Trabajo 1: Diseño de un experimento controlado

Daniel Jiménez García
Sergio Muñoz Gómez
Ángela Caiqing Pousada Morán

Fecha: 2 de noviembre de 2025

Master MITSS. ISE
Universidad Politécnica de Valencia

Resumen

Este trabajo plantea un experimento para comparar dos métodos de evaluación de usabilidad: el *A/B Testing*, un enfoque más cuantitativo basado en datos, y el *Recorrido Pluralista*, un método de revisión en grupo donde participan perfiles distintos para analizar un diseño. La idea surge porque no hay demasiada evidencia que demuestre cuál de los dos funciona mejor a la hora de detectar problemas de usabilidad ni cuál resulta más cómodo o útil para los desarrolladores.

El estudio se divide en dos fases. En la primera, varios equipos de desarrolladores aplican ambos métodos con usuarios reales que prueban dos versiones diferentes de una aplicación web. En esta fase se recogen datos más "objetivos", como los errores cometidos o el tiempo que tardan en completar tareas representativas.

La segunda fase es el núcleo del trabajo, y se centra en comparar qué método les ha resultado más eficaz a los desarrolladores para identificar problemas de usabilidad y cuál les ha gustado más utilizar.

El experimento combina dos enfoques metodológicos: un diseño *between-subjects* para los usuarios de la primera fase (cada usuario solo prueba una versión) y un diseño *within-subjects* para los desarrolladores, que utilizan ambos métodos.

Se plantean dos hipótesis nulas relacionadas con la percepción de errores y la preferencia entre métodos. Las variables dependientes incluyen el número de problemas detectados y el nivel de satisfacción de los desarrolladores, medido mediante el cuestionario SUS, una escala de Likert y preguntas abiertas.

Capítulo 1

Motivación

1.1. Problema a investigar

Actualmente existen diversos métodos a la hora de evaluar la experiencia de usuario a la hora de utilizar 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

El experimento se realiza de manera presencial, ya que en una primera parte se evaluará la interfaz gráfica y después las sensaciones de los desarrolladores aplicando los métodos. Para ello, primero se partirá de un grupo de desarrolladores que será dividio en dos y cada uno de los grupos aplicará cada método de evaluación en 2 grupos independientes de usuarios. El fin del experimento es estudiar cada metodología para comprobar donde es más eficaz el uso de cada una.

1.4. Implicaciones

El experimento tiene implicaciones a nivel académico como profesional dentro del ámbito del desarrollo de software. Desde una perspectiva científica, los resultados pueden llegar a contribuir a una mejor comprensión comparativa entre el **A/B Testing** y el **Recorrido Pluralista**, dos métodos de evaluación de usabilidad ampliamente utilizados. Este conocimiento permitirá orientar futuras investigaciones sobre su efectividad, esfuerzo requerido y aplicabilidad en diferentes contextos de diseño.

En el plano profesional, los hallazgos podrán guiar a los equipos de desarrollo en la selección del método más adecuado según los objetivos y recursos del proyecto. Si el *A/B Testing* demuestra ser más eficiente, podría recomendarse en entornos de validación rápida; mientras que el *Recorrido Pluralista*, al fomentar la discusión colaborativa, resultaría más apropiado para fases de diseño conceptual o evaluación cualitativa.

Por último, el estudio promueve la incorporación sistemática de la evaluación de usabilidad en el proceso de ingeniería del software, impulsando una cultura de diseño centrada en el usuario. Los beneficios potenciales incluyen una reducción de errores, mayor satisfacción de los usuarios finales y una optimización de los costes asociados al rediseño de interfaces.

Capítulo 2

Trabajos relacionados

Diversos estudios 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, en este experimento se parte de un grupo de desarrolladores que serán divididos en dos grupos. A estos desarroladores se les realizará una encuesta demográfica de carácter informativo acerca de la población a estudiar. Una vez recogida la información sobre sus características, se procederá a dividir el grupo en 2. Cada uno aplicará las dos metodoloías en 2 grupos de participanetes independientes. Los participantes sólo serán escogidos una vez y no participarán en el otro experimento de apliacación de la metodología realizado por los desarrolladores para que no afecte en el estudio. Como se muestra en la tabla 3.1 cada equipo desarrollador evaluará a los 2 grupos de usuarios (G1, G2, G3 y G4) independientes para no producir sesgos en el experimento fruto del aprendizaje del funcionamiento de la interfaz. Este enfoque del experimento es contrabalanceado para estudiar ambas metodologías en ambos grupos de desarrolladores.

