UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Doble Grado en Ingeniería Informática y Matemáticas

TRABAJO FIN DE GRADO

EVALUACIÓN EMPÍRICA DE LA USABILIDAD DE UN CHATBOT

Autora: Andrea Nevado Labrador
Tutora: Silvia Teresita Acuña Castillo

JUNIO DE 2019

EVALUACIÓN EMPÍRICA DE LA USABILIDAD DE UN CHATBOT

AUTORA: Andrea Nevado Labrador
TUTORA: Silvia Teresita Acuña Castillo

Grupo de Investigación de Herramientas Interactivas Avanzadas (GHIA)
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2019

Resumen

Los chatbots, agentes de conversación basados en mensajería, han experimentado un gran crecimiento recientemente y están siendo utilizados en diferentes áreas por una amplia variedad de usuarios. Los chatbots están diseñados para hacer las interacciones con el usuario lo más naturales posibles.

Un aspecto crítico en los sistemas software interactivos es la usabilidad, necesaria para proporcionar una experiencia de usuario adecuada. La usabilidad de un chatbot puede ser evaluada a través de experimentos, sin embargo, son pocos los estudios encontrados en la literatura con este propósito.

El objetivo del presente trabajo es diseñar y realizar un experimento para evaluar la usabilidad del chatbot SOCIO, cuya funcionalidad es ayudar en la elaboración de diagramas de clases mediante la interpretación del lenguaje natural. Además, está integrado en las redes sociales Twitter y Telegram, por lo que permite trabajar de manera colaborativa. La evaluación de la usabilidad de SOCIO se realiza mediante la comparación con la usabilidad de otra herramienta que también permite la elaboración de diagramas de clases, la aplicación web Creately. La usabilidad es evaluada con respecto a la eficacia, la eficiencia y la satisfacción desde el punto de vista de usuarios con conocimientos en informática, a su vez se evalúa y compara la calidad de los diagramas de clases obtenidos al emplear dichas herramientas.

El experimento propuesto presenta un diseño *crossover*. En él, la mitad de los sujetos experimentales utilizan el chatbot SOCIO para realizar la primera tarea del experimento y Creately para llevar a cabo la segunda tarea. La otra mitad de los participantes aplica los tratamientos en el orden inverso, realizando la primera tarea con Creately y la segunda con SOCIO. Cada tarea consiste en la elaboración de un diagrama de clases en equipos de tres integrantes.

En el experimento han participado de manera voluntaria 30 estudiantes con conocimientos en Ingeniería Informática. El tamaño muestral se corresponde con el número de equipos formados durante el experimento, en este caso, 10 equipos.

Tras la ejecución del experimento, se analizan estadísticamente los datos recolectados. En primer lugar, se realiza un análisis descriptivo, mediante diagramas de caja, de los datos asociados a las métricas de la eficacia, eficiencia, satisfacción y calidad. En segundo lugar, se ajusta un modelo lineal mixto para cada métrica, y los resultados se complementan calculando el tamaño del efecto del tratamiento, a través del cálculo de la *d* de Cohen y su error estándar.

A partir de los resultados del análisis y su discusión, se concluye que el tamaño muestral del experimento parece insuficiente y ha podido causar la inexistencia de diferencias significativas producidas por las herramientas en las métricas asociadas a la eficacia, la eficiencia y la calidad. Sin embargo, en relación con la eficiencia, medida a través del tiempo y los mensajes de discusión, parece que el chatbot SOCIO es más eficiente que Creately, es decir, se requiere menor esfuerzo al utilizar SOCIO. Este hecho se refleja en los diagramas de caja asociados a dichas métricas. Estos diagramas muestran tiempos más bajos para el chatbot y un menor número de mensajes de discusión intercambiados. La

variable satisfacción sí parece verse afectada por la herramienta utilizada, con resultados favorables para SOCIO. Finalmente, la tarea parece afectar a la satisfacción y a las métricas de la calidad, en concreto, la segunda tarea parece producir peores resultados. También, la secuencia (orden de aplicación) parece afectar a ciertas métricas de la calidad, implicando que la interacción entre la tarea y el tratamiento o los efectos de *carryover* se han materializado, lo cual supone una amenaza a la validez interna. Ante los resultados obtenidos, se requiere de réplicas del experimento base para consolidar los resultados de la investigación.

Palabras clave: Usabilidad, Chatbot, Experimento, Diseño *Crossover*, Eficacia, Eficiencia, Satisfacción, Modelo Lineal Mixto

Abstract

Messaging-based conversational agents, or Chatbots, have recently seen massive growth. They are used in many different areas and by a wide variety of users. Chatbots are designed to make interactions closer and more personal to the user.

Usability is a critical aspect in interactive software systems and it is essential to incorporate usability in chatbots, in order to improve user experience. Chatbot usability can be evaluated through experiments; however, there are few studies in the literature for this purpose.

The aim of this work was to design and perform an experiment to evaluate the usability of the chatbot SOCIO. SOCIO's goal is to help in the elaboration of class diagrams through natural language. Moreover, it is a collaborative tool integrated with Twitter and Telegram. The chatbot's usability is evaluated by comparing it with the usability of the web application Creately, which also allows for the creation of class diagrams. The usability will be evaluated with respect to effectiveness, efficiency and satisfaction from the point of view of users with a computer science background. The quality of the class diagrams obtained using these tools will also be evaluated and compared.

The experiment has a crossover design. During the experiment, half of the experimental subjects use SOCIO to perform the first task and Creately to carry out the second task. The other half applies the treatments in the reverse order, performing the first task with Creately and the second one with SOCIO. Each task consists in the elaboration of a class diagram in groups of three members.

30 students with computer science backgrounds participated voluntarily in the experiment. The sample size corresponds with the number of groups formed during the experiment, in this case, 10 teams.

Once the experiment was carried out, a statistical analysis of the data was performed. First of all, a descriptive analysis through box-plots was carried out for the data associated with the metrics of effectiveness, efficiency, satisfaction and quality. Next, it was fitted one linear mixed model per metric. The results of the statistical analysis were complemented by the size of the treatment effect, through Cohen's *d* for the treatments and their corresponding standard error.

From the results of the analysis and its discussion, it is clear that the sample size of the experiment seems insufficient and there is no significant differences produced by the tool in the metrics associated with effectiveness, efficiency and quality. However, the efficiency, measured through time, and the discussion messages, seem better for SOCIO than for Creately. In other words, less effort is required when using SOCIO. This comes from the fact that box-plots show lower times for the chatbot and a smaller number of discussion messages exchanged. The satisfaction variable does seem to be affected by the tool used, with favorable results for SOCIO. Finally, the task seems to affect satisfaction and quality metrics, specifically, the second task seems to produce worse results. Also, the sequence (order of application) seems to affect certain quality metrics, implying that the interaction between the task and the treatment or carryover effects have materialized,

which poses a threat to internal validity. Given the results obtained, the replication of base experiments is required to consolidate the results of the investigation.

Keywords: Usability, Chatbot, Experiment, Crossover Design, Effectiveness, Efficiency, Satisfaction, Mixed Linear Model

Agradecimientos

A mi tutora Silvia, por su ayuda, su disponibilidad y su guía a lo largo de todo el trabajo.

A Sara y a Adrián, por su ayuda y su tiempo.

A mis compañeros, por participar voluntariamente en el experimento.

A mis amigos de siempre y a los que se han convertido en mis amigos durante estos años, por los buenos momentos compartidos.

A mi familia, por estar ahí siempre, por su cariño, su apoyo y su confianza.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 Introducción

En este primer capítulo, se describe la motivación del presente trabajo, los objetivos del mismo y su estructura.

1.1 Motivación

Los chatbots, agentes de conversación basados en mensajería, han experimentado un gran crecimiento recientemente (Jain et al., 2018). Los chatbots están diseñados para hacer las interacciones con el usuario lo más naturales posibles, similares a una conversación de persona a persona (Nguyen & Sidorova, 2018). Están siendo utilizados en diferentes áreas, como ayudantes en el autocontrol y superación de enfermedades, ayudantes en la planificación de viajes y búsqueda de información, o como asistentes personales (Cheng et al., 2018; Lopatovska et al., 2018; Nguyen & Sidorova, 2018; Pérez et al., 2017; Tielman et al., 2017). Este hecho implica el uso de los chatbots por parte de una amplia variedad de usuarios.

Un aspecto crítico en los sistemas software interactivos, que es necesario incorporar en el desarrollo de los chatbots para proporcionar un experiencia de usuario adecuada, es la usabilidad (Ren et al., 2019). La usabilidad se define como el grado en el que un software puede ser utilizado por usuarios específicos para lograr determinados objetivos con eficacia, eficiencia y satisfacción, en un contexto específico de uso (ISO 9241-11, 1998).

Métodos comunes de investigación para evaluar la usabilidad de los chatbots son las encuestas, los test de usabilidad y los experimentos (Ren et al., 2019). La experimentación es una cuestión clave en la ciencia y en la ingeniería. En la Ingeniería del Software, esta práctica es bastante reciente (Vegas et al., 2016). 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). El presente trabajo lleva a cabo un experimento para evaluar la usabilidad del chatbot SOCIO (Pérez-Soler et al., 2017), realizando una contribución en este ámbito.

El chatbot SOCIO asiste en la elaboración de diagramas de clases mediante la interpretación de mensajes en lenguaje natural. Con el objetivo de beneficiarse de la naturaleza ubicua y colaborativa de las redes sociales, SOCIO está integrado en Twitter y Telegram. Así, permite la creación de diagramas en grupo, en cualquier momento y desde cualquier lugar. Solo se han realizado evaluaciones de este chatbot a pequeña escala (Pérez-Soler et al., 2017; Pérez-Soler et al., 2018), por lo que también se requiere más esfuerzo en esta línea.

1.2 Objetivos

Con el propósito de colaborar en la línea de la experimentación con chatbots para la evaluación de su usabilidad y la obtención de evidencias empíricas y, en particular, en la evaluación del chatbot SOCIO, los objetivos del presente Trabajo de Fin de Grado son los siguientes:

- Realizar un análisis de la literatura, a partir del cual se diseñe un experimento que evalúe la usabilidad del chatbot SOCIO. La usabilidad será evaluada con respecto a la eficacia, la eficiencia y la satisfacción de los usuarios interaccionando con SOCIO en el ámbito académico, con estudiantes del Grado en Ingeniería Informática y del Doble Grado en Ingeniería Informática y Matemáticas de la Escuela Politécnica Superior de la Universidad Autónoma de Madrid. A su vez, el experimento evaluará la calidad de los diagramas de clases generados por el chatbot.

- Llevar a cabo el experimento diseñado.

- Analizar estadísticamente los datos recolectados durante la ejecución del experimento y realizar una discusión de los resultados obtenidos, concluyendo así la evaluación del chatbot SOCIO.

1.3 Estructura del Trabajo

El presente trabajo consta de seis capítulos:

En el primer Capítulo se detalla la motivación del trabajo y los objetivos del mismo.

En el segundo Capítulo se aborda el estado del arte de la experimentación con chatbots para la evaluación de su usabilidad, se presentan las características del chatbot SOCIO y evaluaciones de distintos aspectos del mismo.

En el tercer Capítulo se describe el experimento llevado a cabo con el chatbot SOCIO. Se especifica el diseño experimental, se abordan los objetivos e hipótesis de la investigación, los factores y variables respuesta, y las métricas empleadas para medir cada variable. A su vez, se detalla el perfil de los sujetos experimentales, las herramientas utilizadas, las tareas a realizar y la ejecución del experimento. Por último, se consideran las amenazas a la validez interna y externa del experimento.

En el cuarto Capítulo se muestra el análisis estadístico de los datos recolectados durante el experimento.

En el quinto Capítulo se discuten los resultados obtenidos en el análisis.

En el sexto y último Capítulo se presentan las conclusiones y trabajos futuros.

A continuación de los capítulos, se muestran las referencias empleadas durante la realización del trabajo, un glosario con definiciones y seis anexos:

En el Anexo A se exponen 10 tablas comparativas: ocho de experimentos de la literatura realizados con chatbots y dos de evaluaciones realizadas con el chatbot SOCIO.

En el Anexo B se exponen los documentos utilizados durante las sesiones en las que se llevó a cabo el experimento con el chatbot SOCIO.

En el Anexo C se detalla el funcionamiento de las herramientas empleadas en el experimento.

En el Anexo D se explica cómo se ha llevado a cabo la evaluación de la calidad de los diagramas realizados durante las tareas del experimento.

En el Anexo E se muestran diagramas de caja que complementan el análisis descriptivo de los resultados del experimento.

En el Anexo F se detalla el cálculo del tamaño del efecto realizado en el análisis.

2 Estado del arte

En este capítulo, en la sección 2.1, se analizan ocho artículos de experimentación con chatbots, obtenidos de un *Systematic Mapping Study* más amplio (Ren et al., 2019). En la sección 2.2, se analizan dos artículos, proporcionados por el investigador principal del proyecto bajo el cual se desarrolló el chatbot SOCIO, Juan de Lara Jaramillo, profesor de la Universidad Autónoma de Madrid. Estos artículos describen al chatbot SOCIO y las evaluaciones realizadas de distintos aspectos del mismo. El propósito de conocer el estado del arte de la experimentación con chatbots es definir el diseño del experimento que evaluará la usabilidad de SOCIO. En la sección 2.3, se describe el diseño elegido a grandes rasgos.

2.1 Experimentación con Chatbots para la Evaluación de su Usabilidad

Los ocho artículos de experimentación con chatbots, reportados en la literatura (Ren et al., 2019), describen experimentos llevados a cabo para conocer los factores que afectan en la interacción de los usuarios con los chatbots y evaluar su usabilidad. Los chatbots abordados pertenecen a diferentes ámbitos. Algunos ayudan a auto-controlar enfermedades como la diabetes (Cheng et al., 2018; Sinoo et al., 2018), u ofrecen terapia para pacientes que sufren trastornos de estrés postraumático (Tielman et al., 2017). Otros facilitan la planificación de viajes (Nguyen & Sidorova, 2018), compra de zapatos (Jain et al., 2018), búsqueda de información (Pérez et al., 2017) o son asistentes personales, como Apple Siri y Amazon Alexa (Chen & Wang, 2018; Lopatovska et al., 2018).

La técnica de evaluación empleada en la mayoría de los experimentos consiste en comparar el chatbot con otra aplicación o sistema que presente la misma funcionalidad, o con él mismo, añadiendo o modificando cierta característica (Cheng et al., 2018; Jain et al., 2018; Nguyen & Sidorova, 2018; Pérez et al., 2017; Sinoo et al., 2018). En (Nguyen & Sidorova, 2018), por ejemplo, se comparan un sistema web y un sistema chatbot para estudiar las diferencias en los niveles de satisfacción del sistema y determinar los factores que afectan.

El diseño experimental empleado solo es mencionado en uno de los experimentos (Jain et al., 2018). Se trata de un diseño *within-subject* en el que los sujetos experimentales deben aplicar todos los tratamientos a evaluar. Sin embargo, se deduce que otros experimentos, donde la evaluación del chatbot en cuestión se realiza mediante comparación, también siguen este diseño, pues hacen referencia a la aplicación de todos los tratamientos por parte de todos los participantes (Nguyen & Sidorova, 2018; Pérez et al., 2017). Un aspecto que sí se especifica en los experimentos es el orden aleatorio de aplicación de los tratamientos, para evitar los efectos que pueda producir utilizarlos en un orden en particular.

El perfil de los sujetos experimentales varía dependiendo de la funcionalidad del chatbot. Si un chatbot está orientado a un grupo de usuarios específico, los sujetos reclutados para el experimento deben poseer las características de dicho grupo (Cheng et al., 2018; Sinoo et al., 2018; Tielman et al., 2017). A su vez, se debe tener en cuenta el objetivo del estudio a la hora de seleccionar a los participantes. En (Chen & Wang, 2018),

por ejemplo, el propósito es estudiar cómo afecta la experiencia personal y el conocimiento técnico en la interacción con agentes conversacionales y la usabilidad percibida. Para ello, la muestra escogida es variada, algunos sujetos presentan estos factores y otros no.

Los tamaños muestrales empleados son diversos, desde 7 participantes (Tielman et al., 2017) hasta 41 (Chen & Wang, 2018). La mayoría se concentra en la franja de 10 a 22 sujetos, siendo la media 19,43 y la desviación típica 11,03. En ninguno de los experimentos se justifica el tamaño muestral elegido.

En la mayoría de los experimentos, las tareas que deben realizar los sujetos son sencillas. Por ejemplo, utilizar Apple Sira para buscar un hotel barato en Osaka (Chen & Wang, 2018), buscar un billete de avión y una habitación de hotel empleando un chatbot y una web (Nguyen & Sidorova, 2018) o jugar con un robot y su avatar (Sinoo et al., 2018). En algunos experimentos, los participantes reciben tutoriales o información acerca del chatbot o de la aplicación que van a utilizar, antes de realizar las tareas (Jain et al., 2018; Tielman et al., 2017).

Finalmente, en todos los experimentos están presentes los cuestionarios. A través de ellos, se recolectan datos de la experiencia del usuario y su satisfacción. Los cuestionarios se proporcionan al finalizar el experimento, aunque en algunos casos, también se rellenan cuestionarios después de cada tarea y/o al inicio del experimento para conocer mejor el perfil de los usuarios (Jain et al., 2018; Pérez et al., 2017).

2.2 El Chatbot SOCIO y sus Evaluaciones

El chatbot SOCIO asiste en la elaboración de diagramas de clases (Pérez-Soler et al., 2017). Con el objetivo de obtener beneficio de la naturaleza ubicua y colaborativa de las redes sociales, SOCIO está integrado en Twitter y Telegram. Este enfoque mejora la flexibilidad de la creación de diagramas de clases, pues permite diseñar en grupo, en cualquier momento y en cualquier lugar. Además, se integra perfectamente con el uso normal de las redes sociales con las que los usuarios están familiarizados.

A través de la red social, los usuarios pueden interactuar entre ellos de la manera habitual, enviando mensajes de discusión y colaboración. A su vez, los mensajes pueden dirigirse al chatbot SOCIO, estos serán interpretados y empleados para la construcción del diagrama. En Twitter los usuarios deben ser seguidores de SOCIO y mencionar su nombre de usuario (@modellingBot) en aquellos mensajes cuya intención sea ir dirigidos al chatbot. En Telegram, las personas interesadas en la creación de un diagrama y el chatbot deben crear un grupo. En esta red social, los mensajes enviados a SOCIO comienzan por /.

Los mensajes enviados al chatbot SOCIO pueden ser comandos de gestión o mensajes de actualización del diagrama. Los primeros permiten realizar tareas de gestión del proyecto, como crear un nuevo proyecto o consultar los existentes. Los segundos pueden ser comandos o mensajes descriptivos. Los comandos son acciones imperativas que manipulan el diagrama directamente, para añadir una clase, cambiar un tipo o borrar un elemento. Los mensajes descriptivos son sentencias en lenguaje natural del contexto del diagrama a elaborar, como "*the shop contains products*".

Interactuar a través de mensajes puede ser una manera más fácil y rápida de elaborar un diagrama de clases, en comparación con otras herramientas de creación de diagramas.

Además, por una parte, el uso del lenguaje natural brinda a personas con poco conocimiento sobre diagramas de clases, o sin él, la oportunidad de colaborar de manera activa en labores de diseño. Así, grupos de ingenieros y expertos en otros dominios pueden trabajar de manera conjunta. Por otra parte, en el ámbito educativo, el chatbot SOCIO podría emplearse para la resolución de ejercicios de elaboración de diagramas de clases en grupo.

Ante el problema de diseñar un diagrama, la exploración de diferentes soluciones puede ser necesaria, para ello, SOCIO soporta ramas, y para elegir una de las soluciones, incorpora un mecanismo de consenso.

Las características expuestas del chatbot SOCIO se describen en dos artículos analizados (Pérez-Soler et al., 2017; Pérez-Soler et al., 2018). También, se presentan en ellos dos evaluaciones a pequeña escala, en la primera se valora la idoneidad de este chatbot, en la segunda, su mecanismo de consenso. El número de participantes en ambas evaluaciones es reducido, 10 y 8, respectivamente.

Las tareas de las evaluaciones se realizan en grupos, en un contexto de usuario mayoritariamente informático. Por un lado, los participantes de la primera evaluación tienen formación en Ingeniería Informática (graduados o estudiantes de último año) y realizan la tarea propuesta divididos en 4 grupos de Telegram: 2 grupos de 2 y 2 grupos de 3. Por otro lado, los participantes de la segunda evaluación constituyen un único grupo de Telegram, y en cuanto a sus perfiles, 6 son ingenieros informáticos, 1 ingeniero en telecomunicaciones y 1 físico.

Las tareas propuestas son sencillas. En la primera evaluación, se debe realizar un diagrama para el comercio electrónico en 15 minutos. En la segunda, tras un tutorial de 10 minutos, elegir la mejor de tres opciones para dos proyectos, primero con el mecanismo de consenso ausente y después presente.

Tras las tareas, en ambas evaluaciones se rellena un cuestionario de manera individual. El cuestionario de la primera, presenta una parte para evaluar la satisfacción de los usuarios, preguntas acerca del empleo del lenguaje natural y la integración del chatbot en las redes sociales, y finalmente preguntas abiertas. El cuestionario de la segunda evaluación se centra en el mecanismo de consenso.

Finalmente, los resultados de la evaluación de la idoneidad de la propuesta, integrar el chatbot de elaboración de diagramas de clases en las redes sociales, fueron positivos en cuanto a la satisfacción, el empleo del lenguaje natural como método de interacción y la idea de colaborar a través de las redes (Pérez-Soler et al., 2017). En cuanto al set de comandos de SOCIO y la precisión de la interpretación del lenguaje natural, los resultados fueron buenos, pero no tanto, lo que sugiere la necesidad de mejorar en esta línea.

El mecanismo de consenso fue considerado útil para grupos numerosos y con una salida que refleja la opinión de la mayoría (Pérez-Soler et al., 2018).

2.3 Conclusiones

Son pocos los experimentos que hay en la literatura para evaluar y obtener evidencias empíricas de la usabilidad de los chatbots, por lo que se necesita más esfuerzo en esta

línea. Por otra parte, las evaluaciones realizadas del chatbot SOCIO son a pequeña escala y ninguna de ellas ha tenido como objetivo evaluar su usabilidad en términos de la eficacia, la eficiencia y la satisfacción.

Con el propósito de contribuir tanto a la experimentación para evaluar la usabilidad de los chatbots, como a la evaluación del chatbot SOCIO, a partir del análisis de los experimentos de la sección 2.1 y de las evaluaciones de la sección 2.2, se ha diseñado un experimento para evaluar la usabilidad de SOCIO respecto a la eficacia, la eficiencia y la satisfacción de equipos de usuarios interaccionando con el chatbot. En el Anexo A, se muestran las tablas comparativas de los experimentos y de las evaluaciones analizados en la literatura.

Como la técnica más utilizada en los artículos analizados es la comparación, en el experimento diseñado, se compara este chatbot con otra herramienta de elaboración de diagramas de clases. Se utiliza un diseño experimental *within-subjects*, mencionado en varios de los experimentos analizados, por lo que los participantes deben aplicar estos dos tratamientos. Se considerará también el orden de aplicación, aleatorio, para evitar los efectos que pueda producir un orden en particular. Las tareas a realizar son sencillas, consisten en la elaboración de dos diagramas de clases para gestionar una tienda y un colegio. Como en todos los artículos analizados, se emplearán cuestionarios: de familiaridad al inicio del experimento y de satisfacción tras cada tarea. Se imparte un tutorial de cinco minutos sobre el tratamiento a utilizar antes de cada tarea.

Los sujetos elegidos tienen conocimientos en Ingeniería Informática, en concreto, en la realización de diagramas de clases. Como en las evaluaciones anteriores del chatbot, los participantes llevan a cabo las tareas en equipos, en concreto, de tres miembros.

En total, los sujetos experimentales son 30. El número es superior al de las evaluaciones anteriores del chatbot SOCIO (8 y 10). En comparación con los experimentos analizados, es superior a la media 19,43. Sin embargo, se ha de tener en cuenta que se trabaja en equipos de 3, por lo que el tamaño muestral se corresponde con el número de equipos, en este caso 10. Este número sigue siendo superior a los 4 equipos de la primera evaluación de SOCIO y al único equipo de la segunda, pero está por debajo del tamaño muestral medio de los otros experimentos. Por lo tanto, el tamaño muestral con el que se realiza el experimento es pequeño.

En el siguiente capítulo se describe de manera más específica el diseño experimental, expuesto en esta sección a grandes rasgos.

3 Experimento

En este capítulo se presenta el experimento realizado para evaluar la usabilidad del chatbot SOCIO. Las secciones sucesivas detallan su diseño, los objetivos y las hipótesis de la investigación, los factores y las variables respuesta del experimento, los sujetos experimentales y, finalmente, las amenazas a la validez.

