Trabajo 1: Diseño de un experimento controlado

Daniel Jiménez García
Sergio Muñoz Gómez
Ángela Caiqing Pousada Morán

Fecha: 30 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

Actualmente existen diversos métodos a la hora de evaluar la experiencia de usuario cuando estos utilizan sistemas interactivos. El problema que se aborda en este trabajo es la falta de evidencia empírica que permita comparar la **efectividad** 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 la capacidad para detectar problemas de usabilidad y medir su impacto en la satisfacción y eficiencia de los evaluadores.

1.3. Contexto

Para este experimento se parte de una actividad previa que se supone realizada. Esta actividad previa consiste en la evaluacón de una interfaz gráfica por parte de un equipo desarrollador sobre unos usuarios usando los dos métodos de evaluación **A/B Testing** y **Recorrido Pluralista**. Así pues, nuestro estudio se centra en recopilar información de cómo ha sido la experiencia de estos desarrolladores usando las dos técnicas de evaluación

Capítulo 2

Trabajos relacionados

Diversos estudios previos han comparado métodos de evaluación de usabilidad. Nielsen [1] definió la *Evaluación Heurística* como técnica de inspección informal realizada por expertos, mientras que Rubin [2] 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 [3]. Por su parte, el **Recorrido Pluralista** [4] combina la revisión de expertos, diseñadores y usuarios, permitiendo identificar problemas desde perspectivas diversas.

Capítulo 3

Descripción del diseño

3.1. Hipótesis y variables

Nuestro experimento está enfocado en comprobar qué método ofrece mejores resultados a la hora de encontrar fallos en el diseño de la interfaz y cuál prefieren los desarrollodores.

Hipótesis:

- H_{01}: No existen diferencias significativas en los **errores percibidos** entre A/B Testing y Recorrido Pluralista.
- H_{11}: Existen diferencias significativas en los **errores percibidos** entre A/B Testing y Recorrido Pluralista.
- H_{02}: No existen diferencias significativas en la **preferencia de uso de una metodogía frente a la otra** por parte de los desarrolladores.
- H_{12}: Existen diferencias significativas en la **preferencia de uso de una metodogía frente a la otra** por parte de los desarrolladores.

Variables:

- **Variable independiente:** Método de evaluación (A/B Testing, Recorrido Pluralista).
- **Variables dependientes:**
  - **Errores:** cometidos en el uso de la interfaz
  - **Preferencia:** por parte de los desarroladores
- **Variables de control:** tipo de tarea, complejidad del prototipo, experiencia previa de los participantes.

Los errores producidos por los usuarios podemos definirlos como una variable objetiva que mide el número de tareas realizadas de manera incorrecta sobre el total de las planteadas por el equipo tal y como se muestra en la ecuación 3.1

$$Errores = \frac{Tareas\ realizadas\ mal}{Total\ tareas\ realizadas} \tag{3.1}$$

Por otra parte, la satisfacción de los dessarrolladores es una variable subjetiva que se medirá aplicando el uso de el cuestionario mostrado en el Anexo B.

3.2. Diseño del experimento

Como se ha comentado, este experimento se desarrollaría después realizar la evaluación de la interfaz con las dos metodoglías por parte de los desarrolladores. Estas pruebas deben haber sido llevadas a cabo por dos grupos independites de desarrolladores aplicando cada uno los dos métodos de evaluación. Como se muestra en la tabla 3.1 cada equipo desarrollador evalua a otro dos grupos de usuarios (A. B, C y D) también independientes para que no producir sesgos en el experimento fruto del aprendizaje durante el experimento. Este enfoque del experimento es por un lado *between-subjects* en la parte de los usuarios y por otra parte *within-subjects* en el estudio de la satisfacción de los desarrolladores. Esto es debido a que se aleatoriza la selección de usuarios para cada método, y los desarrolladores aplican ambos para su estudio.

Grupo	**A/B Testing**	**Recorrido Pluralista**
Desarrolladores 1	A	B
Desarrolladores 2	C	D

Tabla 3.1: Diseño del experimento entre grupos (between-subjects)

Las tareas que deberán haber hecho los usuarios incluiyen actividades representativas de un uso real de la aplicación. El grupo de desarrolladores asignado al método de A/B Testing centrará su evaluación en la comparación de dos versiones de la interfaz, registrando métricas cuantitativas como el tiempo medio de ejecución, la tasa de éxito de las tareas y la frecuencia de errores entre otros. Por su parte, el grupo que emplee el Recorrido Pluralista llevará a cabo una revisión colaborativa, analizando las mismas tareas y discutiendo colectivamente los problemas de usabilidad identificados.

Una vez finalizada esa evaluación, se dará paso a nuestro estudio comparativo de ambos métodos. Nuesto estudio involucrará los errores encontrados en el estudio anterior reportado por los desarrolladores y el análisis de una encuesta realizada a los desarrolladores para medir su grado de satisfacción con los métodos empleados.

El proceso de evaluación más concretamente queda descrito como sigue. Primero dos equipos desarrolladores aplican las dos técnicas de evaluación sobre grupos de usuarios independientes. Con esta evaluación recuperan datos de los errores caputarados por estos.

