. S. (2018, April). Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-12). https://doi.org/10.1145/3173574.3174047

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human-Computer Studies*, *132*, 138-161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323-327. https://doi.org/10.1016/j.intcom.2010.04.004

Fornell, C & Larcker, D.F. (1981). Evaluating structural equiations modelswith unoservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Ghosh, D., Foong, P. S., Zhang, S., & Zhao, S. (2018, April). Assessing the Utility of the System Usability Scale for Evaluating Voice-based User Interfaces. In *Proceedings of the Sixth International Symposium of Chinese CHI* (pp. 11-15). https://doi.org/10.1145/3202667.3204844

Hair, F., Anderson, R.E, Tatham, R.L. & Black, W.C. (1998). *Multivariate data analysis with readings*. Prentice Hall.

Hair, F., Anderson, R.E., Tatham, R.L. & Black, W.C. (2005). *Multivariate data analysis*. Prentice Hall.

Herbert, D., & Kang, B. H. (2018). Intelligent conversation system using multiple classification ripple down rules and conversational context. *Expert Systems with Applications*, 112, 342-352. https://doi.org/10.1016/j.eswa.2018.06.049

Hernández R.; Fernández, C. y Baptista, P. (2010). *Metodología de la investigación*. McGraw-Hill. (5ª Ed.)

Hobert, S. & Meyer von Wolff, R. (2019). Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. In *Proceedings of the 14th International Conference on Wirtschaftsinformatik* (S. 301–315). Siege.

Jung, S. (2013). Exploratory factor analysis with small sample sizes: A comparison of three approaches. *Behavioural Processes, 97,* 90–95.

Kaplan, R. M., & Saccuzzo, D. P. (2009). Standardized tests in education, civil service, and the military. *Psychological testing: Principles, applications, and Issues*, 7, 325-327.

Kerlinger, F.N. (1979). *Enfoque conceptual de la investigación del comportamiento.* Nueva Editorial Interamericana.

Kerly, A., Ellis, R., & Bull, S. (2008). Conversational agents in *E-Learning. Proceedings of Artificial Intelligence* 2008 (pp. 169–182).

Kirakowski, J. (1994). The use of questionnaire methods for usability assessment. http://sumi.ucc.ie/sumipapp.html

Landauer, T. K. (1997). Behavioral Research Methods in Human-Computer Interaction. In Helander, M., Landauer, T., Prabhu, P. (Ed.) *Handbook of Human-Computer Interaction* pp. 203-227. Elsevier

Lewis, J. (2018) The System Usability Scale: Past, Present, and Future. *International Journal of Human–Computer Interaction*, *34*(7), 577-590. https://doi.org/10.1080/10447318.2018.1455307

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In *International conference on human centered design* (pp. 94-103). Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-02806-9_12

Lucey, N. M. (1991). *More than Meets the I: User-Satisfaction of Computer Systems*. Unpublished thesis for Diploma in Applied Psychology, University College Cork.

Neumann, A. T., de Lange, P., & Klamma, R. (2019, December). Collaborative Creation and Training of Social Bots in Learning Communities. In *2019 IEEE 5th International Conference on Collaboration and Internet Computing* (CIC) (pp. 11-19). IEEE. https://doi.org/10.1109/cic48465.2019.00011

Nunnally, J.C. (1978). *Psychometric theory*. McGraw-Hill Book.

Park, M., Aiken, M., & Salvador, L. (2019). How do humans interact with chatbots? An analysis of transcripts. *International Journal of Management & Information Technology*, 14, 3338-3350. https://doi.org/10.24297/ijmit.v14i0.7921

Ren, R., Castro, J.W., Acuña, S.T., & de Lara, J. (2019). Usability of Chatbots: A Systematic Mapping Study. In *Proceedings of 31st International Conference on Software Engineering & Knowledge Engineering* (SEKE'19) (pp. 479-484). https://doi.org/10.18293/seke2019-029

Saenz, J., Burgess, W., Gustitis, E., Mena, A., & Sasangohar, F. (2017). The usability analysis of chatbot technologies for internal personnel communications. In *IIE Annual Conference. Proceedings* (pp. 1357-1362). Institute of Industrial and Systems Engineers (IISE).

