Trabajo 1: Diseño de un experimento controlado

Daniel Jiménez García
Sergio Muñoz Gómez
Ángela Caiqing Pousada Morán

Fecha: 9 de octubre de 2025

Master MITSS. ISE
Universidad Politécnica de Valencia

Resumen

Añadir resumen después de terminar el trabajo

Capítulo 1

Motivación

1.1. Problema a investigar

En la actualidad existen numerosos métodos para evaluar la usabilidad y experiencia de usuario (UX) a la hora de utilziar sistemas interactivos. Sin embargo, no existe un consenso claro sobre cuál de ellos resulta más adecuado para cada sistema. El problema que se aborda en este trabajo es la falta de evidencia empírica que permita comparar la **efectividad, eficiencia y satisfacción de los evaluadores** cuando aplican un método empírico (*A/B Testing*) frente a un método de inspección grupal (*Recorrido Pluralista*).

1.2. Definición del experimento

Analizar los métodos de evaluación de usabilidad **A/B Testing** y **Recorrido Pluralista** con el propósito de comparar su capacidad para detectar problemas de usabilidad y medir su impacto en la satisfacción y eficiencia de los evaluadores.

1.3. Contexto

El experimento se llevará a cabo en un entorno controlado de laboratorio, sobre un prototipo web de complejidad media. Se contará con estudiantes de ingeniería de software y diseño UX como participantes. Cada sesión incluirá la realización de tareas específicas dentro del sistema y la evaluación mediante uno de los dos métodos propuestos.

Capítulo 2

Trabajos relacionados

Diversos estudios previos han comparado métodos de evaluación de usabilidad. Nielsen (1994) definió la *Evaluación Heurística* como técnica de inspección rápida, mientras que Rubin y Chisnell (2008) destacaron la importancia de los métodos empíricos como el *User Testing* para observar comportamientos reales.

El **A/B Testing** se utiliza ampliamente en el ámbito web para medir diferencias en satisfacción y eficiencia entre dos versiones de una interfaz (Kohavi et al., 2009). Por su parte, el **Recorrido Pluralista** (Bias, 1994) combina la revisión de expertos, diseñadores y usuarios, permitiendo identificar problemas desde perspectivas diversas.

A pesar de su uso extendido, no existen suficientes estudios comparativos entre ambos métodos bajo condiciones controladas, especialmente considerando variables subjetivas como satisfacción o percepción de facilidad de uso.

Capítulo 3

Descripción del diseño

3.1. Hipótesis y variables

Hipótesis:

- H_{01}: No existen diferencias significativas en la **satisfacción** entre A/B Testing y Recorrido Pluralista.
- H_{11}: A/B Testing genera mayor satisfacción percibida que el Recorrido Pluralista.
- H_{02}: No existen diferencias significativas en la **usabilidad percibida** entre los métodos.
- H_{12}: El Recorrido Pluralista detecta más problemas de usabilidad que A/B Testing.
- H_{03}: No existen diferencias significativas en la **eficiencia** (tiempo medio de tareas).
- H_{13}: Los evaluadores son más eficientes (menor tiempo) usando A/B Testing.

Variables:

- **Variable independiente:** Método de evaluación (A/B Testing, Recorrido Pluralista).
- **Variables dependientes:**
  - **Satisfacción:** medida mediante cuestionario Likert (1–7).
  - **Usabilidad:** puntuación SUS (*System Usability Scale*).
  - **Eficiencia:** tiempo promedio (en segundos) para completar tareas.
- **Variables de control:** tipo de tarea, complejidad del prototipo, experiencia previa de los participantes.

3.2. Diseño del experimento

El experimento seguirá un diseño **intra-sujetos contrabalanceado**. Cada participante aplicará ambos métodos (A/B Testing y Recorrido Pluralista) sobre dos conjuntos de tareas diferentes, en sesiones separadas por 24 horas para mitigar los efectos de aprendizaje y fatiga. El orden de aplicación de los métodos será aleatorizado para contrarrestar sesgos de orden.