3.3. Selección de sujetos

La selección de los participantes se realizará de manera intencionada, buscando garantizar la representatividad del perfil de usuario objetivo del estudio: desarrolladores de software con conocimientos básicos en usabilidad o experiencia de usuario (UX). Dado que el propósito del experimento es comparar la aplicación de los métodos A/B Testing y Recorrido Pluralista desde la perspectiva de evaluadores técnicos, resulta fundamental que los sujetos cuenten con una comprensión general de los principios de diseño de interfaces y evaluación de sistemas interactivos.

El estudio se dirigirá a un grupo de entre 16 y 20 participantes, un tamaño muestral adecuado para un diseño intra-sujetos contrabalanceado, donde cada participante aplica ambos métodos. Este rango permite obtener una potencia estadística suficiente para detectar diferencias de tamaño medio (Cohen's $d \approx 0{,}5$) con un nivel de significación de 0.05, reduciendo a la vez la variabilidad interindividual.

Criterios de inclusión

- Ser desarrollador o estudiante avanzado de ingeniería de software, diseño de interacción o disciplinas afines.
- Poseer conocimientos básicos sobre usabilidad o experiencia de usuario, acreditados mediante formación previa o experiencia práctica.
- Tener familiaridad con interfaces web interactivas.
- Contar con disponibilidad para asistir a dos sesiones experimentales, separadas por al menos 24 horas.
- Aceptar voluntariamente participar en el estudio, firmando el consentimiento informado.

Criterios de exclusión

- Haber participado previamente en el piloto del experimento o en estudios similares que involucren los mismos métodos.
- Poseer un conocimiento profundo o especializado sobre los prototipos empleados, lo que podría sesgar la evaluación.
- Presentar dificultades técnicas o de comunicación que impidan el correcto desarrollo de las tareas experimentales.

Procedimiento de reclutamiento

Los participantes serán reclutados a través de convocatorias internas en facultades de ingeniería y diseño, y mediante invitaciones personales a profesionales en activo del ámbito del desarrollo web. Se ofrecerá información detallada sobre los objetivos del estudio, la naturaleza de las tareas a realizar y la duración estimada de

cada sesión (aproximadamente 60 minutos). La participación será voluntaria y no remunerada, aunque se podrá ofrecer una constancia de participación académica.

Antes de iniciar el experimento, cada participante completará un cuestionario demográfico donde se recogerán datos como edad, nivel educativo, años de experiencia en desarrollo, conocimiento previo de técnicas de evaluación de usabilidad y frecuencia de participación en proyectos web. Esta información servirá para describir la muestra y analizar posibles correlaciones entre la experiencia previa y las percepciones de los métodos evaluados.

Consideraciones éticas

El estudio cumplirá con los principios éticos establecidos por la Declaración de Helsinki y las normas institucionales sobre investigación con participantes humanos. Todos los sujetos firmarán un consentimiento informado en el que se garantice:

- La confidencialidad de los datos recogidos.
- El uso exclusivo de la información con fines académicos y de investigación.
- La posibilidad de abandonar el estudio en cualquier momento, sin consecuencias.

Asimismo, los datos personales serán anonimizados y almacenados de forma segura, cumpliendo con la normativa de protección de datos vigente.

3.4. Objetos e instrumentación

Los objetos experimentales serán dos versiones interactivas de una misma aplicación web, denominadas versión A y versión B y que presentan en su diseño visual y disposición de los elementos de la interfaz idénticos, así pues sólo se estudiará la aplicación de las metodologías. Ambas versiones mantienen la misma complejidad funcional, permitiendo que las diferencias observadas en los resultados se deban exclusivamente al método de evaluación empleado y no a variaciones en la dificultad de las tareas. Estas versiones servirán como base para la comparación de los dos métodos de evaluación de usabilidad propuestos en este experimento.

Durante la aplicación del método A/B Testing, los participantes interactuarán individualmente con ambas versiones del prototipo, realizando un conjunto de tareas representativas de uso cotidiano, como localizar un producto, completar una compra simulada o modificar una configuración en el perfil de usuario y buscar información específica dentro del sistema. En el caso del Recorrido Pluralista, los participantes se organizarán en pequeños grupos junto con un diseñador gráfico y un experto en usabilidad, aplicando el método sobre el mismo conjunto de tareas de forma colaborativa sobre el mismo conjunto de tareas. A lo largo de estas sesiones, se analizarán colectivamente los pasos de interacción, se discutirán los problemas detectados y se propondrán posibles mejoras en el diseño. Las tareas han sido seleccionadas por su relevancia y que sean comparables entre ambos métodos sin generar efectos de fatiga o aprendizaje.