3.1 Diseño Experimental

El experimento posee un diseño ***crossover***, ilustrado en la Tabla 1. Se trata de un caso particular de diseño *within-subjects*. En un diseño *within-subjects*, cada sujeto experimental aplica todos los tratamientos. Cuando diferentes grupos de sujetos aplican los tratamientos en un orden distinto, se obtiene un diseño *crossover* (Vegas et al, 2016).

Los **tratamientos** de este experimento son dos herramientas de elaboración de diagramas de clases: el chatbot SOCIO y la aplicación web Creately.

Para emplear las herramientas, los sujetos se dividen, al azar, en dos grupos experimentales, A y B, y dentro de cada grupo se forman aleatoriamente equipos de tres miembros. Los equipos del grupo A emplean primero Creately y después SOCIO. Los equipos del grupo B emplean primero SOCIO y después Creately. Estos dos órdenes de aplicación de los tratamientos se denominan **secuencias**.

Los momentos en los que se aplica cada tratamiento se denominan **periodos**. Los equipos emplean cada una de las dos herramientas una sola vez, por lo que hay dos momentos de aplicación, es decir, dos periodos. En el momento en el que se aplica la primera herramienta (de acuerdo al grupo experimental), periodo 1, se realiza la tarea 1 y en el momento en el que se utiliza la segunda herramienta, periodo 2, se lleva a cabo la tarea 2.

	PERIODO 1	PERIODO 2	
	TAREA 1	TAREA 2	
GRUPO A	CREATELY	SOCIO	SECUENCIAS
GRUPO B	SOCIO	CREATELY	

Tabla 1: Diseño experimental.

El diseño *crossover* presenta dos ventajas. Por una parte, requiere un menor número de sujetos que un diseño *between-subject*. En un diseño *between-subject*, cada sujeto es asignado a un único tratamiento, por ello, *n* sujetos producen *n* observaciones, mientras

que un diseño cruzado, *n* sujetos generan *n x t*, donde *t* es el número de tratamientos. Por otra parte, controla las diferencias intrínsecas de los participantes al exponerlos a todos los tratamientos.

3.2 Objetivo, Pregunta de Investigación e Hipótesis

El objetivo de la investigación es evaluar la usabilidad del chatbot SOCIO, mediante la comparación con la usabilidad de la aplicación web Creately, con respecto a la eficacia, la eficiencia y la satisfacción desde el punto de vista de usuarios con conocimientos en informática; así como evaluar y comparar la calidad de los diagramas de clases obtenidos al emplear dichas herramientas.

Conforme a este objetivo, la pregunta de investigación es la siguiente:

> **RQ:** ¿El uso del chatbot SOCIO tiene un efecto más positivo en la eficacia, la eficiencia y la satisfacción del usuario al realizar un diagrama de clases, así como en la calidad del mismo, en relación con el uso de la aplicación web Creately?

Las hipótesis de la investigación son:

> **H.1.1.0:** No existe diferencia en la eficacia al realizar el diagrama de clases con SOCIO o con CREATELY.
> **H.1.2.0:** No existe diferencia en la eficiencia al realizar el diagrama de clases con SOCIO o con CREATELY
> **H.1.3.0:** No existe diferencia en la satisfacción al realizar el diagrama de clases con SOCIO o con CREATELY.
> **H.1.4.0:** No existe diferencia en la calidad del diagrama de clases al realizarlo con SOCIO o CREATELY.

Las hipótesis expuestas involucran tanto al chatbot SOCIO como a la aplicación web Creately. Sin embargo, el uso del chatbot SOCIO genera un conjunto de datos empíricos mayor que Creately. Por ello, considerando los datos adicionales proporcionados por SOCIO, durante las dos tareas del experimento, se plantean también las siguientes hipótesis:

> **H.2.1.0:** No existe diferencia en el número de mensajes que los equipos dirigen a SOCIO al realizar la tarea 1 o la tarea 2.
> **H.2.2.0:** No existe diferencia en el número de mensajes erróneos que los equipos dirigen a SOCIO al realizar la tarea 1 o la tarea 2.
> **H.2.3.0:** No existe diferencia en el número de mensajes útiles que los equipos dirigen a SOCIO al realizar la tarea 1 o la tarea 2.
> **H.2.4.0:** No existe diferencia en el número de mensajes descriptivos que los equipos dirigen a SOCIO al realizar la tarea 1 o la tarea 2.
> **H.2.5.0:** No existe diferencia en el número de comandos que los equipos dirigen a SOCIO al realizar la tarea 1 o la tarea 2.
> **H.2.6.0:** No existe diferencia en el número de acciones desencadenadas por SOCIO al realizar la tarea 1 o la tarea 2.

3.3 Factores y Variables Respuesta

Los **factores** de un experimento *crossov*er son el **tratamiento**, la **secuencia** (orden de aplicación de los tratamientos) y el **periodo** (momento en el que se aplica un tratamiento), (Vegas et al., 2016).

En el caso de este experimento, de acuerdo con el diseño experimental descrito en la sección 3.1, cada factor posee dos niveles. Los niveles del tratamiento son el chatbot SOCIO y la aplicación web Creately. Los niveles de la secuencia son Creately-SOCIO y SOCIO-Creately. Por último, el periodo presenta los niveles periodo 1 y periodo 2. Durante el periodo 1 (momento en el que se aplica el primer tratamiento, acorde al grupo experimental) se realiza la tarea 1, en el periodo 2 (momento en el que se aplica el segundo tratamiento) se lleva a cabo la tarea 2. Este hecho implica que el efecto producido por el factor periodo se confunda con el efecto de la tarea realizada.

Las **variables respuesta** abordadas son la **eficacia**, la **eficiencia** y la **satisfacción**, ya que (ISO/IEC 25010, 2010) las define como características comunes para evaluar la usabilidad; y la **calidad**.

La métrica utilizada para medir la eficacia es el grado completitud con el que un equipo ha finalizado la tarea.

Las métricas empleadas para mediar la eficiencia son:

- Rapidez.
  - Tiempo en minutos empleado por un equipo para completar una tarea.

- Fluidez.
  - Número de mensajes de discusión generados por un equipo, con la finalidad de comunicarse entre sí, durante la realización de la tarea.
  - Número de mensajes enviados al chatbot SOCIO por un equipo durante la realización de la tarea.
  - Número de mensajes erróneos dirigidos al chatbot SOCIO por un equipo durante la realización de la tarea. Estos son mensajes cuya intención inicial es ser enviados a SOCIO, pero por errores de escritura no se dirigen al chatbot o en el caso de ser enviados, no son comprendidos.

- Interactividad**.**
  - Número de mensajes útiles dirigidos al chatbot por un equipo durante la realización de la tarea. Estos son mensajes que han supuesto una aportación al diagrama realizado.
  - Dentro del número de mensajes útiles, número de mensajes descriptivos empleados por un equipo para realizar el diagrama de clases de una tarea.
  - Dentro del número de mensajes útiles, número de comandos (mensajes imperativos) empleados por un equipo para realizar del diagrama de una tarea.
  - Número de acciones desencadenas por SOCIO, a partir de aquellos mensajes que han supuesto una aportación al diagrama obtenido.

La satisfacción se mide por equipos mediante las respuestas al cuestionario SUS (*System Usability Scale*), rellenado por los participantes de manera individual (Sinoo et al., 2018). Los valores de las respuestas son ordinales en una escala de Likert de 5 puntos,

equivaliendo el valor 1 a “totalmente en desacuerdo” y el 5 a “totalmente de acuerdo”. A partir de las repuestas de los integrantes, se seleccionará la mediana para obtener la respuesta del equipo. Finalmente con el promedio de las respuestas del cuestionario SUS de cada equipo, se obtienen la puntuación de satisfacción.

Para medir la calidad de los diagramas realizados por los equipos, se toman de referencia las soluciones ideales de ambas tareas, mostradas en el Anexo B. Las métricas de la calidad son precisión, recall, accuracy, aciertos y error (Giraldo et al., 2018). A continuación se muestran las fórmulas asociadas a estas métricas, cuyos términos se calculan a partir de los valores de la matriz de confusión (*true positive* (TP), *false positive* (FP), *true negative* (TN), *false negative* (FN)):

$$Precisión = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad Accuracy = \frac{TN+TP}{TP+TN+FP\ FN}$$

$$Error = \frac{FP+FN}{TP+TN+FP+FN} \qquad Aciertos = \frac{TP}{Nº\ elementos\ diagrama\ ideal}$$

El Anexo D detalla el significado de estas métricas y el cálculo de los valores de la matriz de confusión.

No todas las métricas expuestas están relacionadas con el chatbot SOCIO y la aplicación web Creately. Algunas de ellas solo se generan al interaccionar con el chatbot. La Tabla 2 muestra todas las métricas mencionadas y además, especifica si están asociadas a ambas herramientas o solo a SOCIO.

VARIABLE	MÉTRICA		HERRAMIENTA
Eficacia	Completitud		Ambas
Eficiencia	Rapidez	Tiempo empleado en completar la tarea	
	Fluidez	Número de mensajes de discusión	
		Número de mensajes enviados a SOCIO Número de mensajes erróneos dirigidos a SOCIO	SOCIO
	Interactividad	Número de mensajes útiles	
		Número de mensajes descriptivos	
		Número de comandos	
		Número de acciones desencadenas	
Satisfacción	Respuestas cuestionario SUS		Ambas
Calidad	Precisión, recall, accuracy, aciertos y error		

Tabla 2: Métricas para la eficiencia, la eficacia, la satisfacción y la calidad.

3.4 Sujetos Experimentales

En el experimento han participado de manera voluntaria 30 estudiantes pertenecientes al Grado en Ingeniería Informática y al Doble Grado en Ingeniería Informática y Matemáticas, de la Escuela Politécnica Superior de Universidad Autónoma de Madrid. Todos ellos han cursado las asignaturas de Análisis y Diseño del Software y Proyecto de Análisis y Diseño del Software, por lo que poseen los conocimientos necesarios sobre diagramas de clases, para realizar las tareas del experimento.

A partir del cuestionario de familiaridad, Anexo B, completado por cada participante como primer paso del experimento, cabe destacar las siguientes características:

- La muestra final consta de 30 sujetos de los cuales 21 son hombres y 9 son mujeres. Se trata de una muestra de sujetos jóvenes, se encuentran en una franja de edad de 19 a 23 años, siendo la media 21,97 y la desviación típica de 0,76.

- En general, los participantes son novatos en el conocimiento y uso de los chatbots. El 70% considera tener un nivel de conocimiento bajo o muy bajo. El 43% nunca ha utilizado un chatbot. El 93% indica que su grado de uso de chatbots es bajo o nulo.

- Los sujetos están muy habituados a las redes sociales, en un intervalo de grado de uso de medio a muy alto, se concentra el 77% de los participantes. Todos utilizan al menos dos redes sociales, siendo una de ellas WhatsApp. También, gran parte de los sujetos suelen utilizar Instagram y Twitter, y una proporción menor Facebook y Telegram.

- La mayoría de los participantes, el 83%, indican que han utilizado alguna vez Telegram, red social que se emplea en el experimento. El 53% de los participantes indican que utilizarán la aplicación móvil de Telegram durante el experimento, el 37% la aplicación web y el 10% la aplicación de escritorio.

- El nivel de conocimiento del 70% de los participantes sobre diagramas de clases está entre nivel medio y alto, de acuerdo a su consideración. A su vez, presentan, en general, un buen nivel de inglés, solo un 3% indica poseer un nivel bajo.

Los sujetos experimentales son asociados al azar a un grupo experimental, de acuerdo con el diseño experimental definido, Tabla 1, y realizan las tareas del experimento en equipos aleatorios de tres. Antes de aplicar los tratamientos del experimento, el chatbot SOCIO y la aplicación web Creately, reciben un breve tutorial sobre los mismos.

3.5 Herramientas y Tareas

En el experimento se emplean dos herramientas de creación de diagramas de clases: el chatbot **SOCIO** y la aplicación web **Creately**.

Se recuerda que SOCIO es un chatbot que ayuda en la elaboración de diagramas de clases mediante la interpretación de oraciones en lenguaje natural, en inglés. Se trata de una herramienta colaborativa, integrada en las redes sociales Twitter y Telegram. En el experimento, los sujetos divididos en equipos utilizan SOCIO en un grupo de Telegram. Los miembros de un grupo son: los integrantes de un equipo y SOCIO, cuyo alias es @modellingBot. A través de este canal, los participantes pueden comunicarse entre sí y con el chatbot. La interacción con SOCIO se realiza mediante comandos (por ejemplo, */talk The shop contains products*), él los interpreta y manda una respuesta. Puede consultarse el Anexo C para obtener información más detallada sobre el funcionamiento y los comandos de SOCIO.

Creately (https://creately.com/app/) es una aplicación web que permite la creación de diversos tipos de diagramas de manera colaborativa, entre ellos, diagramas de clases. En el experimento, Creately se combina con un grupo de Telegram para que, como en el caso de SOCIO, los participantes puedan comunicarse durante las tareas. Puede consultarse el Anexo C para ver la apariencia y el funcionamiento de esta aplicación web.

Antes de la aplicación de los tratamientos, al comienzo del experimento, los participantes rellenan, de manera individual, un informe de consentimiento y un cuestionario de familiaridad, Anexo B. Tras ello, cada herramienta se emplea en la realización de una de las dos tareas que conforman el escenario de ejecución del experimento. Los enunciados y soluciones de las tareas se muestran en el Anexo B.

- **Tarea 1**. Los participantes, en equipos de tres, deben diseñar el diagrama de clases (solo clases y atributos) de una aplicación solicitada por una tienda, la cual desea gestionar sus productos y sus clientes. De acuerdo al diseño del experimento, Tabla 1, los equipos pertenecientes al grupo A realizarán esta tarea con Creately. En caso de pertenecer al grupo B, utilizarán SOCIO.

- **Tarea 2**. Los participantes, en equipos de tres, deben diseñar el diagrama de clases (solo clases y atributos) de una aplicación solicitada por un colegio que desea gestionar sus asignaturas, alumnos y profesores. De acuerdo al diseño del experimento, Tabla 1, los equipos pertenecientes al grupo A realizarán esta tarea con SOCIO. En caso de pertenecer al grupo B, utilizarán Creately.

La duración de cada tarea es de máximo 30 minutos. Durante la realización de las tareas, los participantes no pueden hablar en voz alta, los equipos solo pueden comunicarse a través de su grupo de Telegram. Al finalizar cada tarea, los participantes rellenan de manera individual un cuestionario de satisfacción sobre la herramienta que acaban de aplicar, mostrado en el Anexo B.

Las herramientas SOCIO y Creately, los grupos de Telegram empleados durante las tareas, así como los cuestionarios, recolectan los datos empíricos de las métricas asociadas a la eficiencia, la eficacia, la satisfacción y la calidad, a partir de los cuales se realizará el análisis.

3.6 Ejecución del Experimento

El experimento se llevó a cabo en abril de 2019, en el laboratorio 15 de la Escuela Politécnica Superior de la Universidad Autónoma de Madrid. Se realizaron 4 sesiones, de una hora y media cada una. Las 30 personas que participaron en el experimento de manera voluntaria, fueron asignadas a las sesiones en función de su disponibilidad.

En la primera y la segunda sesión, realizaron el experimento 6 y 9 sujetos, respectivamente. Los participantes de estas sesiones constituyeron el grupo A, de acuerdo con el diseño experimental, Tabla 1. En la tercera y la cuarta sesión, participaron 9 y 6 sujetos, respectivamente, y constituyeron el grupo B. Esta asignación de las sesiones a los grupos experimentales se realizó de manera aleatoria. A su vez, en cada sesión los participantes fueron divididos en equipos de tres, al azar. La Tabla 3 muestra la distribución descrita.

FECHAS	SESIÓN	GRUPO	EQUIPOS
09/04/2019	1	A	1 y 2
10/04/2019	2		3, 4 y 5
11/04/2019	3	B	6, 7 y 8
	4		9 y 10

Tabla 3: Sesiones, grupos y equipos del experimento.

En cada sesión, los participantes eran distribuidos en el laboratorio de tal forma que estuviesen separados, para evitar la comunicación en voz alta (necesario al realizar las tareas) y posibles copias entre equipos.

La Figura 1 ilustra la estructura de cada sesión. Al comienzo, los participantes, de manera individual, firmaban el informe de consentimiento, y rellenaban un cuestionario de familiaridad, Anexo B. Después, la sesión se dividía en dos partes. En la primera parte, se impartía el tutorial de la herramienta que los participantes utilizarían en la tarea 1. Se realizaba la tarea 1, los equipos del grupo A empleaban la aplicación web Creately y los equipos del grupo B, el chatbot SOCIO, de acuerdo al diseño experimental, Tabla 1. Al finalizar la tarea 1, los participantes rellenaban de manera individual el cuestionario de satisfacción asociado a la herramienta utilizada, Anexo B. En la segunda parte, de la misma manera, se impartía el tutorial correspondiente. Se realizaba la tarea 2, el grupo A con SOCIO y el grupo B con Creately, y se completaba el cuestionario de satisfacción al finalizar la tarea.

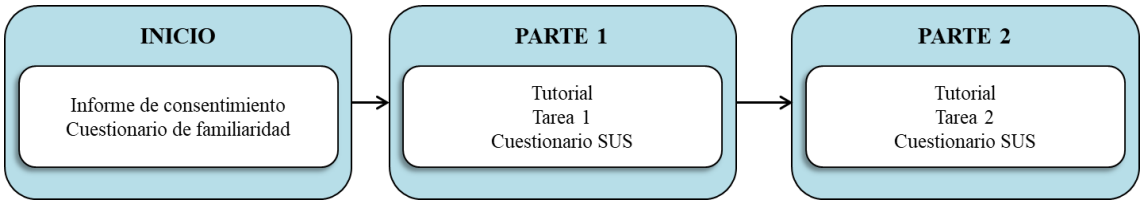


Figura 1: Estructura de una sesión del experimento.

3.7 Amenazas a la Validez

La validez interna se refiere a cuánta confianza tenemos en que los resultados del experimento sean válidos y posibles de interpretar (Campbell & Stanley, 1963). Al diseñar un experimento *crossover*, se ha de tener en cuenta el impacto del número y distribución de los periodos (momentos de aplicación de los tratamientos), y del número y la formación de las secuencias (orden de aplicación de los tratamientos), ya que pueden causar diferentes tipos de amenazas a la validez interna del experimento (Vegas et al., 2016).

Cabe recordar que los tratamientos de este experimento son el chatbot SOCIO y la aplicación web Creately; las secuencias, Creately-SOCIO y SOCIO-Creately; y los periodos son periodo 1, momento en el que se aplica el primer tratamiento (en función de la secuencia) y periodo 2, momento en el que se aplica el segundo tratamiento. Como en el periodo 1 se realiza la tarea 1 y en el periodo 2 se realiza la tarea 2, el efecto que puede causar el periodo se confunde con el efecto que puede causar la tarea realizada.

En este experimento, se abordan las amenazas a la validez interna asociadas al periodo de la siguiente manera:

- Existe una amenaza de aprendizaje por la práctica, ésta tiene lugar cuando las respuestas de los sujetos mejoran a medida que realizan una tarea repetidamente. En este caso, los participantes realizan dos diagramas de clases, y aunque ya disponen de los conocimientos necesarios para elaborarlos, durante el primer diagrama los refrescarán y recuperarán la práctica. La amenaza de aprendizaje puede producir la impresión de que el tratamiento aplicado en último lugar parezca proporcionar mejores resultados. Se puede mitigar comparando los resultados en los dos periodos (de las dos tareas), y estudiando cualquier mejora observada.

- No existe una amenaza de copia entre periodos dado que no hay un intervalo entre ellos (tienen lugar en una misma sesión) y las tareas experimentales son diferentes en cada periodo. Sin embargo, como el experimento se realiza en 4 sesiones diferentes, puede producirse una amenaza de copia si los sujetos de una sesión comentan las tareas llevadas a cabo y sus resultados con participantes de sesiones posteriores. Para mitigar esta posible amenaza, los participantes no fueron informados de que en todas las sesiones se realizarían las mismas tareas. Además, se les pidió que por favor no comentasen nada con los participantes que aún no habían realizado el experimento hasta que se finalizasen todas las sesiones.

- A pesar de que las sesiones no poseen una duración excesivamente larga, una hora y media, se podría producir una amenaza por cansancio o aburrimiento. Los sujetos participan de manera voluntaria y su colaboración no les supone ninguna repercusión en las notas del curso, por lo que podrían sufrir falta de motivación.

En cuanto a las amenazas a la validez interna asociadas a la secuencia:

- No se considera, en principio, que pueda existir una amenaza debida a una secuencia óptima, orden de aplicación de los tratamientos que conduce sistemáticamente a los sujetos experimentales a lograr mejores resultados.

Otra amenaza a la validez interna de los diseños crossover es el *carryover*. Éste tiene lugar cuando un tratamiento es administrado antes de que el efecto del tratamiento anterior haya desaparecido completamente. Como consecuencia, tratamientos aplicados después parecen ser más efectivos que los primeros, si lo primeros potencian la efectividad de los segundos, o menos efectivos si los primeros disminuyen la efectividad de los segundos. Además, en el caso particular de un diseño de crossover como el de este experimento, con el mismo número de periodos que de tratamientos, la interacción entre el período y el tratamiento se confunde intrínsecamente con el efecto del *carryover* y con el efecto de la secuencia, haciendo imposible distinguir cuál de los tres se está produciendo (Vegas et al., 2016).

La validez externa hace referencia a la posibilidad de generalizar los resultados a otros contextos. En este experimento, se trabaja en un contexto de estudiantes universitarios con conocimientos en informática y en concreto, con conocimientos sobre diagramas de clases, por ello, los resultados no son generalizables al ámbito industrial, sino que permanecen en el ámbito académico.

4 Análisis de los Datos

En este capítulo se presenta el análisis de los datos recolectados durante la realización del experimento detallado en el Capítulo 3. En la sección 4.1, se realiza en análisis de datos asociados tanto al chatbot SOCIO como a la aplicación web Creately, empleados para evaluar su eficacia y eficiencia, la satisfacción del usuario y la calidad de los diagramas realizados con ambas herramientas. En la sección 4.2, se analizan los datos generados durante la interacción con SOCIO. Para la realización de todo el estudio estadístico se ha utilizado el lenguaje R (Field et al., 2012).

4.1 Análisis de los Datos de SOCIO y Creately

Sobre los datos recolectados durante el experimento, asociados al chatbot SOCIO y a la aplicación web Creately, en primer lugar, se ha realizado un análisis descriptivo mediante diagramas de caja. Los diagramas muestran los datos de cada métrica agrupados por tratamiento. En el Anexo E, de manera adicional, se analizan los diagramas correspondientes a los datos agrupados por secuencia y tratamiento, y por periodo/tarea y tratamiento.

Tras el análisis descriptivo, los datos han sido analizados mediante el **modelo lineal mixto**, siguiendo (Vegas et al., 2016). Se trata una extensión del modelo lineal general, y es el mejor método para analizar modelos con coeficientes aleatorios (como es el caso de los sujetos del experimento) y dependencia entre datos, debida a medidas repetidas (en un diseño crossover, como el de este experimento, los participantes aplican varios tratamientos, por lo que en cada aplicación se toman las mismas medidas). En particular, se ha ajustado un modelo lineal mixto para cada métrica. Todos los modelos lineales mixtos presentan los mismos factores:

1. Secuencia (Creately-SOCIO o SOCIO-Creately) tienen en cuenta la asignación de los equipos a una combinación de tarea y tratamiento. Los efectos que puede producir la secuencia, los efectos del *carryover* y los de la interacción del tratamiento y la tarea, se confunden en experimentos crossover con un diseño como el de este experimento (Vegas et al., 2016).

2. Periodo (1 o 2), tienen en cuenta la tarea que los equipos tienen que realizar. En este experimento, en el momento de aplicación del primer tratamiento (de acuerdo a la secuencia, Tabla 1), periodo 1, se realiza la tarea 1, y en el momento de aplicación del segundo tratamiento, periodo 2, se realiza la tarea 2. Este hecho produce que los efectos que pueda producir el periodo se confunda con los de la tarea.

3. Tratamiento (Creately o SOCIO), considera la herramienta empleada por los equipos para llevar a cabo las tareas del experimento.

Los resultados del análisis estadístico se complementan calculando el tamaño del efecto del tratamiento, es decir, la magnitud de las diferencias causadas por el tratamiento. Para ello, se calcula la *d* de Cohen para los tratamientos (*d*, en adelante), y su correspondiente

error estándar (*SE*). Se emplean las fórmulas proporcionadas en (Higgins & Green, 2006), mostradas en el Anexo F.