Grupo	**A/B Testing**	**Recorrido Pluralista**
Desarrolladores 1	G1	G2
Desarrolladores 2	G3	G4

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

Las tareas que deberán realizar los usuarios incluyen actividades representativas de un uso real de la aplicación. Para el A/B Testing cada desarrollador evaluará 2 versiones de la interfaz gráfica en los usuarios que tendrán que realizar diversas tareas relacionada con la funcionalidad de la aplicación. En cuanto al estudio del Recorrido pluralista el segundo grupo de desarrollador realizará una entrevista general con el grupo de participantes e irán discutiendo con ellos aspectos de la interfaz mientras ejecutan tareas que harían con la interfaz web.

Una vez finalizada esa evaluación de la interfaz, se dará paso a comparar ambos métodos donde se estudiarán las encuestas SUS hecha por los desarrolladores y se encuestará a estos mismos para determinar la eficiencia y rendimiento percibido utilizando cada uno de las metodologías. Además se tendrá en cuenta la cantidad de errores encontrados en los usuarios en la aplicación de cada una de las metodologías.

3.3. Selección de sujetos

El experimento requiere la participación de dos grupos distintos de participantes: desarrolladores evaluadores y usuarios finales. Cada grupo cumple un rol específico en el diseño experimental y cuenta con sus propios criterios de selección.

La selección de los participantes para desarrolladores evaluadores 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, 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.

Para la selección de los usuarios finales, se empleará un muestreo aleatorio estratificado, asegurando la inclusión de individuos con diferentes niveles de experiencia en el uso de aplicaciones web. Se buscará reclutar entre 40 y 50 usuarios, divididos en dos grupos independientes que participarán en las evaluaciones realizadas por los desarrolladores. Este tamaño muestral permitirá obtener resultados representativos y generalizables sobre la efectividad de los métodos evaluados.

Criterios de inclusión

Criterios para desarrolladores evaluadores:

- 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 para usuarios finales:

- Contar con habilidades básicas de navegación web y uso de interfaces interactivas.
- No haber participado previamente en estudios similares que involucren los mismos métodos de evaluación.
- Aceptar voluntariamente participar en el estudio, firmando el consentimiento informado.

Criterios de exclusión

Criterios para desarrolladores evaluadores:

- 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.

Criterios para usuarios finales:

- Tener experiencia previa significativa con los prototipos empleados, lo que podría influir en la percepción de usabilidad.
- Presentar discapacidades visuales, motoras o cognitivas que dificulten la interacción con las interfaces web.
- Haber participado en estudios similares en los últimos seis meses, para evitar efectos de aprendizaje o familiaridad con los métodos evaluados.

Procedimiento de reclutamiento

- **Desarrolladores evaluadores:**

  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.

- **Usuarios finales:**

  Los usuarios serán reclutados mediante muestreo por conveniencia en entornos académicos y profesionales, asegurando la diversidad en género, edad y nivel de experiencia tecnológica. Se utilizarán anuncios en redes sociales, mailing lists y carteles en espacios públicos.

Como se comentó en el anterior apartado, 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 estarán constituidos por una aplicación web interactiva desarrollada en dos versiones diferenciadas, denominadas versión A y versión B, que presentan las mismas funcionalidades y flujo de tareas, pero difieren en determinados elementos visuales o de disposición. Estas variaciones se introducen con el fin de posibilitar la aplicación del método A/B Testing, que requiere comparar la interacción de los usuarios con dos variantes del mismo sistema. En cambio, para la aplicación del Recorrido Pluralista se empleará una única versión de la aplicación (la versión B), sobre la que se realizará un análisis colaborativo de la interacción.

El objeto principal del experimento es evaluar la percepción de los desarrolladores tras aplicar ambos métodos de evaluación de usabilidad. Los desarrolladores no son los sujetos que ejecutan las tareas, sino los evaluadores que orquestan y supervisan las pruebas con grupos de usuarios reales, aplicando cada metodología y observando los resultados obtenidos.

De esta forma, el estudio busca analizar cómo valoran los desarrolladores la utilidad, facilidad de aplicación, carga de trabajo y eficacia percibida de cada método una vez que han tenido la oportunidad de ponerlos en práctica.

Cada desarrollador aplicará ambos métodos de manera secuencial y contrabalanceada de modo que algunos comiencen con A/B Testing y otros con Recorrido Pluralista con el objetivo de evitar sesgos derivados del orden o de la experiencia acumulada. Después de la aplicación de ambos métodos, los desarrolladores completarán un cuestionario de valoración comparativa, expresando su opinión y grado de satisfacción con cada técnica.

Durante la aplicación del método A/B Testing, cada desarrollador coordinará sesiones individuales en las que los usuarios interactuarán con las versiones A y B del prototipo. Se registrarán métricas como tiempo de ejecución, tasa de éxito y errores cometidos por los usuarios, además de observaciones sobre los problemas detectados.