3.2.1. Chatty

El presente experimento tiene como objetivo comparar la aplicación de dos métodos de evaluación de usabilidad —A/B Testing y Recorrido Pluralista— cuando son utilizados por desarrolladores de software en el contexto de la evaluación de una aplicación web. A través de este diseño se pretende obtener evidencia empírica sobre cuál de los dos métodos resulta más adecuado y eficiente desde la perspectiva de los propios evaluadores, considerando tres variables principales: satisfacción, usabilidad percibida y eficiencia, esta última medida tanto en tiempo invertido como en esfuerzo percibido.

De acuerdo con el marco GQM (Goal-Question-Metric), el propósito del experimento es analizar el comportamiento y las percepciones de los desarrolladores al aplicar ambos métodos de evaluación. La pregunta que guía el estudio es: ¿qué método proporciona una mejor experiencia de uso y resultados más satisfactorios a los evaluadores de usabilidad? Para responderla, se utilizarán métricas tanto objetivas (tiempos de ejecución, número y severidad de los problemas detectados) como subjetivas (puntuaciones de satisfacción y usabilidad percibida obtenidas mediante el cuestionario USE adaptado).

El diseño experimental se estructura en torno a tres hipótesis principales. En primer lugar, se plantea que no existen diferencias significativas en la satisfacción de los desarrolladores al aplicar A/B Testing o Recorrido Pluralista, aunque se espera observar una mayor satisfacción en el método más intuitivo o colaborativo. En segundo lugar, se plantea que no existen diferencias en la usabilidad percibida de los métodos, pero se anticipa que uno de ellos podría ser percibido como más fácil de aplicar. Finalmente, se formula la hipótesis de que no existen diferencias en la eficiencia entre ambos métodos, medida como el tiempo total necesario para realizar las tareas y el esfuerzo subjetivo requerido, aunque se espera que A/B Testing, al ser más automatizado, requiera menos tiempo que el Recorrido Pluralista.

En este experimento, la variable independiente es el método de evaluación de usabilidad empleado, que puede ser A/B Testing o Recorrido Pluralista. Las variables dependientes son la satisfacción del desarrollador, la usabilidad percibida del método, la eficiencia (medida en tiempo y esfuerzo) y la efectividad, entendida como la cantidad y severidad de los problemas de usabilidad detectados. Como variables de control se consideran el nivel de experiencia del participante, la complejidad de las tareas asignadas y el orden en que se aplican los métodos.

El diseño elegido es de tipo intra-sujetos contrabalanceado. Cada participante aplicará ambos métodos de evaluación sobre distintos prototipos de una misma aplicación web, de modo que se puedan comparar directamente las percepciones

individuales sin la influencia de diferencias entre sujetos. Para evitar el efecto de aprendizaje, los participantes se dividirán en dos grupos. El primer grupo aplicará primero el método A/B Testing y posteriormente el Recorrido Pluralista, mientras que el segundo grupo seguirá el orden inverso. Las sesiones estarán separadas por al menos veinticuatro horas, con el fin de mitigar posibles sesgos derivados de la familiaridad con las tareas o el sistema. Además, los prototipos serán contrabalanceados entre los grupos para garantizar que ambos métodos se apliquen sobre versiones equivalentes de la aplicación.

La muestra estará compuesta por entre dieciséis y veinte desarrolladores de software, incluyendo tanto estudiantes de posgrado como profesionales con experiencia en desarrollo web o en diseño de interfaces. Los participantes deberán contar con conocimientos básicos sobre usabilidad o experiencia de usuario y disponibilidad para asistir a dos sesiones experimentales. Se excluirán aquellos que hayan participado en el piloto del experimento o que conozcan previamente los prototipos utilizados. Este tamaño muestral se considera adecuado para un diseño intra-sujetos con un efecto moderado y una potencia estadística aceptable.