4.1.1 Eficacia

El grado de completitud de las tareas es la métrica empleada para medir la eficacia. La Figura 2 muestra el diagrama de caja correspondiente al grado de completitud con el que los equipos finalizaron las tareas, agrupado por tratamiento. Los datos son muy similares para SOCIO y para Creately, aunque SOCIO presenta valores más bajos de completitud.

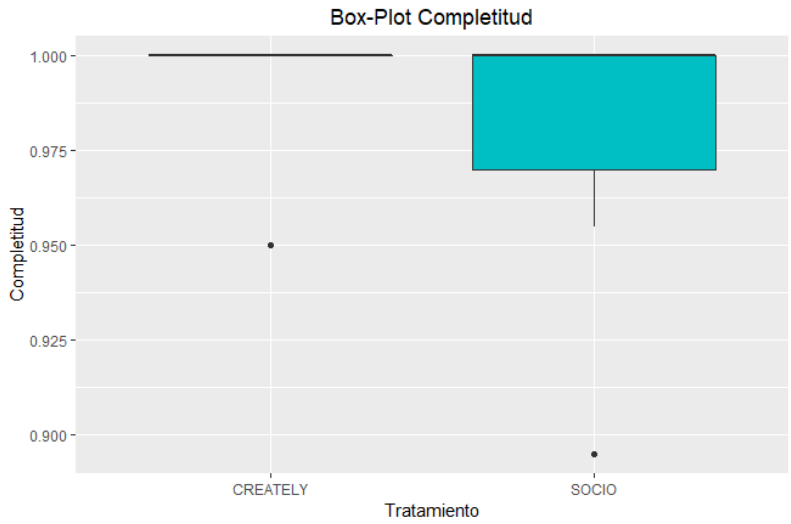


Figura 2: Diagrama de caja del grado de completitud de las tareas para SOCIO y Creately.

La Tabla 4 muestra los resultados del modelo lineal mixto ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	1	0.01	0
Seq	-0.02	0.01	0.23
Treatment	-0.01	0.01	0.28
Period	0	0.01	0.63

Tabla 4: Modelo lineal mixto para la completitud.

Como se puede ver en la Tabla 4, ninguno de los factores es estadísticamente significativo. El grado medio de completitud esperado entre ambos tratamiento es 1, como la diferencia entre ambos tratamiento del grado de completitud es 0.01 (mayor para Creately), no es significativa, lo cual puede ser debido al pequeño tamaño muestral. Finalmente, se obtiene d = -0.52, $SE(d)$ = 0.52. Esto sugiere que el tamaño del efecto es mediano, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos de completitud. En resumen, **parece que Creately y Socio se comportan de manera similar en términos de completitud**.

4.1.2 Eficiencia

El tiempo empleado en realizar las tareas y el número de mensajes de discusión intercambiados entre los miembros de un equipo durante las mismas, son las métricas empleadas para mediar la eficiencia. A continuación, se analizan los resultados obtenidos.

Tiempo

La Figura 3 muestra el diagrama de caja correspondiente al tiempo empleado por los equipos en realizar las tareas, agrupado por tratamiento. Se observan mejores resultados, es decir, tiempos más bajos, para SOCIO.

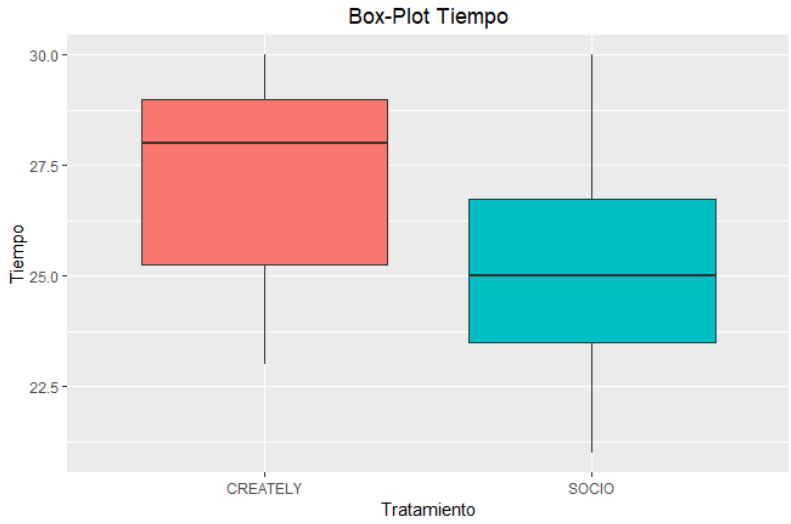


Figura 3: Diagrama de caja para el tiempo empleado en completar la tarea con SOCIO y Creately.

La Tabla 5 muestra los resultados del modelo lineal mixto ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	28.2	1.16	0
Seq	-2	1.16	0.12
Treatment	-1.8	1.16	0.15
Period	-0.2	1.16	0.86

Tabla 5: Modelo lineal mixto para el tiempo empleado en completar la tarea.

Como se puede ver en la Tabla 5, ninguno de los tres factores son estadísticamente significativos. Los tiempos para Creately son de media 1.8 minutos superiores a los de SOCIO, este valor no es significativo comparándolo con 28 minutos, el tiempo medio entre ambos tratamientos. Finalmente, se obtiene d = -0.68 y $SE(d)$ = 0.57, por lo que el tamaño del efecto es mediano, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos del tiempo. En la Figura 3, así como en las Figuras 32 y 33 del Anexo E, los gráficos muestran mejores tiempos para SOCIO, es decir, tiempos más bajos, pero no podemos decir que este hecho afecte a la eficiencia debido a que el número de sujetos es pequeño.

Número de mensajes de discusión

La Figura 4 muestra el diagrama de caja del número de mensajes de discusión agrupados por tratamiento. Se aprecia que se generan más mensajes de discusión empleando Creately que empleando SOCIO.

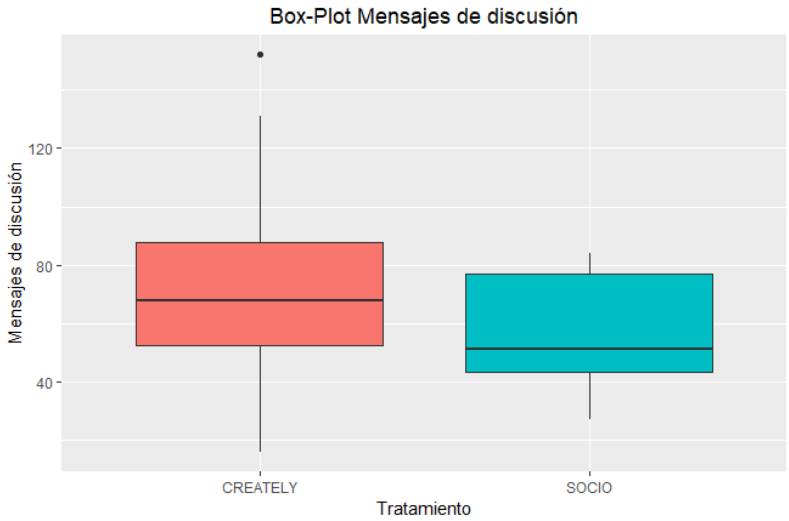


Figura 4: Diagrama de caja para el número de mensajes de discusión de SOCIO y Creately.

La Tabla 6 muestra los resultados del modelo lineal mixto, ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	82.2	15.06	0
Seq	-7.8	18.6	0.7
Treatment	-18.4	10.37	0.11
Period	-5.8	10.37	0.59

Tabla 6: Modelo lineal mixto para el número de mensajes de discusión.

Como se puede ver en la Tabla 6, ninguno de los factores son estadísticamente significativos. El número medio de mensajes de discusión es 82.2, por lo que el hecho de que Creately obtenga de media 18.4 mensajes más que SOCIO no es una diferencia significativa, y puede ser debida a un tamaño muestral pequeño. Finalmente, se obtiene d = -0.57 y $SE(d)$ = 0.28, por lo que el tamaño del efecto es mediano, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos del número de mensajes de discusión. En resumen, **Creately parece generar un mayor número de mensajes de discusión que SOCIO, requiriendo SOCIO menos esfuerzo**.

4.1.3 Satisfacción

La Figura 5 muestra el diagrama de caja correspondiente a las puntuaciones de satisfacción de los equipos, agrupadas por tratamiento. Se observa que las puntuaciones para SOCIO son más altas que para Creately.

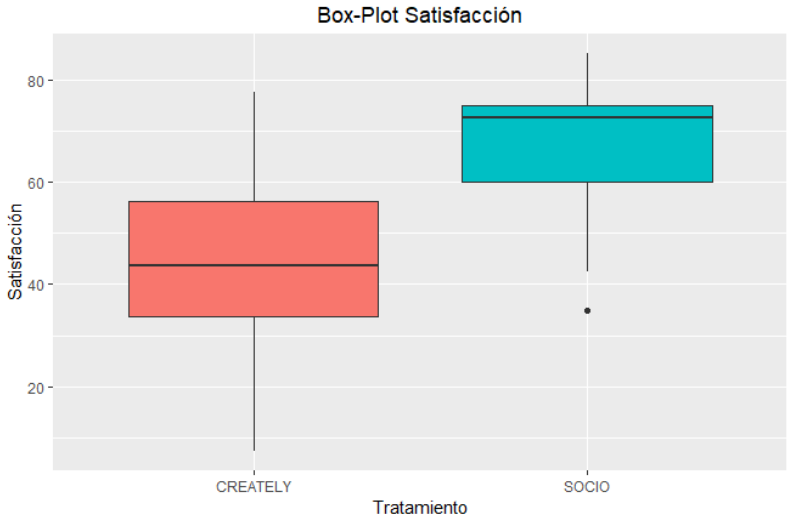


Figura 5: Diagrama de caja para las puntuaciones de satisfacción para SOCIO y Creately**.**

La Tabla 7 muestra los resultados del modelo lineal mixto ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	58	7.17	0
Seq	-8	7.17	0.30
Treatment	22.5	7.17	**0.014**
Period	-21	7.17	**0.019**

Tabla 7: Modelo lineal mixto para la satisfacción.

Como se puede ver en la Tabla 7, el tratamiento y el periodo son estadísticamente significativos, lo cual afecta en la satisfacción. SOCIO obtiene de media 22.5 puntos de satisfacción más que Creately. En relación a las tareas, en la tarea 2 se obtienen 21 puntos menos que en la tarea 1. Esta diferencia de puntuaciones se ve también reflejada en la Figura 5 y en la Figura 37 del Anexo E. Comparando estos datos con los 58 puntos de media, los efectos del tratamiento y del periodo parecen grandes. Finalmente, se obtiene $d = 1.17$ y $SE(d) = 0.67$, hecho que sugiere de nuevo que el tamaño del efecto es grande, de acuerdo a las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos de la satisfacción. En resumen, **SOCIO parece satisfacer a los usuarios en mayor medida que Creately.**

4.1.4 Calidad

La calidad de los diagramas realizados con SOCIO y Creately se evalúa a partir de las métricas accuracy, precisión, recall, aciertos y error. Los datos asociados a estas métricas se analizan a continuación.

Accuracy

La Figura 6 muestra el diagrama correspondiente a las puntuaciones de accuracy asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately. Como

se puede ver, las puntuaciones parecen muy similares para SOCIO y Creately, aunque las de SOCIO parecen más dispersas.

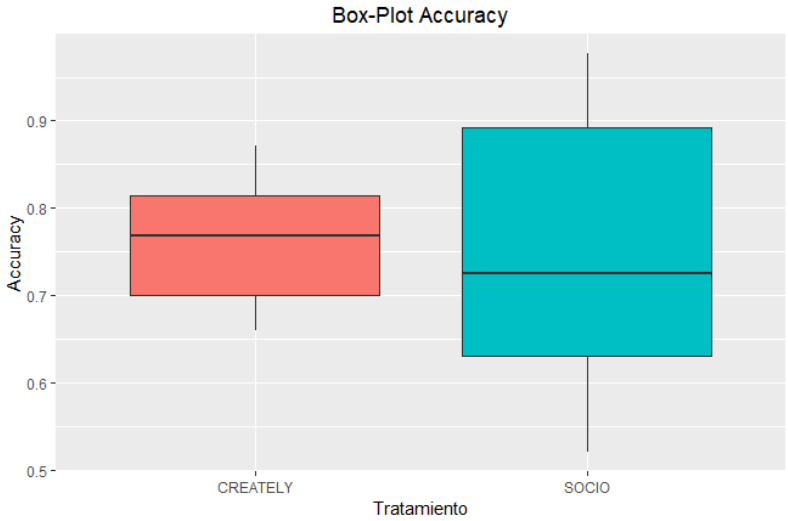


Figura 6: Diagrama de caja de las puntuaciones de accuracy para SOCIO y Creately.

La Tabla 8 muestra los resultados del modelo lineal mixto, ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.78	0.03	0
Seq	0.12	0.03	**0**
Treatment	-0.01	0.03	0.77
Period	-0.17	0.03	**0**

Tabla 8: Modelo lineal mixto para accuracy.

Como se puede ver en la Tabla 8, la secuencia y el periodo son estadísticamente significativos, lo cual afecta a la métrica accuracy. En la secuencia SOCIO-Creately (en adelante, SC-CR), las puntuaciones de accuracy son de media 0.12 más altas que en la secuencia Creately- SOCIO (en adelante, CR-SC). Respecto al periodo, en la tarea 1 las puntuaciones de accuracy son de media 0.17 más altas que en la tarea 2. Estas diferencias se reflejan en la Figura 38 y en la Figura 39 del Anexo E, respectivamente. Comparando estos valores con la media para los dos tratamientos 0.78, parece que los efectos de la secuencia y el periodo son medianos. El hecho de que la secuencia sea estadísticamente significativa implica que la interacción entre la tarea y el tratamiento o los efectos de *carryover* se materializaron (Vegas et al., 2016). A su vez, los efectos del tratamiento se revierten dependiendo de la secuencia. Esta reversión se aprecia en la Figura 38 del Anexo E, la gráfica muestra como en la secuencia CR-SC, Creately obtiene mejores resultados, y la secuencia SC-CR, ocurre lo contrario, los resultados son mejores para SOCIO. Finalmente, se obtiene d = -0.07 y $SE(d)$ = 0.54, lo cual sugiere que el tamaño del efecto es pequeño, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos de accuracy.

Precisión

La Figura 7 muestra el diagrama correspondiente a las puntuaciones de precisión asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately. Como se puede ver, las puntuaciones parecen muy similares para SOCIO y Creately, aunque las de SOCIO se muestran más dispersas.

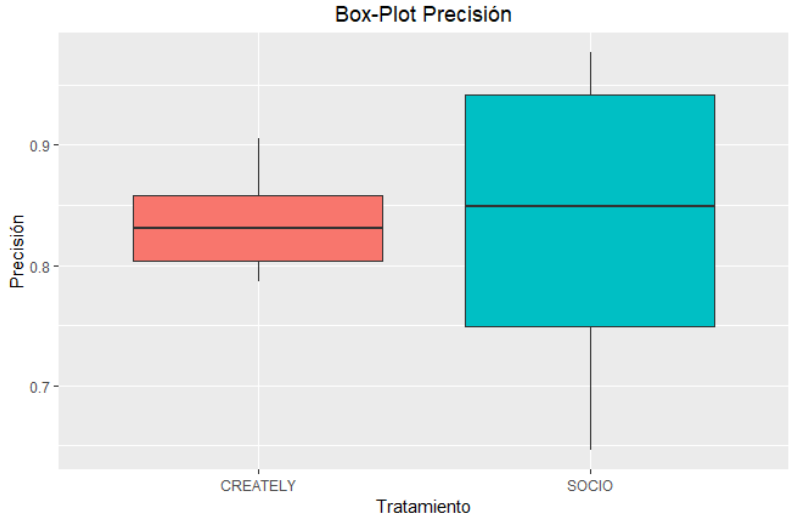


Figura 7: Diagrama de caja de las puntuaciones de precisión para SOCIO y Creately**.**

La Tabla 9 muestra los resultados del modelo lineal mixto, ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.84	0.02	0
Seq	0.11	0.02	**0**
Treatment	0	0.02	0.89
Order	-0.11	0.02	**0**

Tabla 9: Modelo lineal mixto para la precisión.

Como se puede ver en la Tabla 9, la secuencia y el periodo son estadísticamente significativos, lo cual afecta a la precisión. Para la secuencia SC-CR, las puntuaciones son de media 0.11 más altas que para la secuencia CR-SC. En cuanto al periodo, para la tarea 1 las puntuaciones son de media 0.11 más altas que para la tarea 2. Estas diferencias se ven reflejadas en las Figuras 40 y 41 del Anexo E, respectivamente. Comparando estas diferencias, con el valor medio de los tratamientos 0.84, los efectos del efecto y el periodo no parecen grandes. El hecho de que la secuencia sea estadísticamente significativa implica que la interacción entre la tarea y el tratamiento o los efectos de *carryover* se materializaron (Vegas et al., 2016). A su vez, los efectos del tratamiento se revierten dependiendo de la secuencia. La reversión es más significativa para SOCIO que para Creately, ya que Creately mantiene puntuaciones similares en ambas secuencias, Figura 40 del Anexo E. Finalmente, se obtiene d = -0.03 y $SE(d)$ = 0.46, lo cual sugiere que el tamaño del efecto es pequeño, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos de la precisión.

Recall

La Figura 8 muestra el diagrama correspondiente a las puntuaciones de recall asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately. Como se puede ver, las puntuaciones parecen muy similares para SOCIO y Creately, aunque las de Creately se muestran superiores y las de SOCIO más dispersas.

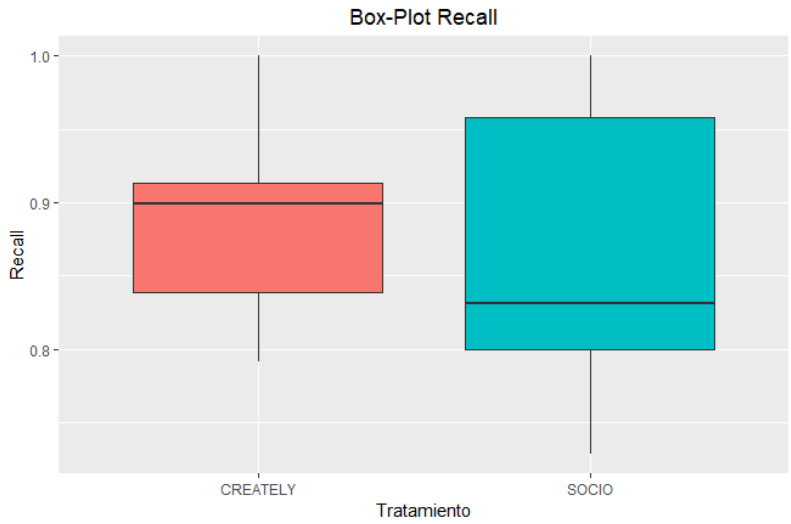


Figura 8: Diagrama de caja de las puntuaciones de recall para SOCIO y Creately.

La Tabla 10 muestra los resultados del modelo lineal mixto, ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.93	0.02	0
Seq	0.04	0.02	0.15
Treatment	-0.02	0.02	0.35
Period	-0.11	0.02	**0**

Tabla 10: Modelo lineal mixto para la variable recall.

Como se puede ver en la Tabla 10, el periodo (la tarea desarrollada) es estadísticamente significativo, lo cual afecta a las puntuaciones de recall. Las puntuaciones para la tarea 1 son de media 0.11 superiores que para la tarea 2, esta diferencia se ve reflejada en la Figura 43 del Anexo E. Sin embargo, por una parte, parece que el efecto del periodo no es grande comparando la diferencia con el valor medio esperado, 0.93. Por otra parte, la Figura 42 del Anexo E refleja que los efectos del tratamiento se revierten dependiendo de la secuencia. Finalmente, se obtiene d = -0.29 y $SE(d)$ = 0.57, lo cual sugiere que el tamaño del efecto es pequeño, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos de recall.

Aciertos

La Figura 9 muestra el diagrama correspondiente a las puntuaciones de aciertos asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately. Como

se puede ver, las puntuaciones parecen muy similares para SOCIO y Creately, aunque las de SOCIO aparecen dispersas.

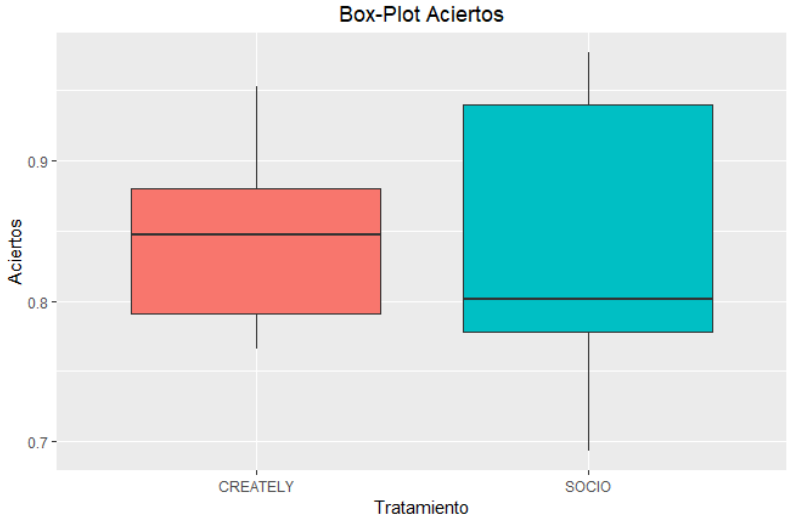


Figura 9: Diagrama de las puntuaciones de aciertos para SOCIO y Creately.

La Tabla 11 muestra los resultados del modelo lineal mixto, ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.89	0.02	0
Seq	0.04	0.02	0.11
Treatment	0	0.02	0.98
Period	-0.12	0.02	**0**

Tabla 11: Modelo lineal mixto para la variable aciertos.

Como se puede ver en la Tabla 11, el periodo (la tarea desarrollada) es estadísticamente significativo, lo cual afecta a las puntuaciones de los aciertos. En la tarea 1, se obtiene de media una puntuación de aciertos 0.12 más elevada que en la tarea 2. Esta diferencia se ve reflejada en la Figura 45 del Anexo E. Comparando esta diferencia con el valor medio esperado, 0.89, por una parte, parece que los efectos del periodo no son muy grandes. Por otra parte, la Figura 44 del Anexo E, refleja que los efectos del tratamiento se revierten dependiendo de la secuencia. Finalmente, se obtiene $d = 0$ y $SE(d) = 0.57$, lo cual sugiere que el tamaño del efecto es pequeño, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos de aciertos.

Error

La Figura 10 muestra el diagrama correspondiente a las puntuaciones de error asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately. Como se puede ver, las puntuaciones parecen similares para SOCIO y Creately, aunque las de SOCIO parecen más dispersas.

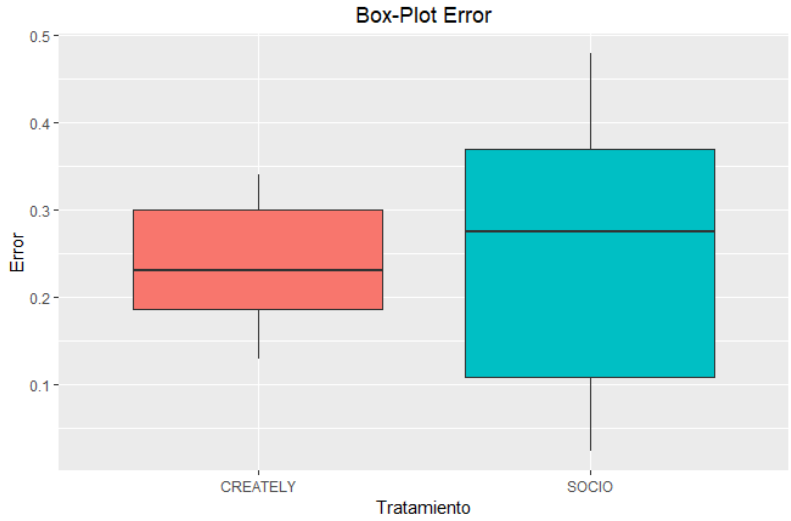


Figura 10: Diagrama de caja de las puntuaciones de error para SOCIO y Creately.

La Tabla 12 muestra los resultados del modelo lineal mixto, ajustado para analizar los datos.

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.21	0.03	0
Seq	-0.12	0.03	**0**
Treatment	0.01	0.03	0.77
Period	0.17	0.03	**0**

Tabla 12: Modelo lineal mixto para el error.