Sauro, J., & Lewis, J. R. (2009). Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability. *In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1609-1618).* https://doi.org/10.1145/1518701.1518947

Schroeder, J., Wilkes, C., Rowan, K., Toledo, A., Paradiso, A., Czerwinski, M., Mark, G., & Linehan, M. M. (2018, April). Pocket skills: A conversational mobile web app to support dialectical behavioral therapy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-15). https://doi.org/10.1145/3173574.3173972

Sinoo, C., van Der Pal, S., Henkemans, O. A. B., Keizer, A., Bierman, B. P., Looije, R., & Neerincx, M. A. (2018). Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management. *Patient Education and Counseling*, *101*(7), 1248-1255. https://doi.org/10.1016/j.pec.2018.02.008

Sumikawa, Y., Fujiyoshi, M., Hatakeyama, H., & Nagai, M. (2020). Supporting creation of FAQ dataset for E-learning chatbot. In *Intelligent Decision Technologies 2019* (pp. 3-13). Springer. https://doi.org/10.1007/978-981-13-8311-3_1

Tabachnik, B.G., & Fidell, L.S., (1989). *Using Multivariate Statistics*. Harper Collins Publishers.

Tegos, S., & Demetriadis, S. (2017). Conversational agents improve peer learning through building on prior knowledge. *Journal of Educational Technology & Society*, *20*(1), 99-111.

Tegos, S., Demetriadis, S., & Tsiatsos, T. (2014). A configurable conversational agent to trigger students' productive dialogue: a pilot study in the CALL domain. *International Journal of Artificial Intelligence in Education*, *24*(1), 62-91. https://doi.org/10.1007/s40593-013-0007-3

Tielman, M.L., Neerincx, M.A., Bidarra, R., Kybartas, B., & Brinkman, W.P. (2017). A therapy system for post-traumatic stress disorder using a virtual agent and virtual storytelling to reconstruct traumatic memories. *Journal of medical systems*, *41*(8), 125-145. https://doi.org/10.1007/s10916-017-0771-y

Tullis, T. S., & Stetson, J. N. (2004). A Comparison of Questionnaires for Assessing Website Usability, *unpublished presentation given at the UPA Annual Conference*. http://home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf

Wargnier, P., Malaisé, A., Jacquemot, J., Benveniste, S., Jouvelot, P., Pino, M., & Rigaud, A. S. (2015, March). Towards attention monitoring of older adults with cognitive impairment during interaction with an embodied conversational agent. In *2015 3rd IEEE VR International Workshop on Virtual and Augmented Assistive Technology (VAAT)* (pp. 23-28). IEEE. https://doi.org/10.1109/vaat.2015.7155406

Wellnhammer, N., Dolata, M., Steigler, S., & Schwabe, G. (2020, January). Studying with the Help of Digital Tutors: Design Aspects of Conversational Agents that Influence the Learning Process. In *Proceedings of the 53rd (pp. 146-155). Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2020.019

Winkler, R., & Söllner, M. (2020, January). Towards Empowering Educators to Create their own Smart Personal Assistants. In *Proceedings of the 53rd (pp. 22-31). Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2020.005

Zeni, J., da Costa, C. & da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications*, 129, 56-67. https://doi.org/10.1016/j.eswa.2019.03.054

Cómo citar:

Artíles-Rodríguez, J., Guerra-Santana, M., Aguiar-Perera, Mª. V., & Rodríguez-Pulido, J. (2021). Agente conversacional virtual: la inteligencia artificial para el aprendizaje autónomo [Embodied conversational agents: artificial intelligence for autonomous learning]. *Pixel-Bit. Revista de Medios y Educación*, 62, 107-144. https://doi.org/10.12795/pixelbit.86171