Los objetos experimentales serán dos prototipos web interactivos que representan versiones diferentes de la misma aplicación, con funcionalidades equivalentes y una complejidad moderada. Las tareas que los desarrolladores deberán realizar serán realistas y de corta duración, por ejemplo, localizar un producto y simular una compra, modificar una configuración en el perfil de usuario o buscar información específica dentro del sistema. En el método A/B Testing, se evaluarán las diferencias entre las dos versiones de la interfaz, registrando el rendimiento y los tiempos de interacción. En el caso del Recorrido Pluralista, los desarrolladores participarán en una sesión grupal junto a un diseñador y un experto en usabilidad, discutiendo colectivamente la ejecución de las tareas y los problemas detectados.

El procedimiento experimental constará de dos sesiones por participante. En la primera, se explicará brevemente el método a utilizar y las tareas que deberán completarse. A continuación, el participante aplicará el método correspondiente sobre uno de los prototipos, registrándose los tiempos, observaciones y hallazgos detectados. Una vez finalizada la sesión, se completará el cuestionario USE adaptado, que mide la percepción de utilidad, facilidad de uso, facilidad de aprendizaje y satisfacción general con el método aplicado. En la segunda sesión, que se llevará a cabo al menos un día después, el participante repetirá el procedimiento aplicando el segundo método sobre el segundo prototipo. Al finalizar, se recogerán de nuevo las métricas objetivas y las percepciones subjetivas. En el caso del Recorrido Pluralista, se registrará tanto el tiempo total de la sesión grupal como la participación individual de cada desarrollador, con el fin de normalizar los datos y poder compararlos con el A/B Testing.

Para la recopilación de datos se utilizarán diversas herramientas. El tiempo por tarea se medirá mediante un software de registro automático; los hallazgos de usabilidad se documentarán en una plantilla estandarizada donde se indique la descripción del problema, su severidad y frecuencia. Los cuestionarios USE y SUS permitirán cuantificar la percepción de usabilidad y satisfacción de los participantes, mientras que un breve formulario demográfico recogerá información sobre la experiencia y el rol de cada desarrollador. Además, se almacenarán observaciones cualitativas y

comentarios de los participantes sobre las ventajas y limitaciones percibidas en cada método.

El análisis de los resultados combinará técnicas cuantitativas y cualitativas. Se realizará un análisis descriptivo de las puntuaciones obtenidas en los cuestionarios y de los tiempos registrados, calculando medias, medianas y desviaciones estándar. Posteriormente, se aplicarán pruebas estadísticas inferenciales, como la prueba t para muestras relacionadas o el test de Wilcoxon cuando los datos no sigan una distribución normal, con un nivel de significación de $\alpha = 0{,}05$. También se calcularán los tamaños del efecto (Cohen's d) para estimar la magnitud de las diferencias observadas. Los comentarios abiertos se analizarán mediante codificación temática, identificando patrones recurrentes en las opiniones de los desarrolladores sobre los métodos comparados.

Con el fin de garantizar la validez interna, el orden de aplicación de los métodos se contrabalanceará y las sesiones se espaciarán en el tiempo para reducir efectos de aprendizaje. La validez externa se abordará seleccionando participantes con distintos niveles de experiencia, lo que permitirá una mejor generalización de los resultados. La validez de constructo se refuerza mediante el uso de instrumentos validados, como las escalas USE y SUS, mientras que la validez de conclusión se preservará mediante un análisis estadístico apropiado y el control de la potencia del estudio.

Antes de la ejecución definitiva se realizará un piloto con tres a cinco desarrolladores, con el objetivo de validar la claridad de las instrucciones, la duración de las sesiones y la adecuación de las tareas. Los resultados del piloto servirán para ajustar los materiales y tiempos, y detectar posibles problemas en la instrumentación o el análisis. Tras el piloto, se planifica un cronograma de cinco semanas: la primera dedicada al diseño y revisión del protocolo, la segunda al piloto y ajustes, las semanas tres y cuatro a la ejecución de las sesiones experimentales y la quinta al análisis de resultados y redacción del informe final.