Como se puede ver en la Tabla 12, la secuencia y el periodo son estadísticamente significativos, lo cual afecta al error. En la secuencia CR-SC se obtiene de media 0.12 puntos de error más que en la secuencia SC-CR. En la tarea 2 se obtienen de media 0.17 puntos de error más que en la tarea 1. Estas diferencias se reflejan en las Figuras 46 y 47 del Anexo E, respectivamente. Comparando estos valores con los 0.21 puntos de media, los efectos de la secuencia y el periodo parecen grandes. El hecho de que la secuencia sea estadísticamente significativa implica que la interacción entre la tarea y el tratamiento o los efectos de *carryover* se materializaron (Vegas et al., 2016). A su vez, los efectos del tratamiento se revierten dependiendo de la secuencia, Figura 46 del Anexo E. Finalmente, se obtiene d = 0.07 y $SE(d)$ = 0.54, lo cual sugiere que el tamaño del efecto es pequeño, de acuerdo con las reglas generales (Borenstein et al., 2011), materializado para los tratamientos en términos del error.

4.2 Análisis de Otros Datos de SOCIO

La interacción de los equipos con SOCIO, durante la realización de las tareas, ha generado más datos: número de mensajes dirigidos al chatbot, mensajes erróneos, mensajes útiles, mensajes descriptivos, comandos y acciones desencadenadas. En primer lugar, se ha realizado un análisis descriptivo mediante diagramas de caja, donde se muestran los datos agrupados por tarea. Tras ello, se han realizado una serie de **test-t** para muestras independientes (Field et al., 2012), uno por cada métrica, para comparar la media de las diferentes interacciones generadas durante las tareas 1 y 2. Los resultados del test-t se

complementan, calculando el tamaño del efecto asociado a la tarea, con la *d* de Cohen y su error estándar (SE), siguiendo la fórmula de (Borenstein et al, 2011), expuestas en el Anexo F.

4.2.1 Mensajes enviados a SOCIO

Dentro del número de mensajes enviados al chatbot SOCIO por un equipo durante la realización de la tarea, se incluye mensajes erróneos, mensajes sin y con aportación al diagrama obtenido, y todo tipo de comandos. Se considera mensaje erróneo a aquel mensaje cuya intención inicial era ir dirigido a SOCIO, pero debido a errores en la escritura, finalmente no es enviado al chatbot o en caso de ser enviado, no es comprendido. A continuación, se realiza el análisis del número de mensajes dirigidos a SOCIO y del número de mensajes erróneos.

Número de mensajes enviados a SOCIO

La Figura 11 muestra el diagrama de caja correspondiente al número de mensajes dirigidos al chatbot. Se puede ver que el número de mensajes dirigidos a SOCIO es menor en la tarea 1 que en la tarea 2.

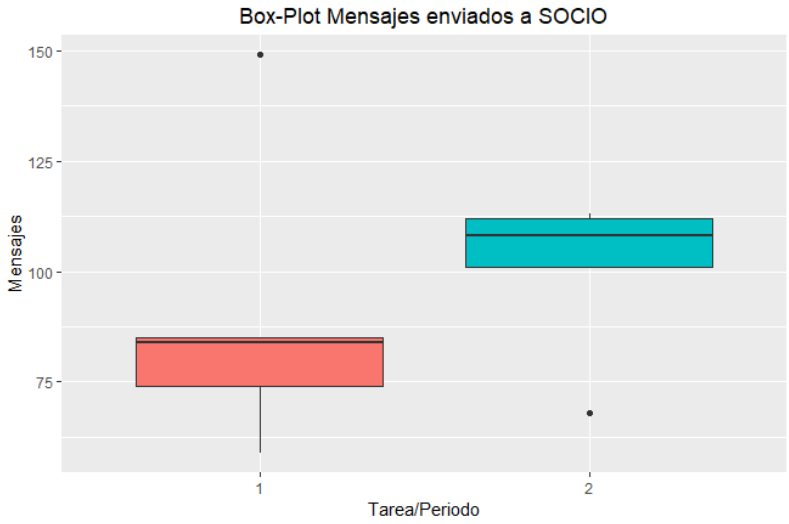


Figura 11: Diagrama de caja para el número de mensajes enviados a SOCIO.

La Tabla 13 presenta los resultados del test-t, comparando la media de mensajes dirigidos a SOCIO en las tareas 1 y 2.

Tarea 1	Tarea 2	95% CI	*p*-valor
90.2	100.4	[-52.86, 32.46]	0.58

Tabla 13: Número medio de mensajes dirigidos a SOCIO en las tareas 1 y 2, resultado del test-t para el número de mensajes enviados a SOCIO durante las tareas.

Como se aprecia en la Tabla 13, la media de mensajes enviados a SOCIO en la tarea 2 es mayor que en la tarea 1. Sin embargo, esta diferencia no es estadísticamente significativa (*p-valor* = 0.58), y un amplio intervalo de confianza (95% *CI* = [-52.86, 32.46]) se materializó. Esto sugiere **que el mayor número de mensajes dirigidos al**

chatbot en la segunda tarea, puede ser debido a una causa aleatoria aislada. Finalmente, un pequeño tamaño del efecto, *d* = -0.37 y *SE(d)* = 0.41, se materializa en el experimento.

Número de mensajes erróneos enviados a SOCIO

La Figura 12 muestra el diagrama de caja asociado al número de mensajes erróneos dirigidos a SOCIO durante la realización de las tareas 1 y 2. Se observa que el número de mensajes erróneos generados es mayor en la tarea 2 que en la tarea 1.

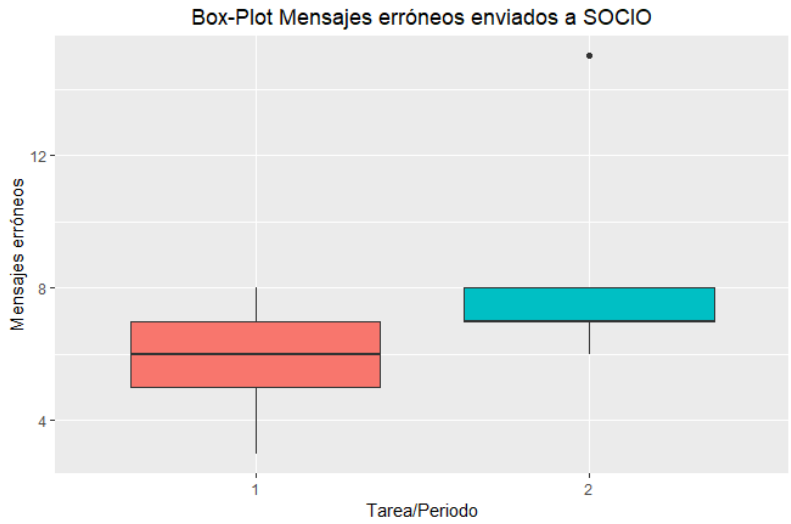


Figura 12: Diagrama de caja para el número de mensajes erróneos dirigidos a SOCIO.

La Tabla 14 presenta los resultados del test-t, comparando la media de mensajes erróneos dirigidos a SOCIO en las tareas 1 y 2.

Tarea 1	Tarea 2	95% CI	*p*-valor
5.8	8.6	[-7.3, 1.7]	0.18

Tabla 14: Media de mensajes erróneos dirigidos al chatbot durante las tareas 1 y 2, resultados del test-t para el número de mensajes erróneos dirigidos a SOCIO durante las tareas.

Como se puede observar en la Tabla 14, la media de mensajes erróneos es mayor en la tarea 2 que en la tarea 1. Sin embargo, de nuevo, esta diferencia no es significativa (*p-valor* = 0.18), y un amplio intervalo de confianza se materializa (95% *CI* = [-7.3, 1.7]). Finalmente, un gran tamaño del efecto, *d* = -0.96 y SE(d) = 0.44, se materializa en el número de mensajes erróneos dirigidos a SOCIO. Esto sugiere, que a pesar de que los resultados no son significativos, **el número de mensajes erróneos generados en la tarea 2 es considerablemente mayor que los generados en la tarea 1.**

4.2.2 Mensajes útiles dirigidos a SOCIO

El número de mensajes útiles enviados a SOCIO hace referencia a aquellos mensajes que han supuesto una aportación al diagrama de clases realizado por los equipos durante las tareas. Los mensajes útiles son de dos tipos: mensajes descriptivos (por ejemplo, */talk*

the house contains rooms) y comandos (mensajes imperativos, por ejemplo, */talk add house*). A continuación, se realiza el análisis del número de mensajes útiles, y del número de mensajes descriptivos y comandos.

Número de mensajes útiles enviados a SOCIO

La Figura 13 muestra el diagrama de caja correspondiente al número de mensajes dirigidos a SOCIO que supusieron un aporte al diagrama de clases. Se observa que el número de mensajes útiles en la tarea 2 es mayor que en la tarea 1.

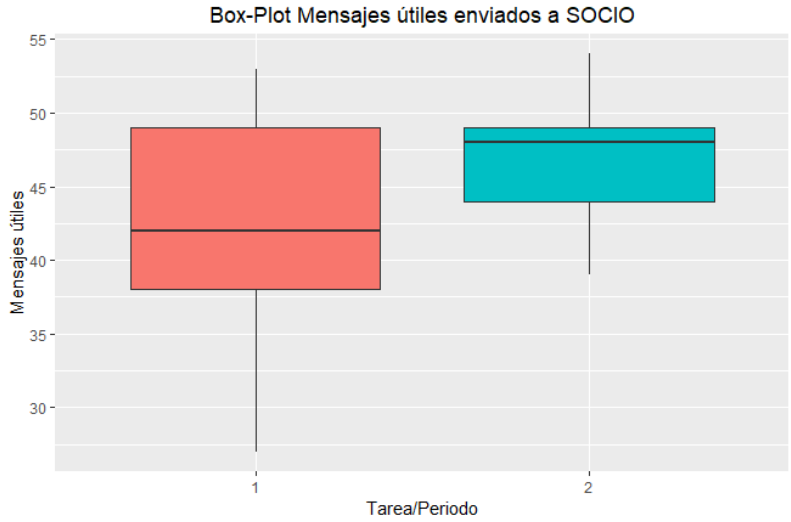


Figura 13: Diagrama de caja para el número de mensajes útiles enviados a SOCIO.

La Tabla 15 presenta los resultados del test-t, comparando la media de mensajes útiles dirigidos a SOCIO en las tareas 1 y 2.

Tarea 1	**Tarea 2**	**95% CI**	***p*-valor**
41.8	46.8	[-17.56, 7.56]	0.37

Tabla 15: Media de mensajes útiles dirigidos al chatbot durante las tareas 1 y 2, resultados del test-t para el número de mensajes útiles enviados a SOCIO durante las tareas.

Como se aprecia en la Tabla 15, la media de mensajes útiles enviados a SOCIO en la tarea 2 es mayor que en tarea 1. Sin embargo, esta diferencia no es estadísticamente significativa (*p-valor* = 0.37), y un intervalo de confianza amplio (95% *CI* = [-17.56, 7.56]) se materializó. Finalmente, un tamaño del efecto medio, *d* = -0.61 y *SE(d)* = 0.41, se materializa en el número de mensajes útiles. Esto sugiere, que a pesar de que los resultados no son significativos, **el número de mensajes útiles generados en la tarea 2 es mayor que los generados en la tarea 1.**

Número de mensajes descriptivos

La Figura 14 muestra el diagrama de caja correspondiente al número de mensajes descriptivos dirigidos a SOCIO que supusieron un aporte al diagrama de clases obtenido. Se observa que el número de mensajes descriptivos es similar en ambas tareas, siendo ligeramente mayor en la tarea 2 y más disperso en la tarea 1.

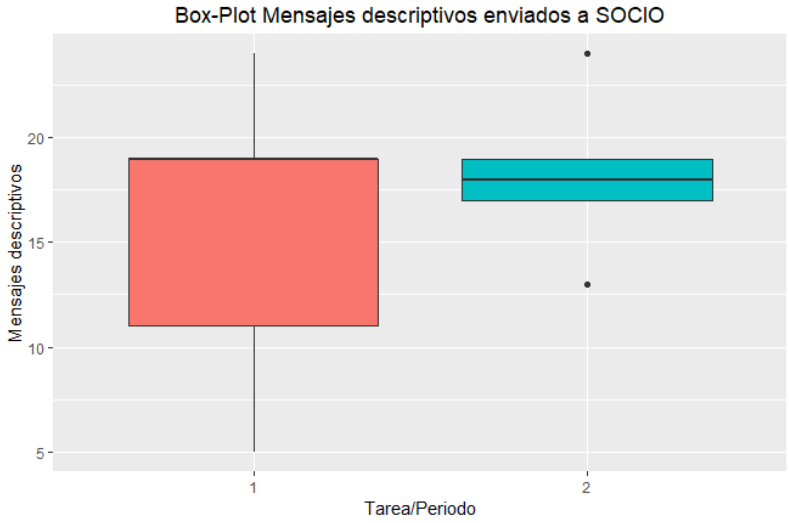


Figura 14: Diagrama de caja para el número de mensajes descriptivos enviados a SOCIO.

La Tabla 16 presenta los resultados del test-t, comparando la media de mensajes descriptivos dirigidos a SOCIO en las tareas 1 y 2.

Tarea 1	**Tarea 2**	**95% CI**	***p*-valor**
15.6	18.2	[-11.9, 6.7]	0.52

Tabla 16: Media de mensajes descriptivos dirigidos al chatbot durante las tareas 1 y 2, resultados del test-t para los mensajes descriptivos enviados a SOCIO durante las tareas.

Como se aprecia en la Tabla 16, la media de mensajes descriptivos enviados a SOCIO en la tarea 2 es mayor que en tarea 1. Esta diferencia no es estadísticamente significativa (*p-valor* = 0.52), y se materializó un amplio intervalo de confianza (95% *CI* = [-11.9, 6.7]). Finalmente, una tamaño del efecto pequeño, *d* = 0.43 y *SE(d)* = 0.41, se materializa en el número de mensajes descriptivos. Esto sugiere, que a pesar de que los resultados no son significativos, **el número de mensajes descriptivos generados en ambas tareas es similar**.

Número de comandos

La Figura 15 muestra el diagrama de caja correspondiente al número de comandos enviados a SOCIO empleados para realizar el diagrama de clases. Se observa que el número de comandos enviados en la tarea 2 en mayor que en la tarea 1.

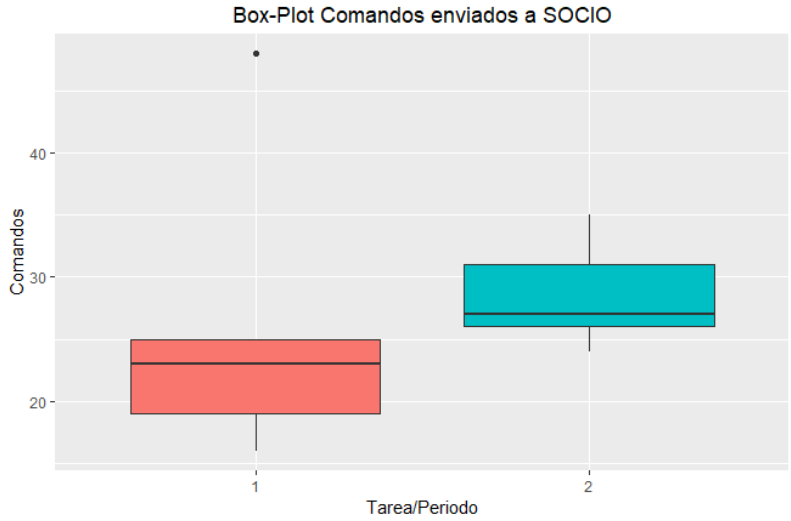


Figura 15: Diagrama de caja para el número de comandos dirigidos a SOCIO.

La Tabla 17 presenta los resultados del test-t, comparando la media de comandos enviados a SOCIO en las tareas 1 y 2.

Tarea 1	Tarea 2	95% CI	*p*-valor
26.2	28.6	[-17.87, 13.07]	0.71

Tabla 17: Media de mensajes comandos dirigidos al chatbot durante las tareas 1 y 2, resultados del test-t para el número de comandos dirigidos a SOCIO durante las tareas.

Como se aprecia en la Tabla 17, la media de comandos enviados a SOCIO en la tarea 2 es mayor que en la tarea 1. Sin embargo, esta diferencia no es estadísticamente significativa (*p-valor* = 0.71), y un amplio intervalo de confianza (95% *CI* = [-17.87, 13.07]) se materializó. Esto sugiere que el mayor número de comandos dirigidos al chatbot en la segunda tarea, puede ser debido a una causa aleatoria aislada. Finalmente, un tamaño del efecto pequeño, *d* = -0.25 y *SE(d)* = 0.40, se materializa en el número de comandos.

4.2.3 Acciones desencadenadas

Los mensajes enviados a SOCIO, para realizar el diagrama de clases, son interpretados por el chatbot. SOCIO puede desencadenar tres tipos de acciones a partir de su interpretación, crear un elemento en el diagrama, modificarlo o eliminarlo.

La Figura 16 muestra el diagrama de caja correspondiente al número de acciones desencadenadas por SOCIO durante la realización del diagrama de clases. Se observa que el número de acciones desencadenadas en la tarea 2 en mayor que en la tarea 1.

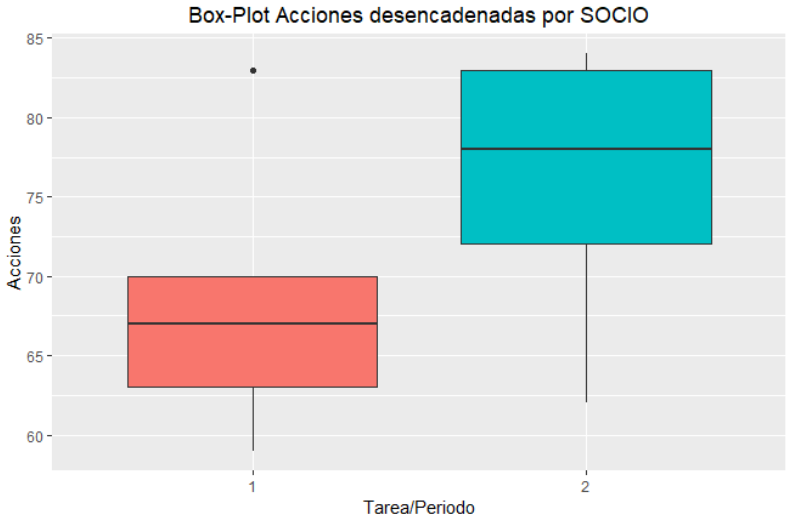


Figura 16: Diagrama de caja para el número de acciones desencadenadas por SOCIO durante la realización del diagrama.

La Tabla 18 presenta los resultados del test-t, comparando la media de acciones desencadenadas por SOCIO en las tareas 1 y 2.

Tarea 1	**Tarea 2**	**95% CI**	***p*-valor**
68.4	75.8	[-20.69, 5.89]	0.23

Tabla 18: Media de acciones desencadenadas por SOCIO durante las tareas 1 y 2, resultados del test-t para el número de acciones desencadenadas por SOCIO durante las tareas.

Como se aprecia en la Tabla 18, la media de acciones desencadenas en la tarea 2 es mayor que en la tarea 1. Esta diferencia no es estadísticamente significativa (*p-valor* = 0.23), y se materializó un intervalo de confianza de amplitud media (95% *CI* = [-20.69, 5.89]). Finalmente, un tamaño del efecto grande, *d* = -0.81 y *SE(d)* = 0.43, se materializa en el número de acciones desencadenadas. Esto sugiere, que a pesar de que los resultados no son significativos, **el número de acciones desencadenadas en la tarea 2 es mayor que los generados en la tarea 1.**

5 Discusión de los Resultados

En este capítulo se discuten e interpretan los resultados obtenidos en el análisis de los datos del experimento, teniendo en cuenta las hipótesis de la investigación. La sección 5.1 expone la discusión e interpretación de los resultados del análisis de los datos de SOCIO y Creately, descrito en la sección 4.1. La sección 5.2 expone la discusión e interpretación de los resultados del análisis de los datos obtenidos durante la interacción con el chatbot SOCIO, detallado en la sección 4.2.

5.1 Resultados para SOCIO y Creately

La Tabla 19 resume los resultados de los modelos lineales mixtos ajustados para cada métrica, el tamaño del efecto que produce el tratamiento, así como la repercusión del análisis sobre las hipótesis asociadas a las variables respuesta. En la columna Hipótesis, el símbolo * denota la existencia de diferencias estadísticamente significativas, el símbolo ~ indica la existencia de diferencias no estadísticamente significativas y el símbolo - indica la existencia de diferencias irrelevantes. En la columna Modelo Lineal Mixto, el símbolo x indica que el factor es estadísticamente significativo, mientras que - indica que no lo es.

Variable	**Hipótesis**		**Métrica**	**Modelo Lineal Mixto**			**Tamaño del efecto**
				Secuencia	**Tratamiento**	**Periodo**	
Eficacia	H.1.1.0	-	Completitud	-	-	-	Mediano
Eficiencia	H.1.2.0	~	Tiempo empleado en completar la tarea	-	-	-	Mediano
			Número de mensajes de discusión	-	-	-	
Satisfacción	H.1.3.0	*	Respuestas cuestionario SUS	-	x	x	Grande
Calidad	H.1.4.0	-	Precisión	x	-	x	Pequeño
			Recall	-	-	x	
			Accuracy	x	-	x	
			Aciertos	-	-	x	
			Error	x	-	x	

Tabla 19: Resumen de los resultados experimentales para SOCIO y Creately.

La **eficacia**, medida a través de la completitud de las tareas, parece similar para el chatbot SOCIO y la aplicación web Creately. Para la métrica completitud, ningún factor es estadísticamente significativo. En concreto, el tratamiento no produce diferencias estadísticamente significativas, lo cual puede ser debido a un tamaño muestral pequeño (10 equipos), pero por el momento la hipótesis H.1.1.0 no puede ser rechazada.

En cuando a la **eficiencia**, medida a través del tiempo y los mensajes de discusión empleados en la realización de la tarea, parece que el chatbot SOCIO es más eficiente que

Creately, es decir, que se requiere menor esfuerzo al utilizar SOCIO. Se ha de recalcar que SOCIO parece más eficiente porque los diagramas de caja del análisis descriptivo muestran tiempos más bajos para el chatbot y un menor número de mensajes de discusión, sin embargo no se puede afirmar que lo sea. Ningún factor afecta al tiempo y a los mensajes de discusión. En particular, las diferencias producidas por el tratamiento no son significativas, lo cual puede ser debido al tamaño muestral, que es pequeño. Por lo tanto, ante diferencias no significativas, la hipótesis H.1.2.0 no puede ser rechazada.

En términos de **satisfacción**, el chatbot SOCIO parece satisfacer en mayor medida a los usuarios que Creately. Por una parte, el tratamiento es un factor significativo en los resultados de satisfacción, las diferencias que produce parecen grandes y el tamaño del efecto obtenido también es grande, por lo tanto, se rechaza la hipótesis H.1.3.0. Por otra parte, el factor periodo/tarea (ambos elementos se confunden en este experimento), es también significativo en los resultados de satisfacción y las diferencias que produce parecen grandes. Las puntuaciones tanto de SOCIO como de Creately cuando son aplicados en la tarea 2, son menores en comparación con sus respectivas puntuaciones para la tarea 1. Este hecho podría deberse a que la tarea 2 haya resultado más difícil, y esta dificultad haya generado niveles de satisfacción menores. También, podría estar afectando el aplicar el tratamiento en segundo lugar o quizás, que los efectos de aplicar la primera herramienta no hayan desaparecido al aplicar la segunda (*carryover*), sobre todo en la secuencia SC-CR. En esta secuencia, las puntuaciones de Creately son, con mucha diferencia, inferiores a las de SOCIO.

Los resultados de satisfacción se pueden contrastar con los aspectos positivos y negativos de los tratamientos, destacados por los participantes en las preguntas abiertas de los cuestionarios de satisfacción, y con sus preferencias. El 80% indicó que prefiere SOCIO a Creately. Muchos de los participantes señalaron que el chatbot es cómodo, sencillo y permite la creación de diagramas de manera rápida. Valoran también positivamente la interacción a través del lenguaje natural, la visualización de las modificaciones del diagrama de manera inmediata (a través de imágenes) y su integración en redes sociales. Sobre Creately destacaron el poder trabajar en equipo en tiempo real respetando el trabajo de los demás (se bloquean los elementos que están siendo modificados). A su vez, consideran que se pueden crear clases y atributos de manera rápida y sencilla. Reflejando los resultados del análisis, en la secuencia SC-CR, el número de aspectos positivos mencionados sobre Creately fue menor, en comparación con la secuencia CR-SC. Además, consideraciones positivas realizadas por los participantes de la secuencia CR-SC (interfaz intuitiva, agradable, fácil de usar), son destacadas como carencias por los participantes de la secuencia SC-CR (poco intuitiva, incómoda, difícil modificar ciertos elementos).