Para la recogida de datos se emplearán distintos instrumentos de medición que permitirán obtener información tanto cuantitativa como cualitativa:

- **Registro del tiempo empleado.** El tiempo empleado en la ejecución de las tareas se registrará mediante un cronómetro o software de registro automático, lo que permitirá calcular la eficiencia de cada participante en ambos métodos.
- **Documentación de problemas de usabilidad.** Los problemas de usabilidad detectados se documentarán en una plantilla estructurada en la que constará su descripción, severidad y frecuencia de aparición.
- **Cuestionario SUS.** Una vez finalizada cada sesión experimental, los participantes completarán el cuestionario, compuesto de diez preguntas con una escala de respuesta de 1 a 5, que permitirá evaluar de manera estandarizada la percepción de usabilidad del sistema.
- **Escala de satisfacción.** Además, se incluirá una escala de satisfacción general de tipo Likert(1–7) que permitirá valorar el grado de satisfacción del participante con el método aplicado.
- **Preguntas abiertas.** Finalmente, se plantearán preguntas abiertas orientadas a recoger impresiones cualitativas sobre la experiencia, tales como la percepción de facilidad de uso, las ventajas y limitaciones de cada método o las dificultades encontradas durante la evaluación.

Antes de comenzar con las sesiones experimentales, se administrará un breve cuestionario demográfico con el fin de recoger información sobre el perfil de los participantes, incluyendo edad, formación, experiencia previa en desarrollo de software y familiaridad con técnicas de evaluación de usabilidad. Durante las sesiones del Recorrido Pluralista, se realizarán además observaciones directas y se registrarán comentarios verbales de los participantes, con el objetivo de complementar los datos cuantitativos con información cualitativa sobre las percepciones, actitudes y dinámicas de grupo. Todos los datos recopilados serán tratados de forma confidencial y anonimizados, garantizando la protección de la información personal y el cumplimiento de los principios éticos establecidos para investigaciones con participantes humanos.

3.5. Evaluación de la validez

En este apartado se evalúa la validez del diseño experimental propuesto, identificando posibles amenazas y las medidas adoptadas para solucionarlas, con el fin de garantizar la robustez y fiabilidad de los resultados.

Validez interna: Se han identificado varias amenazas que podrían afectar a la relación causal entre el método de evaluación y los resultados:

- **Efecto del orden:** Dado que los desarrolladores aplican ambos métodos, se empleará un diseño contrabalanceado para controlar el efecto del orden de aplicación.

- **Efecto aprendizaje:** La exposición previa a un método podría influir en el desempeño. Para minimizarlo, se utilizarán prototipos y tareas equivalentes pero distintas en cada método, y se garantizará que los grupos de usuarios sean independientes.

- **Fatiga:** La sesiones de evaluación estarán separadas por al menos 24 horas para reducir el cansancio de los participantes.

- **Experiencia previa:** Se recogerán datos demográficos y de experiencia mediante un cuestionario inicial, y se formarán grupos homogéneos en cuanto a conocimientos en usabilidad.

Validez de conclusión estadística: Para asegurar que las diferencias observadas en los errores detectados y la satisfacción de los desarrolladores se deben a los métodos de evaluación comparados y no a factores aleatorios, se utilizarán pruebas estadísticas adecuadas. En concreto, se aplicarán pruebas t pareadas para comparar los resultados intra-sujetos (satisfacción de desarrolladores) y pruebas t independientes para los datos entre grupos (errores de usuarios). En caso de que no se cumpla el supuesto de normalidad, se recurrirá a pruebas no paramétricas equivalentes (Wilcoxon y Mann-Whitney respectivamente). Se adoptarán un nivel de significación de $\alpha = 0{,}05$ y se calcularán tamaños del efecto (Cohen's d) para cuantificar la magnitud de las diferencias encontradas.

Validez externa: El contexto de experimento permite generalizar los resultados a entornos similares de evaluación de usabilidad en en ámbito wev con evaluadores semi-expertos. Sin embargo, la selección intencionada de participantes y el uso de un único tipo de interfaz pueden limitar la generalización a otros dominios o perfiles de usuario. Para futuras réplicas, se recomienda ampliar la variedad de prototipos y perfiles de evaluadores.

Validez de constructo: Se han seleccionado instrumentos de medición para operacionalizar las variables de estudio:

- **Errores detectados:** La variable errores se medirá de forma objetiva mediante la fórmula definida en la ecuación 3.1, asegurando una evaluación cuantitativa y comparable entre métodos.

- **Satisfacción y percepción de usabilidad:** La satisfacción y percepción de usabilidad se medirán con el cuestionario SUS y una escala Likert de 7 puntos, ambos con alta fiabilidad y validez contrastada en estudios de usabilidad.

- **Preguntas abiertas:** Permitirán capturar aspectos cualitativos que enriquezcan la interpretación de los resultados.

Bibliografía

[1] Jakob Nielsen. «Usability inspection methods». En: *Conference companion on Human factors in computing systems*. 1994, págs. 413-414.

[2] J. Rubin, D. Chisnell y J. Spool. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. Wiley, 2011. ISBN: 9781118080405. URL: https://books.google.es/books?id=l_e1MmVzMb0C.

[3] Ron Kohavi et al. «Controlled experiments on the web: survey and practical guide». En: *Data mining and knowledge discovery* 18.1 (2009), págs. 140-181.

[4] Randolph G. Bias. «The pluralistic usability walkthrough: coordinated empathies». En: *Usability Inspection Methods*. USA: John Wiley & Sons, Inc., 1994, págs. 63-76. ISBN: 0471018775.

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.