En conjunto, este diseño experimental permitirá comparar de forma rigurosa y controlada la experiencia de los desarrolladores al aplicar los métodos A/B Testing y Recorrido Pluralista, evaluando sus ventajas, limitaciones y adecuación para ser empleados en contextos de desarrollo de software orientado a la usabilidad.

3.3. Selección de sujetos

EL ESTUDIO VA DIRIGIDO A DESARROLLADORES QUE SEAN DESARROLLADORES

Se seleccionarán entre 16 y 20 participantes, con formación básica en usabilidad o experiencia de usuario. Los criterios de inclusión son:

- Familiaridad con interfaces web.
- No haber participado previamente en experimentos similares.

Cada participante firmará un consentimiento informado y completará un cuestionario demográfico inicial.

3.4. Objetos e instrumentación

Los objetos experimentales serán dos versiones de una aplicación web (versión A y versión B) con diferencias en diseño o disposición visual.

- **Instrumentos de medida:** - Cronómetro o software de registro (tiempo por tarea). - Cuestionario SUS (Brooke, 1996). - Escala de satisfacción Likert (1–7). - Registro de observaciones y errores detectados.

3.5. Evaluación de la validez

Validez interna: se controlará el efecto orden mediante contrabalanceo. Se mantendrá la misma complejidad de tareas y condiciones experimentales.

Validez externa: los resultados serán aplicables a contextos similares (evaluaciones de usabilidad de aplicaciones web con participantes semiexpertos).

Validez de constructo: las métricas SUS y Likert se seleccionan por su fiabilidad y validez en estudios de usabilidad.

Validez de conclusión: se emplearán pruebas t pareadas (o Wilcoxon si no hay normalidad) con $\alpha = 0{,}05$, y tamaños del efecto (Cohen's d) para cuantificar diferencias.

Referencias

- Esto es referncia buscada sobre qué es el pluralistic walk...[1]

Lo de abajo es puro ChatGPT(no me escondo)

- Bias, R. (1994). *The Pluralistic Usability Walkthrough: Coordinated Empathies.* In Nielsen & Mack (Eds.), Usability Inspection Methods.
- Brooke, J. (1996). *SUS: A Quick and Dirty Usability Scale.* Usability Evaluation in Industry.
- Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. (2009). *Controlled experiments on the web: Survey and practical guide.* Data Mining and Knowledge Discovery, 18(1).
- Nielsen, J. (1994). *Usability Engineering.* Morgan Kaufmann.
- Rubin, J., & Chisnell, D. (2008). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests.* Wiley.
- Wohlin, C., et al. (2012). *Experimentation in Software Engineering.* Springer.

Bibliografía

[1] Chauncey Wilson. «Chapter 5 - Pluralistic Usability Walkthrough». En: *User Interface Inspection Methods*. Ed. por Chauncey Wilson. Boston: Morgan Kaufmann, 2014, págs. 81-97. ISBN: 978-0-12-410391-7. DOI: https://doi.org/10.1016/B978-0-12-410391-7.00005-1. URL: https://www.sciencedirect.com/science/article/pii/B9780124103917000051.

Apéndice A

Anexo I: Tareas de evaluación

Cada participante completará un conjunto de tareas representativas en la aplicación web.

- **Tarea 1:** Localizar un producto y completar una compra simulada.
- **Tarea 2:** Cambiar una configuración de usuario en el perfil.

En A/B Testing, se compararán tiempos y errores entre versiones A y B. En el Recorrido Pluralista, se discutirán los pasos de interacción con un grupo de tres evaluadores (usuario, diseñador y experto).

Apéndice B

Anexo II: Cuestionario de satisfacción y usabilidad

- Escala SUS (10 ítems, 1–5).
- Escala de satisfacción general (Likert 1–7).
- Preguntas abiertas sobre percepción del método aplicado.