En cuanto a los aspectos negativos, respecto al chatbot SOCIO, en general, los participantes consideran difícil organizarse y mantener una conversación en la versión móvil. Si colaboran muchas personas en la elaboración del diagrama, las imágenes enviadas por SOCIO como respuesta se van encadenando y resulta complicado visualizar el resto de mensajes. Valoraron también negativamente el comando */undo*, pues deshace la última acción realizada por el grupo, no por el usuario que envía el comando. Sugieren acortar el comando */talk* (*/t*), la incorporación de comandos para acciones como crear o borrar *(/make*, */remove*) y sobre todo, una modificación del comando */undo* (deshacer acción propia o añadir mensaje de confirmación antes de efectuarse). Por otra parte, sobre Creately, indican que las actualizaciones son lentas, en ocasiones la herramienta no

reacciona y se generan elementos incómodos imposibles de borrar (cuadrados de colores). Sugieren, en general, corregir los fallos de la aplicación y actualizaciones más rápidas. En el apartado de sugerencias de mejora, también es notable el descontento de los usuarios de la secuencia SC-CR. Varios participantes proponen como mejora no volver a utilizar la herramienta, utilizar otra o eliminarla y empezar de nuevo.

Finalmente, respecto a la **calidad** de los diagramas, el tratamiento no es estadísticamente significativo para ninguna de sus métricas (precisión, recall, accuracy, aciertos y error), lo cual puede ser debido a un tamaño muestral pequeño. Por el momento, al no generar el tratamiento diferencias significativas, la hipótesis H.1.4.0 no puede ser rechazada. Sin embargo, el periodo (la tarea) es un factor estadísticamente significativo para todas las métricas. La tarea 1 parece obtener resultados más favorables que la tarea 2, es decir, mayor puntuación de precisión, recall, accuracy y aciertos, y menor puntuación de error. Este resultado puede ser debido a una diferencia en la dificultad de las tareas. La tarea 1 puede ser más sencilla y más común, o quizás la tarea 2 es más complicada o posee un enunciado menos claro. Como el tamaño muestral es pequeño, no podemos realizar ninguna afirmación, pero este indicio se ha de tener en cuenta de cara a futuras réplicas. Si con un tamaño muestral mayor los resultados siguen mostrando diferencias debidas a las tareas, deberán ser modificadas. La reversión de los efectos del tratamiento en función de la secuencia se produce por el efecto de la tarea. En la secuencia CR-SC, Creately obtiene mejores resultados porque se aplica en la tarea 1. En la secuencia SC-CR, ocurre lo contrario, como la tarea 1 se realiza con SOCIO, el chatbot obtiene mejores resultados. Por último, cabe destacar que secuencia es estadísticamente significativa para las métricas precisión, accuracy y error. Al realizar la tarea 1 con SOCIO y la tarea 2 con Creately parecen obtenerse mejores resultados para las métricas mencionadas. El hecho de que la secuencia sea estadísticamente significativa implica que la interacción entre la tarea y el tratamiento o los efectos de *carryover* se han materializado, lo cual supone una amenaza a la validez interna.

5.2 Resultados para SOCIO

La Tabla 20 resume los resultados de los test-t realizados, uno por cada métrica, para comparar la media de las diferentes interacciones generadas durante las tareas 1 y 2 con el chatbot SOCIO. También, resume el tamaño del efecto producido por la tarea. En la columna Hipótesis, el símbolo - denota la existencia de diferencias irrelevantes.

Métrica	Hipótesis	Test-t	Tamaño del efecto
Mensajes enviados a SOCIO	H.2.1.0	-	Pequeño
Mensajes erróneos enviados a SOCIO	H.2.2.0	-	Grande
Mensajes útiles enviados a SOCIO	H.2.3.0	-	Mediano
Mensajes descriptivos enviados a SOCIO	H.2.4.0	-	Pequeño
Comandaos enviados a SOCIO	H.2.5.0	-	Pequeño
Acciones desencadenadas por SOCIO	H.2.6.0	-	Grande

Tabla 20: Resumen de los resultados experimentales obtenidos para los datos de la interacción con el chatbot SOCIO.

En los test-t para muestras independientes realizados, la tarea no causa diferencias estadísticamente significativas sobre las métricas, por lo tanto, las hipótesis mostradas en la Tabla 20 no pueden ser rechazas. Aunque las diferencias no sean estadísticamente

significativas (lo cual puede ser debido a un tamaño muestral pequeño), mirando el tamaño del efecto, parece que durante la tarea 2 se genera un mayor número de mensajes erróneos, de mensajes útiles y de acciones desencadenadas. Por tanto, parece que requiere mayor esfuerzo.

6 Conclusiones y Trabajos Futuros

En este último capítulo, se presentan las conclusiones del presente trabajo y se proponen posibles trabajos futuros.

6.1 Conclusiones

Este estudio empírico comenzó con el análisis de 10 artículos: ocho de experimentación para evaluar la usabilidad de diferentes chatbots y dos sobre el chatbot SOCIO. A partir del análisis, se diseñó un experimento para evaluar la usabilidad de SOCIO y la calidad de los diagramas de clases que genera. La evaluación se realizó por comparación con la usabilidad de la aplicación web Creately, con respecto a la eficacia, la eficiencia y la satisfacción. El experimento fue realizado por 30 sujetos experimentales con conocimientos en Ingeniería Informática. Los sujetos trabajaron en equipos de 3, lo que supuso una reducción del tamaño muestral a 10 equipos. Tras el experimento, los datos recolectaron se analizaron y los resultados fueron discutidos. Como resultado de este proceso, se pueden señalar las siguientes conclusiones:

- Son pocos los experimentos que hay en la literatura para evaluar la usabilidad de los chatbots, por lo que se necesita más esfuerzo en esta línea.
- El tamaño muestral (10 equipos), utilizado en el experimento realizado con SOCIO y Creately, no es suficiente. Este hecho ha podido causar que no haya diferencias significativas producidas por la herramienta utilizada, en ninguna de las métricas asociadas a la eficiencia, la eficacia y la calidad.
- La herramienta genera diferencias estadísticamente significativas para la variable satisfacción. El chatbot SOCIO parece satisfacer en mayor medida a los usuarios. Además, los participantes señalan que el chatbot es cómodo, sencillo y permite la creación de diagramas de manera rápida. Valoran también de manera positiva la interacción a través del lenguaje natural, la visualización de las modificaciones del diagrama de manera inmediata y su integración en las redes sociales.
- La tarea parece generar diferencias estadísticamente significativas en la satisfacción y en todas las métricas asociadas a la calidad. En concreto, la tarea 2 parece producir peores resultados. A su vez, los efectos de la secuencia o el *carryover* se materializan en algunas métricas de la calidad, lo que supone una amenaza a la validez interna. Ante esta situación, se debe realizar el experimento bajo las mismas condiciones con un tamaño muestral más grande. Si los resultados siguen reflejando los efectos de la tarea, la secuencia o el *carryover*, el diseño del experimento deberá modificarse.

6.2 Trabajos Futuros

El presente trabajo supone una contribución a la evaluación de la usabilidad de los chatbots a través de la experimentación y en particular, colabora en la evaluación del chatbot SOCIO. Sin embargo, en el experimento realizado con SOCIO y con la aplicación

web Creately, las variables eficacia, eficiencia y calidad no parecen estar afectadas por la herramienta utilizada. Este hecho hace necesario continuar con la investigación, con el objetivo de obtener resultados concluyentes. Se consideran los siguientes trabajos futuros:

- Realizar réplicas del experimento, bajo las mismas condiciones, y agregar los resultados para obtener un tamaño muestral mayor.
- Modificar el diseño del experimento, si con un tamaño muestral adecuado se verifica que la tarea y la secuencia o el *carryover* afectan en los resultados.
- Realizar réplicas del experimento con usuarios pertenecientes a diferentes áreas.

Referencias

Borenstein, M., Hedges, L.V., Higgins, J.P., & Rothstein, H.R. (2011). *Introduction to Meta-Analysis*. John Wiley & Sons.

Campbell, D.T., & Stanley, J.C (1963). *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin Company.

Chen, M.L., & Wang, H.C. (2018). How Personal Experience and Technical Knowledge Affect Using Conversational Agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (pp. 53-58).

Cheng, A., Raghavaraju, V., Kanugo, J., Handrianto, Y.P., & Shang, Y. (2018). Development and Evaluation of a Healthy Coping Voice Interface Application Using the Google Home for Elderly Patients with Type 2 Diabetes. In *Proceedings of the 15th IEEE Annual Consumer Communications & Networking Conference* (pp. 1-5).

Deci, E.L, & Ryan, R.M.(2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4): 227-268.

Giraldo, F.D., España, S., Giraldo, W.J., & Pastor, O. (2018). Evaluating the quality of a set of modelling languages used in combination: A method and a tool. *Information Systems*, 77: 48-70.

Field, A. Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. Sage.

Higgins, J.P., & Green, S. (2006). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.

ISO 9241-11. (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part II Guidance on Usability.

ISO/IEC 25010. (2010). Systems and Software Engineering. System and Software Product. Quality Requirements and Evaluation (SQuaRE). System and Software Quality Models.

Jain, M., Kota, R., Kumar, P., & Patel, S.N. (2018). Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 468-477).

Nguyen, Q.N, & Sidorova, A.C. (2018). Understanding user interactions with a chatbot: A self-determination theory approach. In *Proceeding of the 24th Americas Conference on Information Systems 2018: Digital Disruption* (pp. 1-5).

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., & Martinez, A. (2018). Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science*, pp. 1-14.

Pérez, J., Sánchez, Y., Serón, F.J., & Cerezo, E. (2017). Interacting with a Semantic Affective ECA. *In Proceedings of the International Conference on Intelligent Virtual Agents* (pp. 374-384).

Pérez-Soler, S., Guerra, E., & de Lara, J. (2018). Collaborative modelling and group decision-making using chatbots within social networks. *IEEE Software*, 38(2): 48-54.

Pérez-Soler, S., Guerra, E., de Lara, J., & Jurado, F. (2017). The Rise of the (Modeling) Bots: Towards Assisted Modeling via Social Networks. In *Proceedings of the 33nd IEEE/ACM International Conference on Automated Software Engineering* (pp. 723-728).

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).

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.

Vegas, S., Apa, C., & Juristo, N. (2016). Crossover designs in software engineering experiments: Benefits and perils. *IEEE Transactions on Software Engineering*, 42(2), 120-135.

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), p.125.

Glosario

Creately	Aplicación web que permite la elaboración de diagramas, (entre ellos, diagramas de clases) de manera colaborativa.
Diseño crossover	Diseño experimental en el que grupos de sujetos diferentes aplican los tratamientos a evaluar en órdenes diferentes.
Periodo	Momento de aplicación de un tratamiento.
Secuencia	Orden de aplicación de los tratamientos.
SOCIO	Chatbot que permite la elaboración de diagramas de clases mediante la interpretación de sentencias en lenguaje natural. Está integrado en redes sociales.
Tratamiento	Elemento a evaluar en un experimento.

Anexos

A Tablas Comparativas

En este primer anexo se presentan tablas comparativas de experimentos que evalúan la usabilidad de diferentes chatbots y de evaluaciones del chatbot SOCIO (Tablas 21-30).

ID	1
Título	Understanding User Interactions with a Chatbot: A Self-Determination Theory Approach
Autores	Quynh N. Nguyen, Anna Sidorova
Tipo de chatbot	Hipmunk, asistente virtual de planificación de viajes impulsado por inteligencia artificial, ayuda a buscar opciones de vuelo y alojamiento, brinda consejos y recomendaciones.
Objetivo de la investigación	Estudiar las diferencias en la satisfacción del sistema entre sistemas web y sistemas chatbot y qué factores determinan la satisfacción.
Hipótesis	Existe una relación entre la autonomía percibida, la competencia percibida, la carga cognitiva, la satisfacción de desempeño, la satisfacción del proceso y la satisfacción del sistema.
Variables independientes	Sistema: web o chatbot.
Variables respuesta	Satisfacción del sistema, competencia percibida, autonomía percibida, satisfacción del desempeño, carga cognitiva, satisfacción del proceso.
Métricas para cada variable	La satisfacción del rendimiento, del proceso y del sistema son medidas mediante un elemento desarrollado por el autor. La autonomía y competencia percibidas son medidas mediante escalas adaptadas de (Deci & Ryan, 2000). Para medir el esfuerzo cognitivo se adapta una escala (no mencionada) de la literatura existente.
Sujetos	Estudiantes de una universidad pública. El número de sujetos no se menciona.
Instrumentos para la recolección de datos	Experimento online, encuesta.
Tareas	Emplear el sitio web Hipmunk y el chatbot Hipmunk para completar 2 tareas: búsqueda de un billete de avión y búsqueda de habitación de un hotel. El orden de aplicación de los tratamientos y de las tareas a realizar es aleatorio.
Prueba estadística	ANOVA y modelos de ecuaciones estructurales.
Principales resultados	No se mencionan, el experimento aún no se ha llevado a cabo.

Tabla 21: Experimento con el chatbot Hipmunk.

ID	2
Título	Talk to Me: Exploring User Interactions with the Amazon Alexa
Autores	Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, Adrianna Martinez
Tipo de chatbot	Amazon Alexa, aplicación controlada por voz, desarrollada por la compañía de Amazon para sus dispositivos Echo, Echo Dot y Echo Show. Se trata de un asistente personal inteligente, empleado para reproducir música, responder preguntas generales, configurar alarmas y temporizadores o controlar dispositivos domésticos inteligentes.
Objetivo de la investigación	Comprender las interacciones de usuario con Alexa: tipos, variables que afectan y alternativas.
Hipótesis	Las interacciones con Alexa se verán afectadas por la edad, el nivel de competencia (usuario avanzado y no avanzado), el día de la semana, la ubicación del dispositivo. La cantidad de tiempo durante el cual un hogar ha tenido a Alexa podría afectar en la frecuencia de uso. Los recuerdos posteriores al uso podrían afectan en la actitud de los usuarios hacia la tecnología y su uso.
Variables respuesta	Satisfacción, tipos de interacción, variables que afectan a la interacción, alternativas a Alexa.
Métricas para cada variable	Respuestas al cuestionario demográfico y al diario online (compresión de los comandos y tareas completadas) para la satisfacción; respuestas al diario online para los tipos de interacción; respuestas al cuestionario demográfico y diario online para la variables que afectan en la interacción; respuestas del correo electrónico para las alternativas.
Sujetos	19 participantes, usuarios de Alexa de 9 hogares diferentes. Las edades se encuentran entre 4 y 55 años. 12 con un empleo profesional, 5 estudiantes y 2 o desempleados o no respondieron a la pregunta. Se asume una homogeneidad en las características socio-económicas basados en la adquisición tempana de Alexa.
Instrumentos para la recolección de datos	Cuestionario demográfico online, diario online y correo electrónico.
Tareas	Antes del estudio cuestionario demográfico online. Durante el estudio, completar un diario online al final de cada día. Dos semanas después, escribir un email listando las aplicaciones o dispositivos similares a Alexa que ha empleado el participante y su uso. Duración: 4 días (viernes, sábado, domingo y lunes).
Prueba estadística	Porcentajes.
Principales resultados	Alexa se utilizó principalmente para consultar el tiempo, escuchar música y controlar otros dispositivos. Varios participantes indicaron emplear también Apple Siri y Google Now con el mismo propósito, salvo para controlar dispositivos. Los usuarios utilizan más Alexa durante el fin de semana, y su uso tiende a disminuir con el tiempo. Los usuarios indicaron estar satisfechos con Alexa, incluso cuando no proporcionó la información buscada.

Tabla 22: Experimento con Amazon Alexa.

ID	3
Título	Development and Evaluation of a Healthy Coping Voice Interface Application Using the Google Home for Elderly Patients with Type 2 Diabetes
Autores	Amy Cheng, Vaishnavi Raghavaraju, Jayanth Kanugo, Yohanes P. Handrianto, Yi Shang
Tipo de chatbot	Agente conversacional que ayuda a pacientes ancianos con el autocontrol de la diabetes mellitus tipo 2. Pertenece a la aplicación de uso doméstico Healthy Coping in Diabetes, emplea la interfaz de voz de Google Home para alojar al agente y una interfaz web para la visualización de datos.
Objetivo de la investigación	Comparar la aplicación Healthy Coping con las aplicaciones móviles con su misma funcionalidad, a fin de determinar el cumplimiento de los requisitos oficiales de autogestión de la Asociación Americana de Educadores de la Diabetes (AAED). Evaluar la usabilidad práctica de la aplicación mediante el cálculo de las métricas estándar, con el fin de mejorar sus capacidades.
Hipótesis	El empleo de una interfaz de voz en la aplicación Healthy Coping, en lugar de una pantalla de móvil, permitirá mantener una funcionalidad similar a la funcionalidad de otras aplicaciones móviles de autocontrol, pero aumentando la usabilidad y, por lo tanto, los beneficios que puede proporcionar esta aplicación.
Variables independientes	Aplicación: Healthy Coping in Diabetes u otras aplicaciones móviles con la misma funcionalidad.
Variables respuesta	Eficacia, satisfacción.
Métricas para cada variable	Satisfacción: respuestas encuesta cualitativa de satisfacción.
Sujetos	10 participantes ancianos (evaluación inicial de la usabilidad) y expertos en el cuidado de personas mayores y usuarios potenciales (evaluación posterior de la usabilidad)
Instrumentos para la recolección de datos	Encuesta de satisfacción, comentarios.
Tareas	Las personas que llevan a cabo el estudio: revisión de aplicaciones de autocontrol de la diabetes, comparación de sus características con las de Healthy Coping, para evaluar que cubre los requisitos AADE del autocontrol. Los participantes de evaluación inicial de usabilidad: encuesta de detección de depresión de Healthy Coping y encuesta de satisfacción (enfocada a la interfaz de voz). Los participantes de la segunda evaluación de usabilidad: visualizar una demostración de las características y funcionalidades de la aplicación y realizar comentarios sobre la misma.
Prueba estadística	Porcentajes encuesta satisfacción.
Principales resultados	La mayoría de las aplicaciones móviles para el autocontrol de la diabetes no cumplen con todos los requisitos de AAED. El 80% de los participantes prefiere utilizar Healthy Coping a emplear un teléfono móvil. Los expertos llegaron a la conclusión del que Healthy Coping tiene potencial para ayudar. Se han determinado deficiencias y posibles soluciones que mejorarán la usabilidad.

Tabla 23: Experimento con la aplicación Healthy Coping in Diabetes.

ID	4
Título	How Personal Experience and Technical Knowledge Affect Using Conversational Agents
Autores	Mei-Ling Chen, Hao-ChuanWang
Tipo de chatbot	Apple Siri, asistente personal controlado por voz disponible para usuarios de Apple. Siri es un agente conversacional que hace llamadas, manda mensajes, realiza tareas cotidianas como poner alarmas, controla dispositivos dómoticos, etc.
Objetivo de la investigación	Examinar cómo la comprensión de los usuarios afecta en las percepciones y experiencias de uso de agentes conversacionales.
Hipótesis	La experiencia de uso previa y el conocimiento técnico afectan en la usabilidad percibida y el modelo mental de cada usuario.
Variables independientes	Experiencia personal en el uso de agentes conversacionales y conocimiento técnico sobre el modelo del sistema de los agentes conversacionales.
Variables respuesta	Satisfacción.
Métricas para cada variable	Satisfacción: puntuación del cuestionario SUMI.
Sujetos	41 participantes, 24 hombres y 17 mujeres de entre 19 y 31 años, siendo la media de edad 23,34 años y la desviación 2,48. 19 usuarios con experiencia en agentes conversacionales, de los cuales 9 poseen conocimientos técnicos y 10 no. 22 usuarios sin experiencia, de los cuales 9 poseen conocimientos técnicos y 13 no.
Instrumentos para la recolección de datos	Grabación del proceso de uso de Siri, entrevista, cuestionario SUMI.
Tareas	Tareas de planificación de viajes (por ejemplo, “encontrar un hotel económico en Osaka”), usando solo Apple Siri en un teléfono móvil durante 30 minutos. Una vez finalizadas, *think-aloud* retrospectivo (viendo la grabación de la realización de las tareas), entrevista y cuestionario SUMI.
Prueba estadística	ANOVA de dos factores sobre la usabilidad percibida y análisis post-hoc utilizando el test-t.
Principales resultados	Efecto de interacción entre la experiencia previa y el conocimiento técnico sobre la usabilidad global. La experiencia previa y el conocimiento técnico sobre la usabilidad global. En general, la experiencia de uso anterior y el conocimiento técnico afectan de manera diferente en cómo se sienten las personas respecto a los agentes conversacionales en contextos de uso real.

Tabla 24: Experimento con Apple Siri.

ID	5
Título	Convey: Exploring the Use of a Context View for Chatbot
Autores	Mohit Jain, Ramachandra Kota, Pratyush Kumar, Shwetak Patel
Tipo de chatbot	Chatbot para comprar calzado, diseñado para entender y filtrar zapatos basándose en el precio, el color, el material, el estilo y la marca, ayuda a los usuarios en el proceso de decisión.
Objetivo de la investigación	Evaluar la usabilidad de Convey, ventana que muestra el contexto de la conversación y le proporciona al usuario una interacción sencilla con los valores del mismo (confirmación, modificación, eliminación).
Hipótesis	Existe una discrepancia entre el contexto real de chatbot y la percepción que el usuario tiene de él, Convey mejorará la interacción del usuario con el chatbot.
Variables independientes	Interfaz: por defecto o con Convey.
Variables respuesta	Eficiencia y satisfacción.
Métricas para cada variable	Eficiencia: número de zapatos visualizados e imágenes ampliadas antes de añadirlos al carrito, tiempo en completar las tareas, número de palabras introducidas, longitud de los mensajes, tiempo invertido en la interacción con los elementos de Convey, demanda física, demanda mental, esfuerzo. Satisfacción: facilidad de uso, satisfacción percibida durante las tareas, entretenimiento y frustración.
Sujetos	16 participantes, 11 hombres y 5 mujeres, con una media de edad 32,5 años y una desviación típica de 7,4 años. 14 poseen una formación en ingeniería y 2 en finanza y ciencias sociales. Ninguno habla inglés nativo, pero sí de manera fluida. Todos entienden los chatbots a nivel conceptual, 5 han interaccionado con chatbots en la plataforma de Facebook Messenger. 2 participantes declararon que la comprensión adecuada del contexto fue una de las principales dificultades que enfrentaron al interactuar con chatbots en el pasado.
Instrumentos para la recolección de datos	Cuestionario, formulario online, servidor.
Tareas	Con cada interfaz llevar a cabo una de estas dos tareas: seleccionar un par de zapatos de fiesta para sí mismos con la primera interfaz asignada; seleccionar un par de zapatillas de deporte para el género contrario con la segunda interfaz asignada. El orden de las interfaces es aleatorio. Tutorial de cada interfaz antes de su uso. Al finalizar cada tarea, cuestionario para medir la experiencia de usuario (escala Likert de 5 puntos) y opinión subjetiva sobre las interfaces en un formulario online. Al finalizar ambas, comparación de interfaces y preferencia. Duración: 45 minutos de media.
Prueba estadística	Test-t pareado para el tiempo empleado en completar la tarea, total de palabras utilizadas por el usuario, éxito percibido en las tareas y uso potencial en el futuro. Media y desviación típica de los resultados de las métricas.
Principales resultados	Los participantes prefieren Convey, les resulta más fácil de usar, requiere menos esfuerzo mental, es más rápido e intuitivo.

Tabla 25: Experimento con Convey.

ID	6
Título	Friendship with a Robot: Children's Perception of Similarity between a Robot's Physical and Virtual Embodiment that Supports Diabetes Self-Management
Autores	Claudia Sinoo, Sylvia van der Pal, Oliver A. Blason Henkemans, Anouk Keizer, Bert P.B. Bierman, Rosemarijn Looije, Mark A. Neerincx
Tipo de chatbot	NAO, agente conversacional con una encarnación física (robot) y virtual (avatar), apoya el autocontrol de la diabetes de los niños. El robot, su avatar y la aplicación MyPAL (de la que es parte el avatar) constituyen el sistema PAL (*Personal Assistant for a Healthy Lifestyle*).
Objetivo de la investigación	Evaluar el efecto de la similitud percibida entre el robot y el avatar, en la amistad de los niños hacia el avatar. Evaluar el efecto de esta amistad en la usabilidad de la aplicación de autogestión que contiene el avatar y en la motivación de los niños para jugar con ella.
Hipótesis	Al percibir un alto grado de similitud entre el robot y el avatar, también se experimentará una amistad más fuerte con el avatar y una mayor motivación para jugar con MyPAL, así como una mayor usabilidad de la aplicación.
Variables independientes	Encarnación: robot o avatar.
Variables respuesta	Amistad, similitud, motivación.
Métricas para cada variable	Realización de una calificación cuantitativa de las variables en las escalas de Likert con emoticonos (cuestionario SUS adaptado), y una calificación de la similitud de manera cualitativa mediante preguntas abiertas.
Sujetos	21 niños de un campamento para niños con diabetes, 13 niños y 8 niñas de entre 8 y 11 años, siendo la media de edad 9,1 y la desviación típica 1,1. Poseen diabetes desde que tenían entre 1 y 7 años (media 3,5 y la desviación 1,7).
Instrumentos para la recolección de datos	Cuestionarios.
Tareas	Interacciones con el robot y el avatar durante 4 días: jugar con MyPAL, actividades dirigidas por los robots, 4 juegos de pregunta y calificaciones (2 con el robot y después, los mismos, con el avatar). 2 cuestionarios (primer y último día).
Prueba estadística	Alfa de Cronbach, test-t para la diferencia entre la amistad con el robot y el avatar y para los cambios en la amistad, motivación y usabilidad entre el primer día y el último. Con los datos del último cuestionario: ANOVA para el efecto de la similitud percibida en la amistad con el avatar. Post-hoc de Gabriel para las diferencias en la puntuación de amistad, entre las categorías de respuestas de similitud, w^2 para la magnitud del efecto de la similitud en la amistad, análisis de regresión logística para el efecto de la amistad en la motivación y de regresión lineal para el efecto sobre la usabilidad.
Principales resultados	Preferencia por el robot, hacia el que sienten una mayor amistad. Aumento en la amistad con el avatar cuando la similitud percibida aumenta. Efecto positivo de la amistad con el avatar en la motivación para jugar con MyPAL y correlación positiva con la usabilidad. Los niños declararon que el robot era más (inter) activo, más presente y más capaz de hacer cosas diferentes, como bailar.

Tabla 26: Experimento con el robot NAO y su avatar.

ID	7
Título	Interacting with a Semantic Affective ECA
Autores	Joaquín Pérez, Yanet Sánchez, Francisco J. Serón, Eva Cerezo
Tipo de chatbot	E-VOX es un agente conversacional personificado, (en inglés, *Embodied Conversational Agent*, ECA), semántico afectivo mejorado. Trabaja como asistente para proporcionar información útil de Wikipedia, respaldando la sensación real en la Interacción Persona-Ordenador.
Objetivo de la investigación	Estudiar la usabilidad general del sistema E-VOX, la idoneidad y el realismo del agente y el impacto de su comportamiento en la interacción y percepción de los usuarios. Se hace especial énfasis en tratar de evaluar el impacto de los componentes emocionales.
Hipótesis	La personalidad del agente tiene un impacto en la interacción usuario-agente. El comportamiento emocional del agente es percibido por el usuario. El usuario percibe el comportamiento emocional del agente como realista.
Variables independientes	Personalidad del agente: positiva, negativa, sin personalidad.
Variables respuesta	Realismo de apariencia del agente, realismo del comportamiento del agente, grado en el que el comportamiento del agente es apropiado y satisfacción.
Métricas para cada variable	Satisfacción: puntuaciones cuestionarios de satisfacción. Las demás: puntuaciones cuestionarios rellenados al finalizar cada tarea.
Sujetos	22 participantes, 11 hombres y 11 mujeres de entre 17 y 62 años.
Instrumentos para la recolección de datos	Cuestionarios.
Tareas	Cuestionario de familiaridad (edad, sexo, educación, experiencia con ECA), y prueba de personalidad emocional del usuario antes de las tareas. Búsqueda de información sobre el cáncer infantil, búsqueda de información libre. Las tareas se repiten 3 veces, usando un agente con una personalidad positiva negativa y sin personalidad. El orden de personalidades es aleatorio y el usuario no está informado de la personalidad del agente. Cuestionario tras la realización de cada tarea y cuestionario de satisfacción tras finalizar todas las tareas.
Prueba estadística	Cálculo de la media y la desviación típica de las puntuaciones obtenidas en cada pregunta del cuestionario post-tarea y del cuestionario final.
Principales resultados	El sistema se percibe rápido, fácil de usar, útil y entretenido. Los usuarios percibieron la apariencia física del agente casi igual de realista para cada tarea. La personalidad del agente tiene un impacto en la interacción usuario-agente. El comportamiento emocional del agente no es percibido en gran medida por el usuario, dado las limitaciones de expresividad emocional, pero es considerado realista.

Tabla 27: Experimento con E-VOX.

ID	8
Título	A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories
Autores	Myrthe L. Tielman, Mark A. Neerincx, Rafael Bidarra, Ben Kybartas, Willem Paul Brinkman
Tipo de chatbot	Agente virtual perteneciente al sistema 3MR_2, que ofrece terapia a pacientes que presentan un trastorno de estrés postraumático (TEPT). El sistema contiene un diario digital y un 3D WorldBuilder en el que se pueden recrear recuerdos.
Objetivo de la investigación	Evaluación de la usabilidad en dos etapas: prueba de usabilidad inicial del agente virtual y el diario digital, y estudio para evaluar la usabilidad del sistema y sus componentes, la contribución de los elementos del sistema a la terapia, y la utilidad y adecuación de las preguntas realizadas para recordar experiencias traumáticas.
Hipótesis	En el segundo estudio, las preguntas generadas por el módulo de preguntas ayudan a los pacientes a recordar. El sistema es útil.
Variables respuesta	En el primer estudio, usabilidad. En el segundo, utilidad de las preguntas para recordar el trauma, utilidad y comprensión de las funciones del programa, y satisfacción.
Métricas para cada variable	En el segundo estudio, la satisfacción se mide con la puntuación del cuestionario general (SUS), las demás variables con las preguntas con opciones.
Sujetos	En el primer estudio, 3 participantes sanos, investigadores o estudiantes del departamento de informática, 2 poseen conocimientos en psicología y 1 en informática. En el segundo estudio, 4 participantes que habían seguido la terapia para el trastorno de estrés postraumático, 2 son hombres veteranos de guerra y 2 mujeres que habían sufrido abuso sexual en la infancia.
Instrumentos para la recolección de datos	Think-aloud en el primer estudio, cuestionarios en el segundo.
Tareas	En el primer estudio, realización de la primera sesión de terapia (recuerdo positivo) pensando en voz alta. La sesión no incluye preguntas del módulo de preguntas o del entorno 3D. En el segundo estudio, primero, información general sobre el sistema 3MR_2, cómo sería una terapia completa y qué se esperaba de ellos durante el experimento. Después, realización de la primera sesión de terapia y parte de dos sesiones en las que se describe una experiencia traumática, respondiendo también preguntas con opciones. Al finalizar, cuestionario general.
Prueba estadística	En el segundo estudio, análisis multinivel.
Principales resultados	En el primer estudio, se detectan pequeños problemas de usabilidad que son resueltos. Se realizaron dos videos instructivos, uno describe el sistema general y otro el entorno 3D. En el segundo estudio, los participantes consideran las preguntas útiles para ayudarles a recordar, y al sistema útil y usable. Se observa que hay personas que pueden preferir un sistema con menos orientación.

Tabla 28: Experimento con el agente del sistema 3MR_2.

ID	9
Título	The Rise of the (Modelling) Bots: Towards Assisted Modelling via Social Networks
Autores	Sara Pérez-Soler, Esther Guerra, Juan de Lara, Francisco Jurado
Tipo de chatbot	SOCIO, a través de Twitter y Telegram, interpreta los mensajes en lenguaje natural (en inglés) de los usuarios para construir diagramas de clases de manera colaborativa.
Objetivo de la investigación	Evaluar la idoneidad del empleo de las redes sociales y el lenguaje natural para la construcción de diagramas de clase.
Hipótesis	El uso de las redes sociales y el lenguaje natural son útiles para elaborar diagramas de clases.
Variables respuesta	Satisfacción, idoneidad lenguaje natural, idoneidad redes sociales.
Métricas para cada variable	Cuestionario SUS para la satisfacción; puntuaciones del cuestionario, número de mensajes descriptivos y número de comandos, para la idoneidad del lenguaje natural; puntuaciones del cuestionario, número de mensajes de discusión y número de mensajes dirigidos al chatbot para la idoneidad de las redes sociales.
Sujetos	10 participantes con conocimientos en informática (estudiantes de posgrado o de último año). El promedio declarado de experiencia en modelado fue de 62,5%, y el nivel de inglés fue de 72,5%.
Instrumentos para la recolección de datos	Cuestionario y chatbot.
Tareas	Los participantes se dividen en 4 grupos de Telegram (2 grupos de 2 personas y 2 grupos de 3 personas) para crear un diagrama de clases de comercio electrónico en 15 minutos, sin ninguna otra restricción. Al finalizar, completar un cuestionario (familiaridad, SUS, idoneidad lenguaje natural y precisión en la interpretación, funcionalidad del set de comandos, herramienta en redes sociales o no).
Prueba estadística	Porcentajes, coeficiente de correlación (entre bajas puntuaciones en la usabilidad, el nivel de inglés y la experiencia en modelado).
Principales resultados	Resultados positivos en cuanto a la satisfacción, el empleo del lenguaje natural como método de interacción y la idea de colaborar a través de las redes. En cuanto al set de comandos de SOCIO y la precisión de la interpretación del lenguaje natural, los resultados fueron buenos, pero no tanto, lo que sugiere la necesidad de mejorar en esta línea.

Tabla 29: Evaluación del chatbot SOCIO.

ID	10
Título	Collaborative Modelling and Group Decision-Making Using Chatbots within Social Networks
Autores	Sara Pérez-Soler, Esther Guerra, Juan de Lara
Tipo de chatbot	SOCIO, a través de Twitter y Telegram, interpreta los mensajes en lenguaje natural (en inglés) de los usuarios para construir diagramas de clases de manera colaborativa.
Objetivo de la investigación	Evaluación del mecanismo de consenso incorporado al chatbot.
Hipótesis	El mecanismo de consenso facilitará la toma de decisiones en grupos grandes y heterogéneos.
Variables independientes	Mecanismo de consenso: ausencia o presencia.
Variables respuesta	Utilidad del mecanismo de consenso.
Métricas para cada variable	Respuestas al cuestionario para la utilidad del mecanismo.
Sujetos	8 participantes de máster y doctorado, 6 son ingenieros informáticos, 1 ingeniero en telecomunicaciones y 1 físico.
Instrumentos para la recolección de datos	Cuestionario.
Tareas	Tras un tutorial de 10 minutos, elegir la mejor de tres opciones para dos proyectos, primero con el mecanismo de consenso ausente y después presente. Al final, cuestionario sobre el mecanismo.
Prueba estadística	Media.
Principales resultados	El mecanismo de consenso fue considera útil para grupos numerosos y con una salida que refleja la opinión de la mayoría.

Tabla 30: Evaluación del mecanismo de consenso del chatbot SOCIO.

B Documentos del Experimento

Este anexo presenta los documentos utilizados durante las sesiones en las que se desarrolló el experimento. Al inicio de la sesión, los concursantes rellenaban el informe de consentimiento y el cuestionario de familiaridad. Durante las dos tareas del experimento, se les proporcionaba por escrito el enunciado de las mismas. Al finalizar cada tarea, rellenaban el cuestionario de satisfacción sobre la herramienta que habían utilizado para llevarla a cabo, el chatbot SOCIO o la aplicación web Creately.

B.1 Informe de Consentimiento y Cuestionario de Familiaridad

La Figura 17 muestra el informe de consentimiento firmado por los participantes del experimento al inicio de cada sesión, y la Figura 18, el cuestionario de familiaridad rellenado a continuación.

Informe de Consentimiento

Vas a participar en un estudio empírico llevado a cabo por Andrea Nevado y Ranci Ren para evaluar la usabilidad del chatbot SOCIO. Realizarás dos tareas en las cuales habrá que diseñar por equipos un diagrama de clases. Estas tareas no tendrán ninguna repercusión en la calificación de las asignaturas que estés cursando.

Tu participación en el experimento es completamente voluntaria. Gracias por tu colaboración.

Equipo e información de contacto

Andrea Nevado (andrea.nevado@estudiante.uam.es)
Ranci Ren (ranci.ren@gmail.com)
Silvia Teresita Acuña (silvia.acunna@uam.es)

Acuerdo

Firma _____

Figura 17: Informe de consentimiento.

GRUPO ___

Cuestiones generales: Para cada una de las siguientes cuestiones rellena o marca la casilla correspondiente.

Edad. . . . . ☐

Sexo. . . . . Hombre ☐ Mujer ☐

¿Eres estudiante o graduado en informática? . . . . . Sí ☐ No ☐

¿Has utilizado alguna vez Telegram?. . . . . Sí ☐ No ☐

¿Has utilizado alguna vez un chatbot?. . . . . Sí ☐ No ☐

¿Qué redes sociales sueles utilizar?. . . WhatsApp ☐ Telegram ☐ Twitter ☐ Facebook ☐ Instagram ☐

Puntúa tu grado de uso de las redes sociales (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Puntúa tu grado de uso de Telegram (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Puntúa tu nivel de inglés (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu grado de conocimiento sobre diagramas de clases (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu nivel de conocimiento sobre chatbots (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu grado de uso de chatbots (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Versión de Telegram empleada durante la sesión: . . . . . SmartPhone o Tablet ☐ Web ☐ Escritorio ☐

Figura 18: Cuestionario de familiaridad.

B.2 Enunciados de las Tareas del Experimento y Soluciones

La Figuras 19 y 20 muestran el enunciado asociado al diagrama de clases que los equipos tienen que elaborar en las tareas 1 y 2 respectivamente.

TAREA 1

A shop requests an application to manage their products and their clients. They have three types of products: clothes, shoes and bags. All products have an identifier, a name, a color, a description, a price and a category. In some seasons, products may have a discount. The clothes and shoes have a size, and the shoes can be of different heights. The shop wants to visualize all this information about their products, and also, a photo and the number of units.

The shop has the name, address and telephone number of its clients. Each client has an identifier. Clients can place orders. The shop wants to be able to register the orders of each client in the application, in order to see the date on which the order will be made, its identifier and the products it contains.

Figura 19: Enunciado de la tarea 1 del experimento.

TAREA 2

A school, whose name and address are known, requests an application to organize its teachers, students and subjects. The school teaches different subjects depending on the academic year. Each subject has several lessons that can be managed from the application. Exams are performed to evaluate each subject. The school wants to be able to specify the questions, the date and the weight of the exams in the subject through the application. Several classes are taught per subject, in a specific classroom, at a specific day and time. Each class has several students and is given by a single teacher. The school has the full name, address, telephone number and date of its teachers and students. In addition, every person belonging to the school has an identifier.

Figura 20: Enunciado de la tarea 2 del experimento.

Las Figura 21 muestra la solución de la primera tarea y la Figura 22 la solución de la segunda. Estos diagramas se consideran la solución ideal y se toman como referencia para evaluar la calidad de los diagramas elaborados por los participantes.

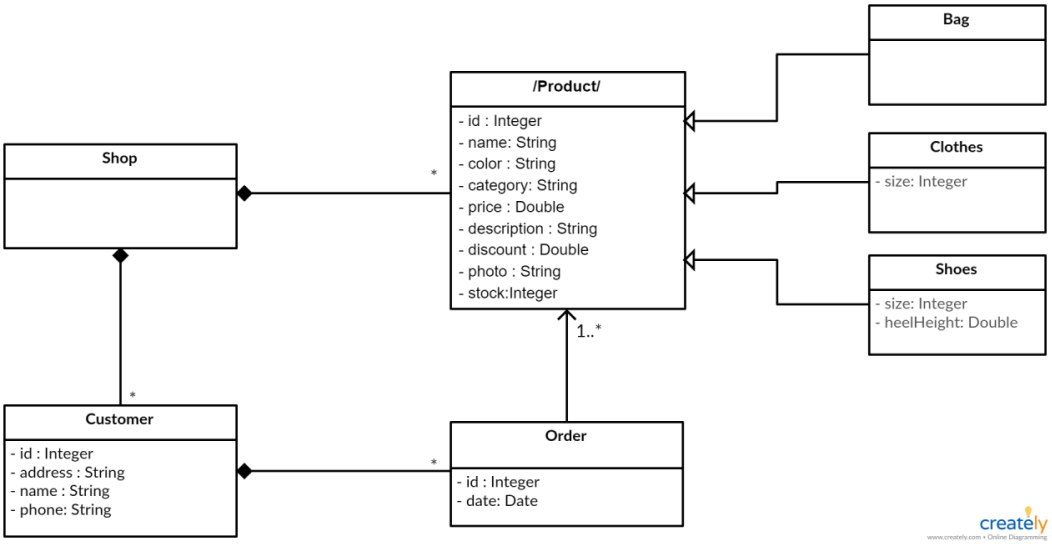


Figura 21: Solución de la tarea 1 del experimento.

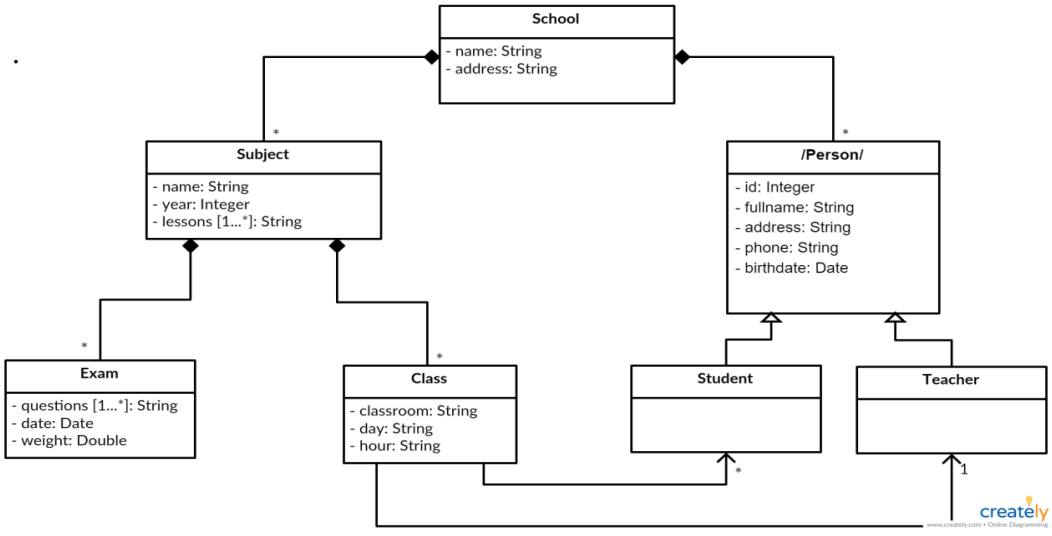
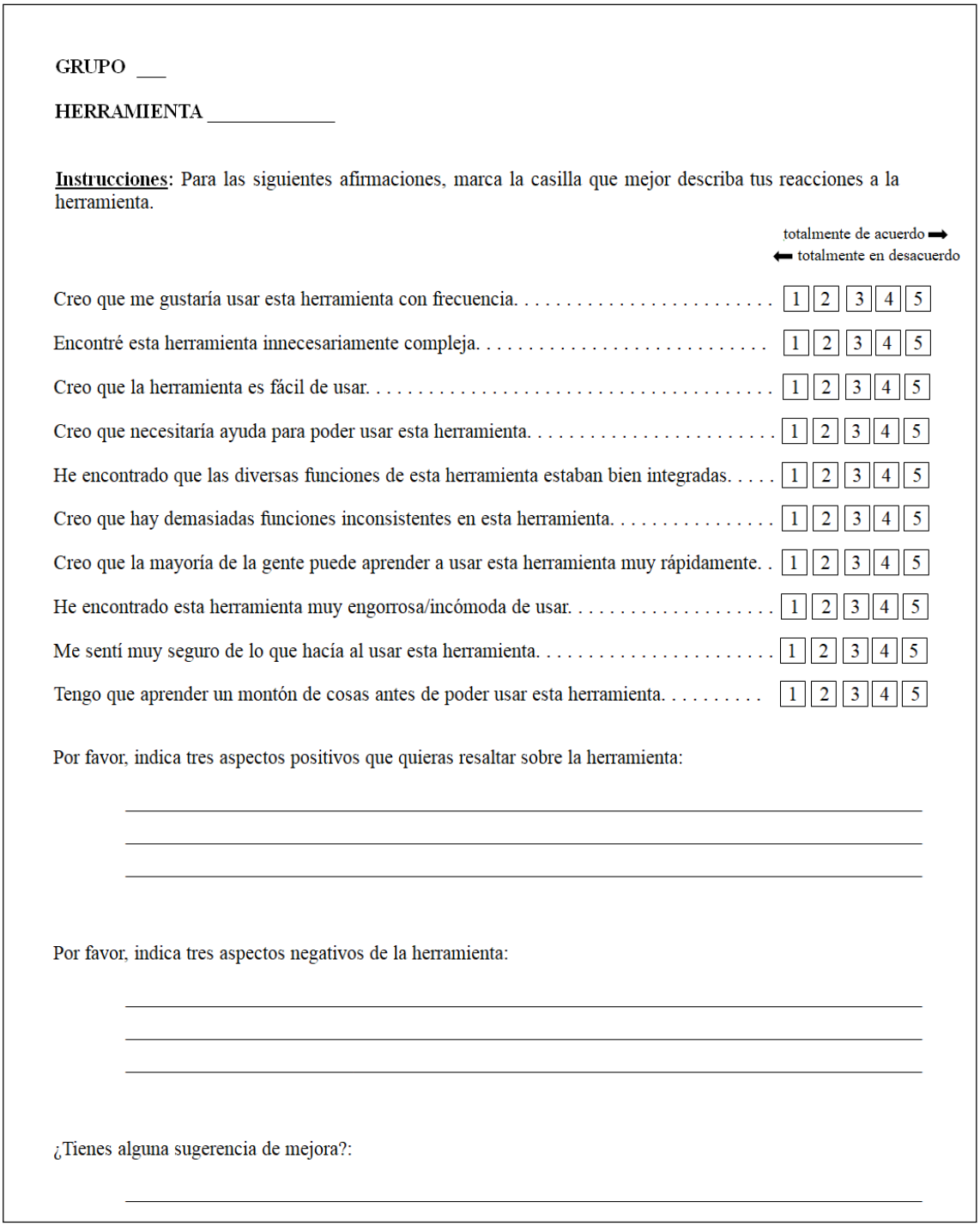


Figura 22: Solución de la tarea 2 del experimento.

B.3 Cuestionario de Satisfacción

La Figura 23 muestra el cuestionario de satisfacción rellenado por los participantes al final de cada tarea del experimento. En él se les pregunta por la herramienta que acaban de aplicar. El cuestionario de satisfacción entregado tras la última tarea incorpora una pregunta más, en la cual deben indicar si prefieren el chatbot SOCIO o la aplicación web Creately.



GRUPO ___

HERRAMIENTA _____

Instrucciones: Para las siguientes afirmaciones, marca la casilla que mejor describa tus reacciones a la herramienta.

totalmente de acuerdo ➡
⬅ totalmente en desacuerdo

Creo que me gustaría usar esta herramienta con frecuencia.	1	2	3	4	5
Encontré esta herramienta innecesariamente compleja.	1	2	3	4	5
Creo que la herramienta es fácil de usar.	1	2	3	4	5
Creo que necesitaría ayuda para poder usar esta herramienta.	1	2	3	4	5
He encontrado que las diversas funciones de esta herramienta estaban bien integradas.	1	2	3	4	5
Creo que hay demasiadas funciones inconsistentes en esta herramienta.	1	2	3	4	5
Creo que la mayoría de la gente puede aprender a usar esta herramienta muy rápidamente. .	1	2	3	4	5
He encontrado esta herramienta muy engorrosa/incómoda de usar.	1	2	3	4	5
Me sentí muy seguro de lo que hacía al usar esta herramienta.	1	2	3	4	5
Tengo que aprender un montón de cosas antes de poder usar esta herramienta.	1	2	3	4	5

Por favor, indica tres aspectos positivos que quieras resaltar sobre la herramienta:

_____
_____
_____

Por favor, indica tres aspectos negativos de la herramienta:

_____
_____
_____

¿Tienes alguna sugerencia de mejora?:

_____

Figura 23: Informe de satisfacción.

C Herramientas del Experimento

Este anexo detalla el funcionamiento de las herramientas utilizadas en el experimento, el chatbot SOCIO y la aplicación web Creately.

C.1 SOCIO

SOCIO es un chatbot que ayuda en la construcción de diagramas de clases mediante la interpretación de oraciones en lenguaje natural, en inglés. Está integrado en las redes sociales Twitter y Telegram, por lo que es una herramienta colaborativa (Figura 24).

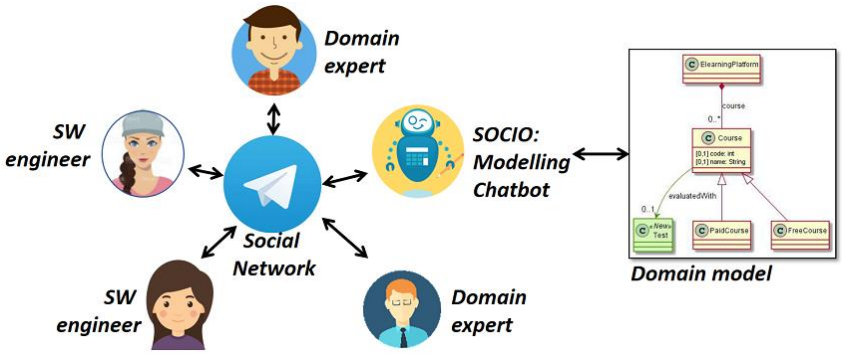


Figura 24: Interacción con el chatbot SOCIO.

En este experimento se ha utilizado el chatbot en la aplicación de Telegram. En Telegram, la interacción con SOCIO se puede realizar a través de un chat o a través de un grupo del que SOCIO sea miembro, su alias es @modellingBot. En este experimento se ha optado por emplear un grupo dado que los participantes realizan las tareas por equipo. La comunicación con SOCIO se establece a través de comandos, Tabla 31.

COMANDO	DESCRIPCIÓN
/start	Muestra todos los comandos
/newproject	Crea un nuevo proyecto
/delproject	Elimina un proyecto
/projects	Lista todos los proyectos existentes
/projectmanager	Permite gestionar los usuarios y la visibilidad del proyecto
/setproject	Permite seleccionar el proyecto en el que se desea trabajar
/help	Enlace a la página de ayuda de SOCIO: https://saraperezsoler.github.io/ModellingBot/
/talk	Permite enviar mensajes en lenguaje natural a SOCIO para construir el diagrama. Los mensajes pueden ser comandos (oraciones imperativas) o mensajes descriptivos
/undo	Deshace la última acción
/redo	Rehace la última acción deshecha
/show	Muestra el estado del diagrama en ese momento
/validate	Valida el diagrama
/get	Envía un fichero con el diagrama en formato ecore
/history	Muestra estadísticas, proporciona el historial de mensajes
/branch	Crea una nueva rama en el proyecto

Tabla 31: Comandos de SOCIO.

El comando */talk* permite enviar mensajes a SOCIO en lenguaje natural para construir el diagramas de clases. SOCIO recibe estos mensajes, los interpreta, modifica el diagrama y manda una imagen mostrando los cambios realizados. Los mensajes pueden ser de dos tipos: mensajes descriptivos y comandos. Los mensajes descriptivos son oraciones del tipo */talk Bathroom, bedroom, kitchen and livingroom are rooms*, Figura 25. Los comandos son oraciones imperativas, precedidas de verbos como *add*, *create*, *make*, *remove*, *erase* o *delete*, por ejemplo, */talk Add width, length and height in room*, Figura 26.

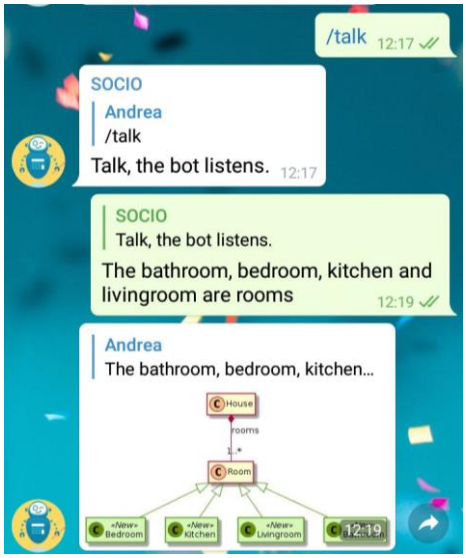


Figura 25: Mensaje descriptivo enviado al chatbot SOCIO empleando el comando */talk*.

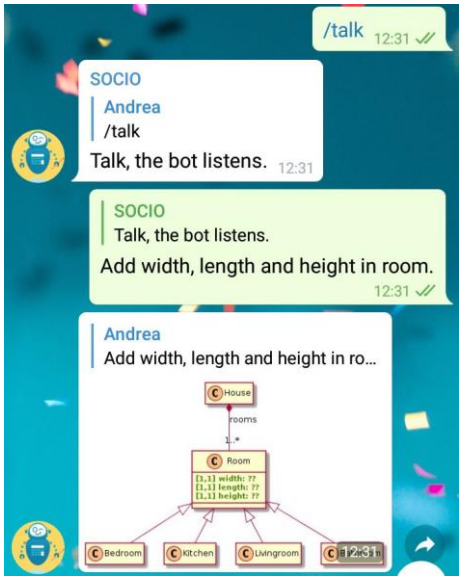


Figura 26: Comando enviado al chatbot SOCIO empleando */talk.*

C.2 Creately

Creately es una aplicación web que permite la creación de diversos tipos de diagramas de manera colaborativa, pues se puede invitar a otras personas a colaborar en el diagrama vía correo electrónico. Esta herramienta está disponible en https://creately.com/app/.

Creately permite crear diagramas de clases. La aplicación tiene una apariencia como la mostrada en la Figura 27. Presenta un menú en el lateral izquierdo con los elementos necesarios para generar el diagrama, basta con arrastrarlos al lienzo para poder utilizarlos

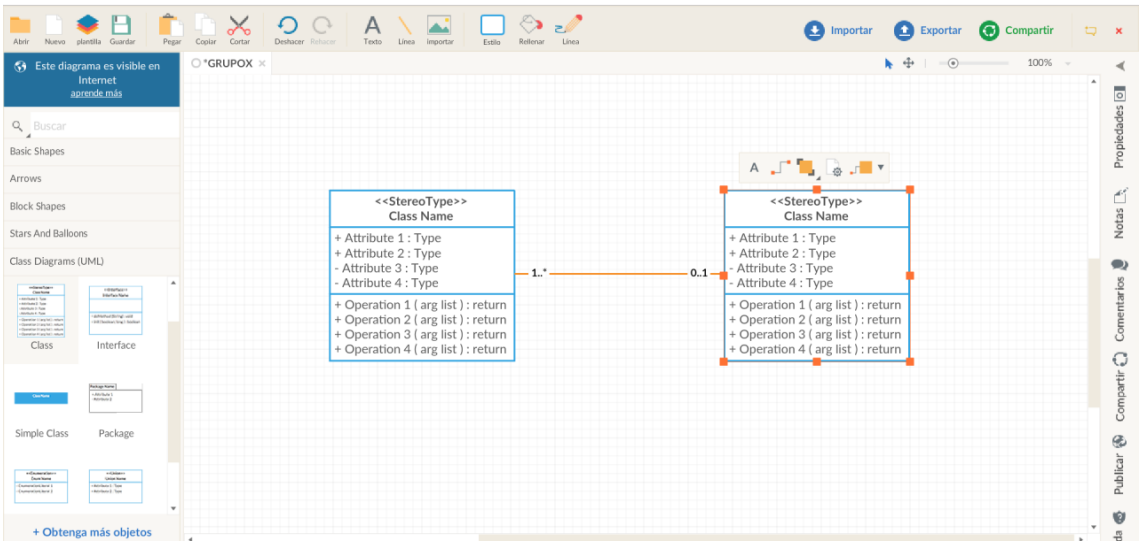


Figura 27: Apariencia de la aplicación web Creately.

Al seleccionar una clase aparece un menú que permite establecer relaciones entre las clases y editarlas, Figura 28. Al seleccionar una relación, se muestra un menú a través del cual se puede determinar el tipo de relación, Figura 29.

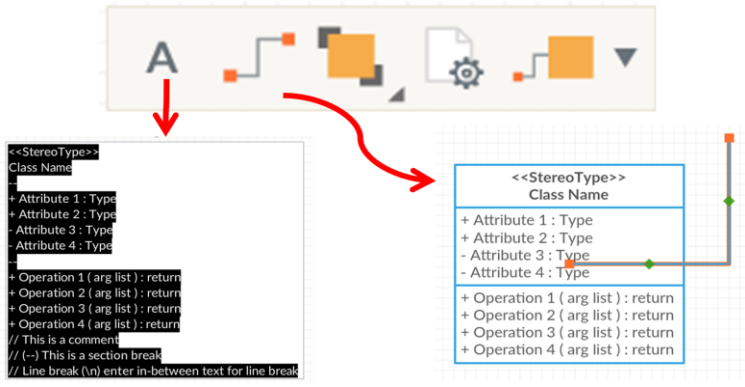


Figura 28: Menú de una clase en la aplicación Creately.

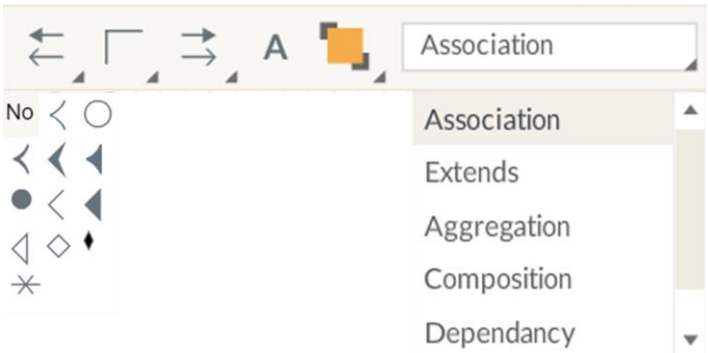


Figura 29: Menú de una relación en la aplicación Creately.

D Evaluación de la Calidad

En este anexo se especifica cómo se ha evaluado la calidad de los diagramas de clases elaborados por los equipos durante las tareas del experimento.

La calidad ha sido evaluada mediante su comparación con los diagramas considerados como la solución ideal, Figuras 19 y 20. Para ello, la evaluación de un diagrama se calcula mediante la matriz de confusión mostrada en la Tabla 32.

		Predicción	
		+	-
Actual	+	TP	FN
	-	FP	TN

Tabla 32: Matriz de confusión.

En la matriz, *Actual* hace referencia al diagrama ideal y *Predicción* al diagrama elaborado por un equipo durante la realización de una tarea. En cuanto a los elementos:

- TP (*true positive*) es el número de elementos que se encuentran en el diagrama ideal y en el diagrama elaborado por un equipo.
- FP (*false positive*) hace referencia al número de elementos que se encuentran en el diagrama elaborado por un equipo, pero no en el diagrama ideal.
- FN (*false negative*) es el número de elementos que se encuentran en el diagrama ideal, pero no se encuentran en el diagrama realizado por un equipo.
- TN (*true negative*), en la comparación de diagramas no hay *true negative*, el valor es siempre cero.

Los elementos del diagrama de clases constan de varias partes, por ejemplo, las relaciones tienen tipo y cardinalidad, las clases tienen tipo y nombre. Por esta razón, se ha empleado el siguiente sistema mostrado en la Tabla 33, para contar los elementos:

- Si una clase existe se cuenta como 0,65. Si el nombre es correcto (está en singular), se suma 0,1. Si su tipo es el adecuado (abstracta o no), se añade 0,25. Si todo es correcto, el elemento valdrá 1.
- Si un atributo existe, se cuenta como 0,75. Si el tipo es el correcto, se suma 0,25. De esta forma, si todo es correcto, el elemento valdrá 1.
- Si existe una relación, se cuenta como 0,5. Si el tipo es el adecuado, se añade 0,25. Si la cardinalidad es la correcta, se suma 0,25. Así, si todo es correcto, el elemento valdrá 1.

Elemento	**Clase**			**Atributo**		**Relación**		
Partes	Existencia	Nombre	Tipo	Existencia	Tipo	Existencia	Tipo	Cardinalidad
Puntuación	0,65	0,1	0,25	0,75	0,25	0,5	0,25	0,25

Tabla 33: Sistema de puntuación de los elementos de un diagrama, empleado para determinar los elementos de la matriz de confusión.

Cabe destacar otras consideraciones que se han tenido en cuenta a la hora de asignar la puntuación:

- Si la parte de un elemento es incorrecta (por ejemplo, si una clase no es abstracta y se ha considerado abstracta), la puntuación asociada a esa parte se sumará como FP.
- Si un elemento carece de alguna de sus partes (por ejemplo, no se ha indicado la cardinalidad en una relación), la puntuación asociada a esa parte se sumará como FN.

El número de TP, FP, FN obtenido se emplea para calcular las métricas de la calidad:

$$Precisión = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad Accuracy = \frac{TN+TP}{TP+TN+FP\ FN}$$

$$Error = \frac{FP+FN}{TP+TN+FP+FN} \qquad Aciertos = \frac{TP}{Nº\ elementos\ diagrama\ ideal}$$

En cuanto a estas métricas, la métrica precisión indica el porcentaje de acierto de un equipo, en función de todos los elementos que han considerado que formaban parte del diagrama. La métrica recall indica el porcentaje de acierto de un equipo, en función de todos los elementos que debería formar parte del diagrama. Esta métrica parece que coincide con la métrica aciertos, pero dado el sistema utilizado para calcular los valores de la matriz de confusión, no coinciden. La métrica accuracy es otra forma de medir la precisión, pero se dividen los aciertos entre todos los elementos de la matriz de confusión.

E Diagramas de Caja

Este anexo presenta los diagramas de caja de los datos asociados a las métricas de la eficacia, la eficiencia, la satisfacción y la calidad, obtenidos durante la realización de las tareas del experimento. Los diagramas muestras los datos agrupados en función del tratamiento (el chatbot SOCIO o la aplicación web Creately) y la secuencia (Creately-SOCIO o SOCIO-Creately), y el tratamiento y la tarea/periodo (1 o 2). A lo largo del anexo la secuencia Creately-SOCIO se abreviará como secuencia CR-SC, y la secuencia SOCIO-Creately como SC-CR.

E.1 Diagramas de Caja para las Métricas de la Eficacia

La eficacia se mide en función del grado de completitud con el que los equipos finalizaron las tareas. La Figura 30 muestra el diagrama de caja correspondiente al grado de completitud con el que los equipos finalizaron la tarea, agrupado por secuencia y tratamiento. Los datos son muy similares para SOCIO y para Creately, aunque SOCIO presenta valores más bajos de completitud en la secuencia SC-CR.

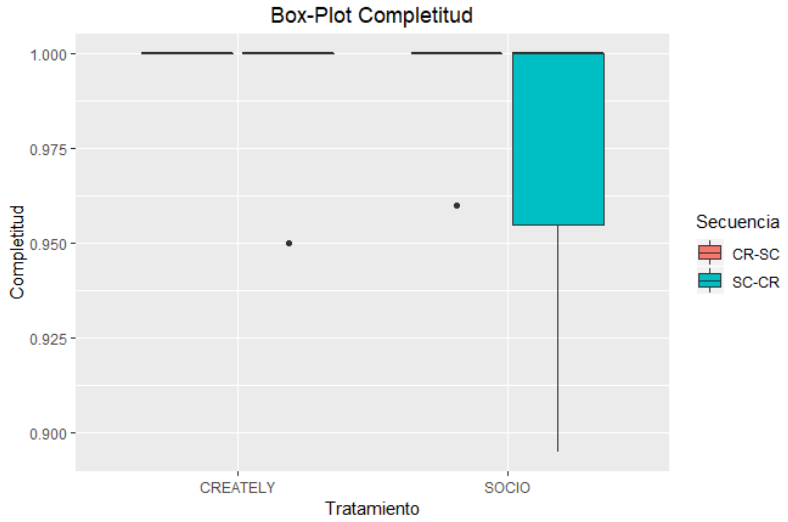


Figura 30: Diagramas de caja para el grado de completitud de las tareas, agrupado por tratamiento y secuencia.

La Figura 31 muestra el diagrama correspondiente al grado de completitud en que los equipos finalizaron la tarea, agrupado por tratamiento y tarea. Los datos son muy similares para SOCIO y para Creately, aunque SOCIO presenta valores más bajos en la tarea 1.

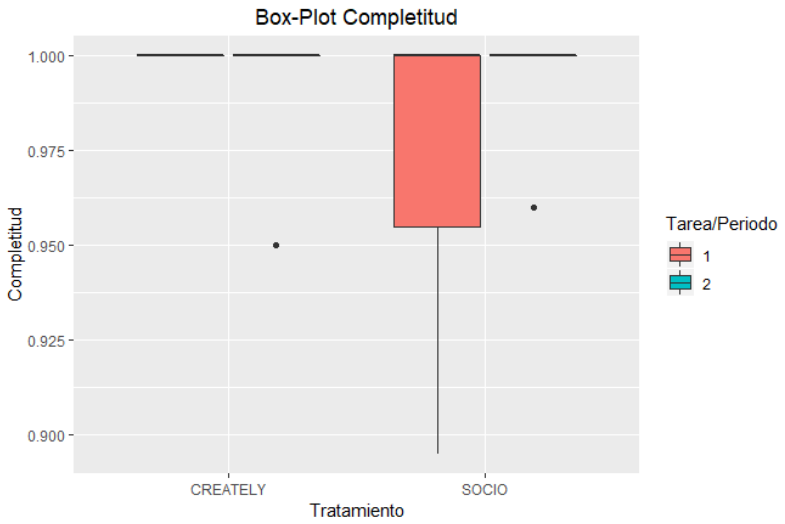


Figura 31: Diagrama de caja para el nivel de completitud de las tareas, agrupado por tratamiento y tarea.

E.2 Diagramas de Caja para las Métricas de la Eficiencia

El tiempo empleado por los equipos para realizar las tareas del experimento, así como los mensajes de discusión generados durante las mismas, son las métricas empleadas para medir la eficiencia. A continuación, se muestra el diagrama de cajas de los datos agrupados por secuencia y tratamiento, y por secuencia y tarea, para las métricas mencionadas.

Tiempo

La Figura 32 muestra el diagrama de caja correspondiente al tiempo empleado por los equipos para completar la tarea, agrupado por tratamiento y secuencia. En ambas secuencias, se observa que los tiempos asociados a SOCIO son más bajos que los asociados a Creately. También, se puede observar que los tiempos de la secuencia SC-CR son inferiores a los de la secuencia CR-SC para ambos tratamientos.

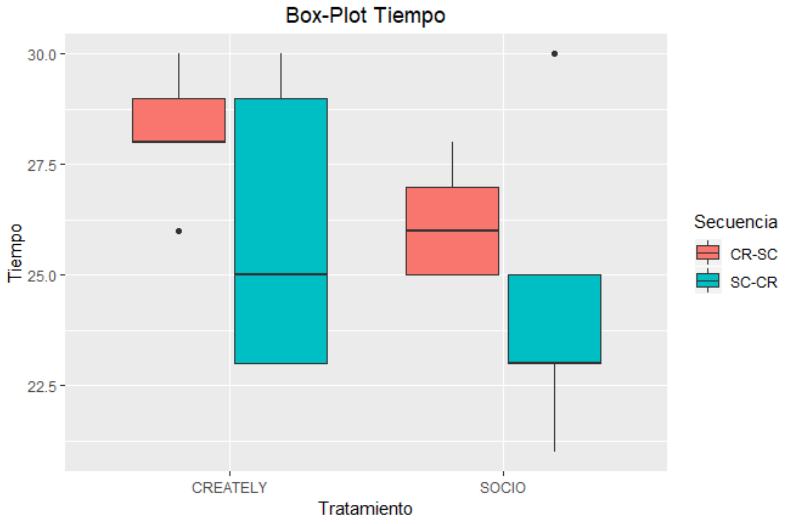


Figura 32: Diagrama de caja para el tiempo empleado en completar la tarea empleando SOCIO y Creately, agrupado por tratamiento y secuencia.

La Figura 33 muestra el diagrama de caja correspondiente al tiempo agrupado por tratamiento y tarea. Se observa que Creately obtiene tiempos más bajos cuando es aplicado en la segunda tarea, mientras que SOCIO obtiene mejores resultados cuando es aplicado en la primera. Comparando las herramientas, SOCIO presenta para la tarea 1 tiempos más bajos que Creately, mientras que Creately presenta tiempos más bajos que SOCIO para la tarea 2, pero no distan tanto los valores.

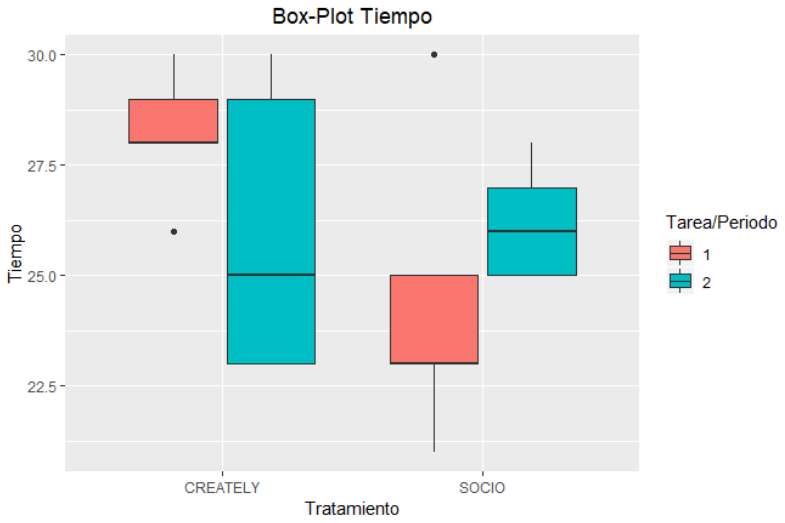


Figura 33: Diagrama de caja para el tiempo empleado en completar la tarea empleando SOCIO y Creately, agrupado por tratamiento y por tarea**.**

Número de mensajes de discusión

La Figura 34 muestra el diagrama de caja correspondiente al número de mensajes de discusión generados por los equipos durante la realización de las tareas, agrupados por tratamiento y secuencia. Se observa que los equipos de la secuencia CR-SC generaron más mensajes para Creately que los de la secuencia SC-CR. Para SOCIO, el número de mensajes en ambas secuencias es similar, aunque más disperso para la secuencia CR-SC. También, se puede ver que en la secuencia CR-SC el número de mensajes para Creately es mayor que para SOCIO, mientras que en la secuencia SC-CR, el número de mensajes es similar aunque más disperso para Creately.

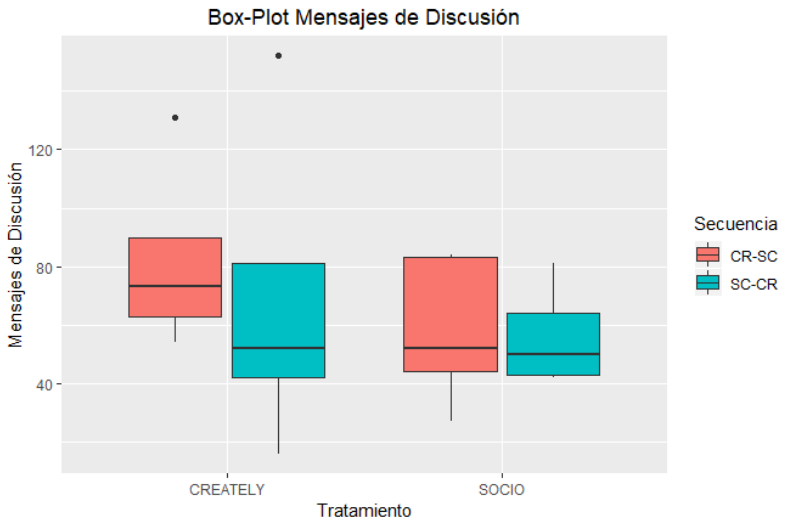


Figura 34: Diagrama de caja para el número de mensajes de discusión de SOCIO y Creately, agrupado por tratamiento y secuencia.

La Figura 35 muestra el diagrama de caja correspondiente al número de mensajes de discusión, agrupados por tratamiento y tarea. Para Creately, se puede observar que se generan más mensajes en la tarea 1 que en la tarea 2, mientras que para SOCIO el número de mensajes es similar, aunque más disperso para la tarea 2. Comparando los resultados de ambas herramientas por tarea, se puede ver que en la tarea 1, el número de mensajes para Creately es mayor que para SOCIO. En la tarea 2, el número de mensajes es similar para ambos tratamientos.

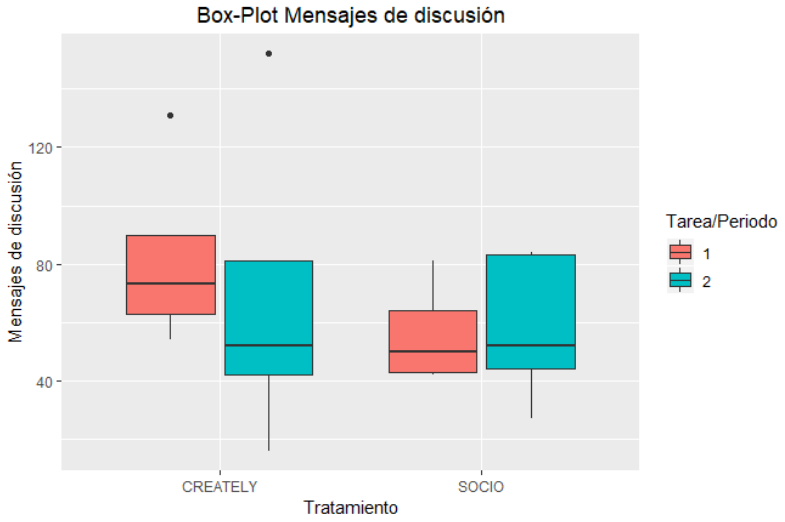


Figura 35: Diagrama de caja para el número de mensajes de discusión para SOCIO y Creately, agrupado por tratamiento y tarea.

E.3 Diagramas de Caja para las Métricas de la Satisfacción

La Figura 36 muestra el diagrama de caja correspondiente a las puntuaciones de satisfacción de los equipos, agrupadas por tratamiento y secuencia. En la secuencia CR-SC, SOCIO obtiene mejores puntuaciones, pero no distan tanto de las puntuaciones de

Creately como ocurre en la secuencia SC-CR. A su vez, se observa que Creately obtiene mejores resultados en la secuencia A, y SOCIO en la secuencia B.

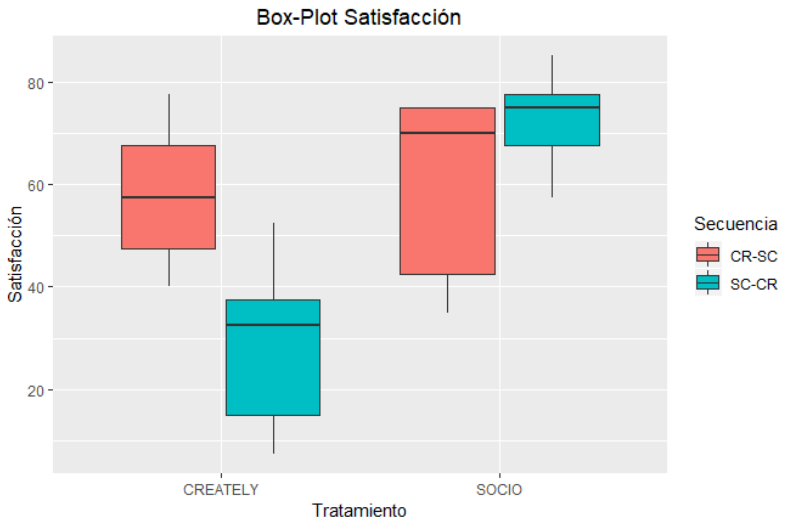


Figura 36: Diagrama de caja para las puntuaciones de satisfacción de SOCIO y Creately, agrupadas por tratamiento y secuencia.

La Figura 37 muestra el diagrama de caja correspondiente a las puntuaciones de satisfacción de los equipos, agrupadas por tratamiento y tarea. Las puntuaciones que la herramienta Creately obtiene al ser empleada en la tarea 1 son superiores a las obtenidas al ser empleada en la tarea 2, lo cual ya se observaba en la Figura 36 (Creately obtiene mejores resultados en la secuencia CR-SC). También, SOCIO muestra mejores puntuaciones cuando se aplica en la tarea 1. Comparando las herramientas, los resultados de SOCIO son más altos que los de Creately en ambas tareas.

Figura 37: Diagrama de caja para las puntuaciones de satisfacción de SOCIO y Creately, por tratamiento y tarea.

E.4 Diagramas de Caja para las Métricas de la Calidad

Las métricas para evaluar la calidad son accuracy, precisión, recall, error y aciertos. A continuación se presenta el diagrama de caja de los datos asociados a las mismas agrupados por tratamiento y secuencia, y por tratamiento y tarea/periodo.

Accuracy

La Figura 38 muestra el diagrama de caja correspondiente a las puntuaciones de accuracy asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por secuencia y tratamiento. Los diagramas elaborados con Creately obtienen mejores puntuaciones en la secuencia CR-SC que en la SC-CR, mientras que los elaborados con SOCIO obtienen mejores resultados en la secuencia SC-CR que en la secuencia CR-SC. Comparando ambos tratamientos, Creately obtiene puntuaciones superiores a las de SOCIO en la secuencia CR-SC, mientras que SOCIO obtiene puntuaciones superiores a las de Creately en la secuencia SC-CR.

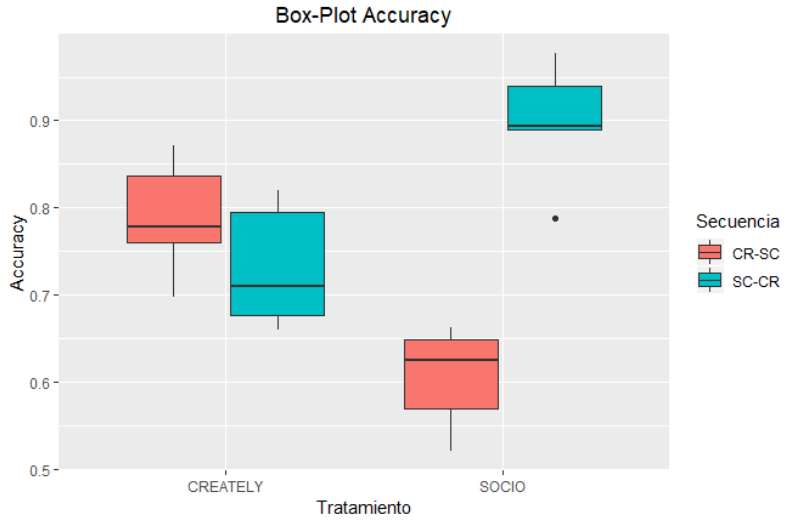


Figura 38: Diagrama de caja para las puntuaciones de accuracy para SOCIO y Creately, agrupadas por tratamiento y secuencia.

La Figura 39 muestra el diagrama de caja correspondiente a las puntuaciones de accuracy asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por tratamiento y tarea. Se observa que ambas herramientas obtienen mejores puntuaciones en la tarea 1. La diferencia entre las puntuaciones de la tarea 1 y la tarea 2 para Creately no es tan significativa coma la que se aprecia para SOCIO. En la tarea 1 las puntuaciones de SOCIO son más altas que las de Creately. En la tarea 2, Creately obtiene mejores puntuaciones que SOCIO.

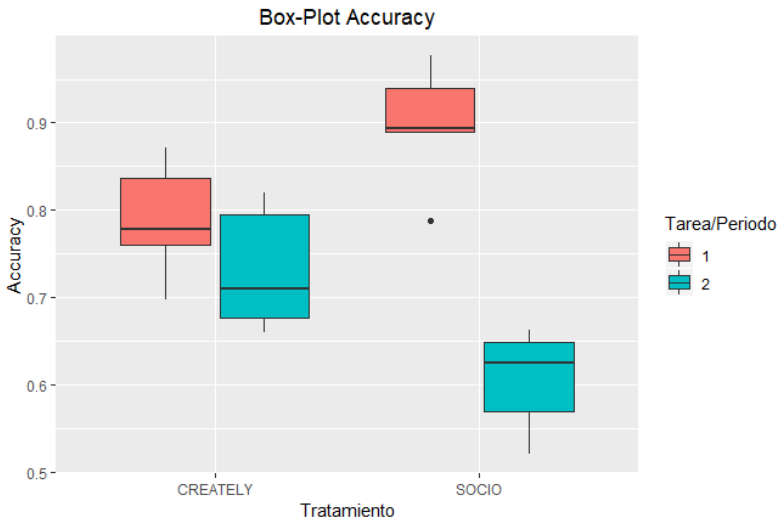


Figura 39: Diagrama de caja para las puntuaciones de accuracy para SOCIO y Creately, agrupadas por tratamiento y tarea.

Precisión

La Figura 40 muestra el diagrama de caja correspondiente a las puntuaciones de precisión asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por secuencia y tratamiento. Se puede observar que Creately obtiene puntuaciones similares en ambas secuencias, mientras que SOCIO obtiene puntuaciones más elevadas en la secuencia SC-CR que en la secuencia CR-SC. Comparando las herramientas, Creately obtiene mejores puntuaciones que SOCIO en la secuencia CR-SC, mientras que SOCIO obtiene mejores puntuaciones que Creately en la secuencia SC-CR.

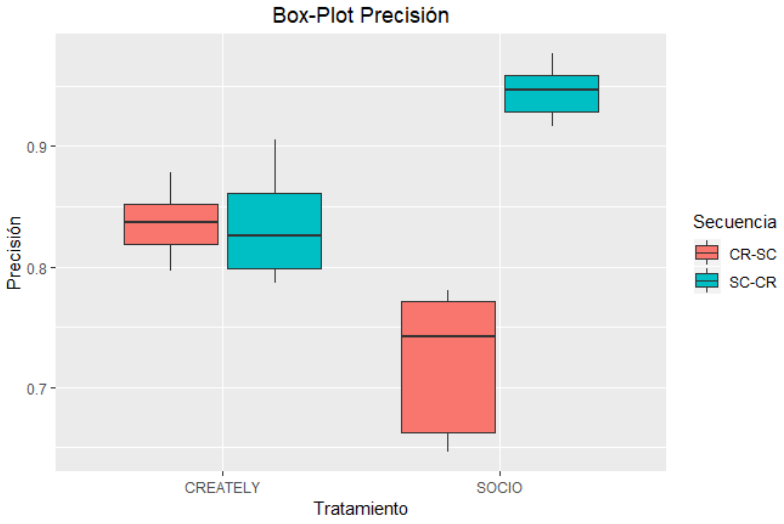


Figura 40: Diagrama de caja de las puntuaciones de precisión para SOCIO y Creately, agrupadas por tratamiento y secuencia.

La Figura 41 muestra el diagrama de caja correspondiente a las puntuaciones de precisión asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por tratamiento y tarea. Se observa que Creately presenta puntuaciones similares en ambas tareas, aunque en la tarea 2 están más dispersas. SOCIO obtiene

mejores puntuaciones para la tarea 1. Comparando ambas herramientas, SOCIO muestra puntuaciones más altas que Creately en la tarea 1, mientras que Creately muestra puntaciones más altas que SOCIO en la tarea 2.

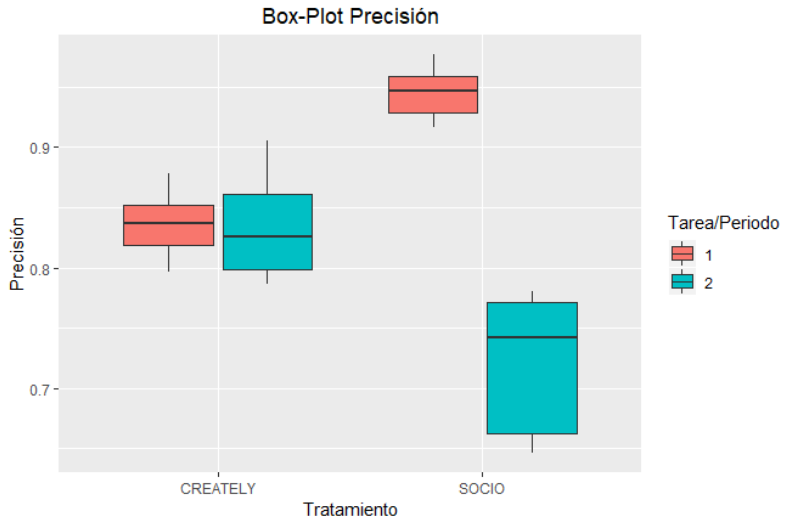


Figura 41: Diagrama de caja de las puntuaciones de precisión para SOCIO y Creately, agrupadas por tratamiento y tarea.

Recall

La Figura 42 muestra el diagrama de caja correspondiente a las puntuaciones de recall, asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por secuencia y tratamiento. Los diagramas elaborados con Creately obtienen mejores puntuaciones en la secuencia CR-SC que en la SC-CR, mientras que los elaborados con SOCIO obtienen mejores resultados en la secuencia SC-CR que en la CR-SC. Comparando ambos tratamientos, Creately obtiene puntuaciones de recall superiores a las de SOCIO en la secuencia CR-SC, mientras que SOCIO obtiene puntuaciones superiores a las de Creately en la secuencia SC-CR.

Figura 42: Diagrama de caja para las puntuaciones de recall para SOCIO y Creately, agrupadas por tratamiento y secuencia.

La Figura 43 muestra el diagrama de caja correspondiente a las puntuaciones de recall asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por tratamiento y tarea. Tanto SOCIO como Creately muestran mejores puntuaciones de recall en la tarea 1 que en la tarea 2. Comparando ambas herramientas, las puntuaciones en la tarea 1 son similares, siendo superiores las de SOCIO. En la tarea 2, son superiores las de Creately a las de SOCIO.

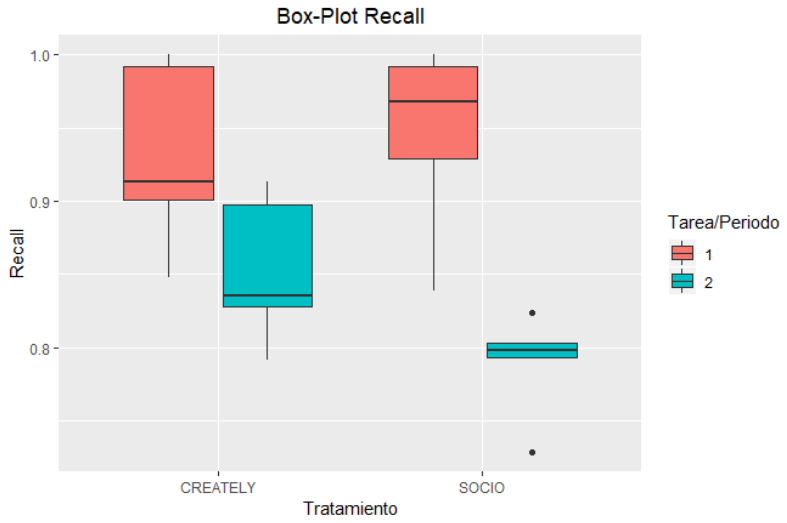


Figura 43: Diagrama de caja de las puntuaciones de recall para SOCIO y Creately, agrupadas por tratamiento y tarea.

Aciertos

La Figura 44 muestra el diagrama de caja correspondiente a las puntuaciones de aciertos asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por secuencia y tratamiento. Los diagramas elaborados con Creately obtienen mejores puntuaciones de aciertos en la secuencia CR-SC que en la SC-CR, mientras que los elaborados con SOCIO obtienen mejores resultados en la secuencia SC-CR que en la CR-SC. Comparando ambos tratamientos, Creately obtiene puntuaciones superiores a las de SOCIO en la secuencia CR-SC, mientras que SOCIO obtiene puntuaciones superiores a las de Creately en la secuencia SC-CR.

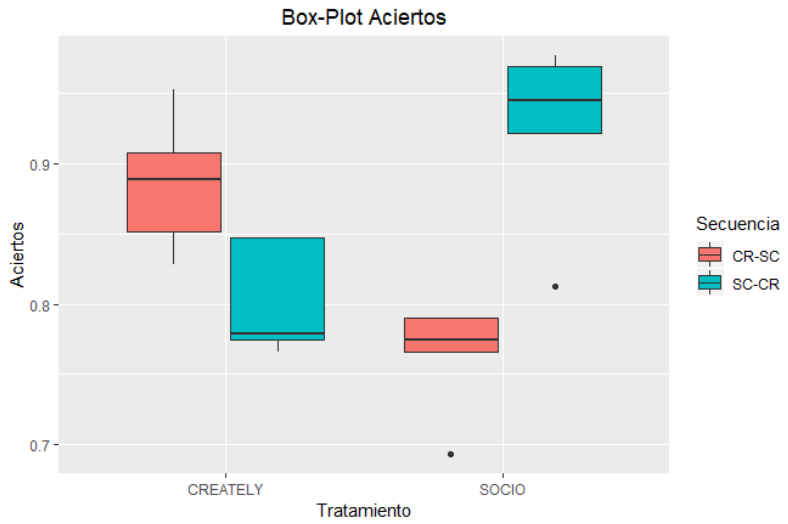


Figura 44: Diagrama de caja de las puntuaciones de aciertos para SOCIO y Creately, agrupadas por tratamiento y secuencia.

La Figura 45 muestra el diagrama de caja correspondiente a las puntuaciones de aciertos asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por tratamiento y tarea. Se observa que ambas herramientas obtienen mejores puntuaciones en la tarea 1. La diferencia entre las puntuaciones de la tarea 1 y la tarea 2 para Creately, no es tan significativa coma la que se aprecia para SOCIO. En la tarea 1 las puntuaciones de SOCIO son más altas que las de Creately. En la tarea 2, aunque son similares, Creately obtiene puntuaciones más altas y dispersas que SOCIO.

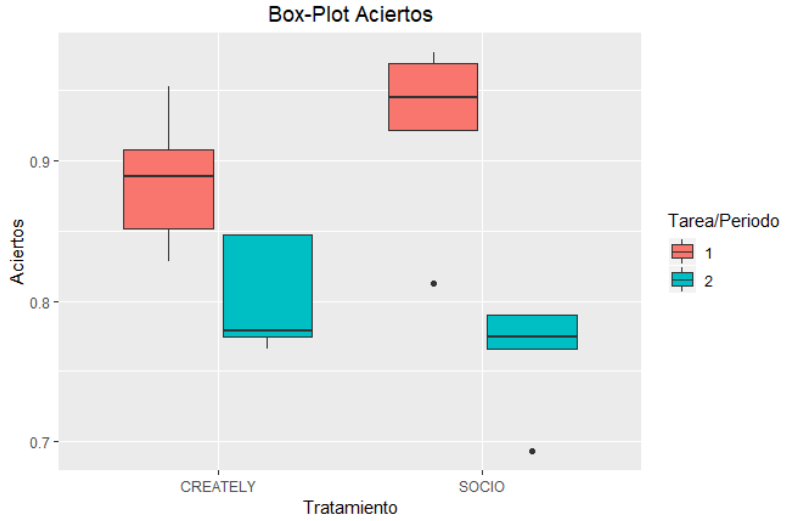


Figura 45: Diagrama de las puntuaciones de aciertos para SOCIO y Creately, agrupadas por tratamiento y tarea.

Error

La Figura 46 muestra el diagrama de caja correspondiente a las puntuaciones de errores cometidos en los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por secuencia y tratamiento. Se observa que Creately presenta puntuaciones de

error mayores en la secuencia SC-CR que en la CR-SC, mientras para SOCIO ocurre lo contrario, obtiene mayor puntuaciones de error en la secuencia CR-SC. Las puntuaciones entre las secuencias distan más para SOCIO que para Creately. También, se aprecia en la secuencia CR-SC, puntuaciones de error más elevadas para SOCIO, sin embargo, en la secuencia SC-CR ocurre lo contrario, es Creately la que obtiene valores de error más altos.

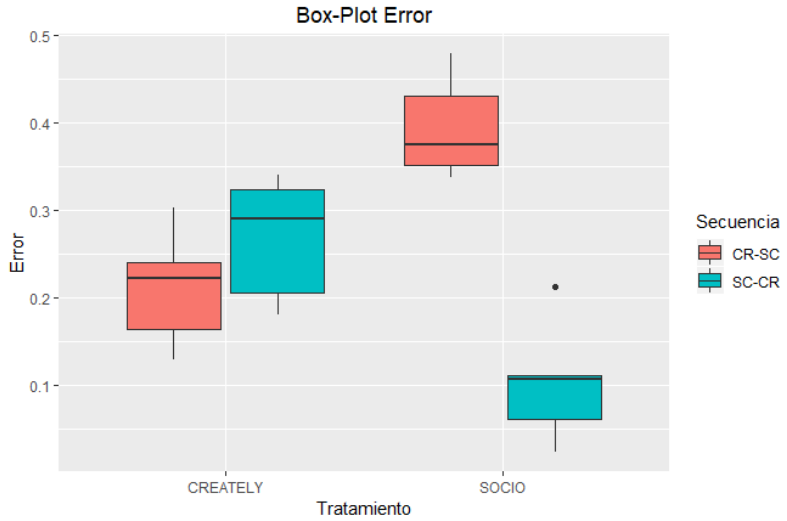


Figura 46: Diagrama de caja de las puntuaciones de error para SOCIO y Creately, agrupadas por tratamiento y secuencia.

La Figura 47 muestra el diagrama de caja correspondiente a las puntuaciones de error asociadas a los diagramas elaborados por los equipos empleando SOCIO y Creately, agrupadas por tratamiento y tarea. Se observa que ambas herramientas obtienen puntuaciones de error, es decir, peores puntuaciones, en la tarea 2. La diferencia entre las puntuaciones de la tarea 1 y la tarea 2 para Creately, no es tan significativa como la que se aprecia para SOCIO. En la tarea 1 las puntuaciones de Creately son más altas que las de SOCIO. En la tarea 2, SOCIO obtiene peores puntuación que Creately.

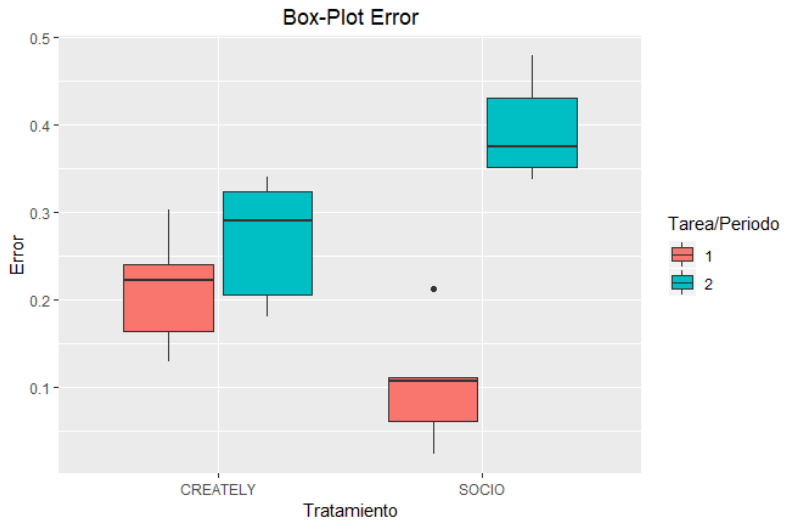


Figura 47: Diagrama de caja de las puntuaciones de error para SOCIO y Creately.

F Tamaño del Efecto

En el análisis de los datos, el tamaño del efecto de un factor, es decir, la magnitud de las diferencias generados por el factor, ha sido medido a través de *d* de Cohen y su error estándar.

En el análisis de los datos asociados tanto a SOCIO como a Creately, se ajusta para cada variable un modelo lineal mixto (Vegas et al., 2016) y se complementa calculando el tamaño del efecto. El cálculo del tamaño del efecto del tratamiento (en este caso, el chatbot SOCIO o la aplicación web Creately) sobre cada métrica, se realiza mediante la *d* de Cohen para los tratamientos y su error estándar, *SE*. Se emplean las fórmulas proporcionadas en (Higgins & Green, 2006):

$$d = \frac{\mathrm{MD}}{\sqrt{\frac{SD_{SC}^2 + SD_{CR}^2}{2}}} \qquad \mathrm{SE(d)} = \sqrt{\frac{1}{N} + \frac{d^2}{N}} \times \sqrt{2(1 - Corr)}$$

En estas fórmulas:

- MD, media de las diferencias de los valores obtenidas con SOCIO y con Creately para una métrica.
- $SD_{SC,}$ desviación típica de los valores de una métrica obtenidos para SOCIO.
- SD_{CR}, desviación típica de los valores de una métrica obtenidos para Creately.
- N, número de valores de cada métrica, en este caso, N =10.
- Corr, correlación entre los valores de una métrica obtenidos para SOCIO y para Creately.

En el análisis de los datos asociados a la interacción con el chatbot SOCIO, se realizan una serie de tests-t independientes (Field, 2013), uno por cada métrica, para comparar la media de las diferentes interacciones generadas durante las tareas 1 y 2. Los resultados del test-t se complementan, calculando el tamaño del efecto asociado a la tarea, con la *d* de Cohen y su error estándar, *SE*, siguiendo la fórmula de (Borenstein et al., 2011):

$$d = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{2}}} \qquad \mathrm{SE(d)} = \frac{N_1 + N_2}{N_1 \times N_2} + \frac{d^2}{2 \times (N_1 + N_2)}$$

En estas fórmulas:

- $M_{1,}$ media de los valores de una métrica correspondientes a la tarea 1.
- $M_{2,}$ media de los valores de una métrica correspondientes a la tarea 2.
- $SD_{1,}$ desviación típica de los valores de una métrica obtenidos en la tarea 1.
- $SD_{2,}$ desviación típica de los valores de una métrica obtenidos en la tarea 2.
- N_1, número de valores de una métrica correspondientes a la tarea 1.
- N_2, número de valores de una métrica correspondientes a la tarea 2.

En ambos análisis, el tamaño del efecto se considera pequeño si la *d* de Cohen es menor de 0,5, mediano si están entre 0,5 y 0,8, y grande si el mayor de 0.8 (Borenstein et al., 2011).