En el caso del Recorrido Pluralista, el desarrollador dirigirá sesiones grupales en las que los usuarios trabajarán junto con un diseñador gráfico y un experto en usabilidad. Durante estas sesiones se recorrerán de forma colaborativa las tareas de

la aplicación, discutiendo los pasos de interacción, los problemas encontrados y las posibles mejoras de diseño.

Una vez completadas ambas fases, los desarrolladores valorarán su experiencia como evaluadores, indicando qué método consideran más eficaz, intuitivo y práctico para la identificación de problemas de usabilidad y para su aplicación en contextos reales de desarrollo.

Para la obtención de información se emplearán instrumentos orientados a recoger datos cuantitativos y cualitativos, tanto de los usuarios como de los desarrolladores:

- Registro del tiempo empleado por los usuarios en la ejecución de las tareas, medido mediante software o cronómetro digital, con el fin de evaluar la eficiencia del método.
- Documentación de problemas de usabilidad, recogida por los desarrolladores en una plantilla estructurada con la descripción, severidad y frecuencia de cada incidencia.
- Cuestionario SUS, administrado a los usuarios tras la interacción con la aplicación, para evaluar la percepción de usabilidad de las versiones A y B.
- Cuestionario de valoración del método, cumplimentado por los desarrolladores al finalizar la aplicación de ambas metodologías, diseñado específicamente para medir su satisfacción, facilidad de aplicación y preferencia entre los métodos.
- Escala de satisfacción general tipo Likert (1–7), para cuantificar la valoración global de los desarrolladores sobre la experiencia de uso de cada técnica.
- Preguntas abiertas, destinadas a recoger percepciones cualitativas sobre las ventajas, limitaciones y dificultades encontradas durante la aplicación de los métodos.

Antes del inicio del experimento, los desarrolladores y usuarios completarán un cuestionario demográfico donde se recogerán datos relativos a edad, formación, experiencia en desarrollo de software y familiaridad con técnicas de evaluación de usabilidad. Durante las sesiones del Recorrido Pluralista, se realizarán observaciones directas y registro de comentarios verbales por parte de los desarrolladores, con el fin de complementar los datos cuantitativos con información cualitativa sobre las percepciones 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

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.

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: Cuestionario para usuarios

Instrucciones: Rellene el cuestionario tras completar las tareas con la aplicación.

- Escala SUS (System Usability Scale) — 10 ítems, respuestas 1 (totalmente en desacuerdo) a 5 (totalmente de acuerdo).
- Preguntas de rendimiento percibido:
  - ¿Pudo completar las tareas propuestas? (Sí/No)
  - Tiempo estimado para completar cada tarea (segundos/minutos).
  - Observaciones sobre dificultades encontradas (respuesta abierta).
- Comentarios abiertos sobre la experiencia de uso: ¿qué aspectos le resultaron más confusos o útiles?

Apéndice B

Anexo II: Cuestionario para desarrolladores

El cuestionario se realizó en el siguiente enlace.

Instrucciones: Complete este cuestionario después de haber aplicado ambos métodos de evaluación de usabilidad (*A/B Testing* y *Recorrido Pluralista*) con los grupos de usuarios asignados. Las respuestas deben reflejar su experiencia como desarrollador al coordinar y aplicar cada metodología.

- **Sección 1 — Facilidad de aplicación (escala Likert 1 = totalmente en desacuerdo, 5 = totalmente de acuerdo):**
  - Entendí rápidamente cómo aplicar el método A/B Testing.
  - Entendí rápidamente cómo aplicar el método de Recorrido Pluralista.
  - La documentación o guía del método A/B Testing fue clara y suficiente.
  - La dinámica del Recorrido Pluralista fue fácil de seguir.
  - Me resultó sencillo identificar pasos concretos al aplicar A/B Testing.
  - Me resultó sencillo coordinar y ejecutar las sesiones del Recorrido Pluralista.
  - En general, considero que el método [A/B Testing / Recorrido Pluralista] es fácil de aplicar.
- **Sección 2 — Eficiencia percibida (escala Likert 1 = totalmente en desacuerdo, 5 = totalmente de acuerdo):**
  - El tiempo necesario para preparar un A/B Testing es razonable.
  - El tiempo necesario para preparar un Recorrido Pluralista es razonable.
  - Aplicar A/B Testing requiere menos esfuerzo que otros métodos de evaluación.
  - El Recorrido Pluralista requiere demasiado esfuerzo de coordinación.
  - El análisis de resultados en A/B Testing me resultó rápido y claro.
  - El análisis de resultados del Recorrido Pluralista fue sencillo de interpretar.

- **Sección 3 — Utilidad y efectividad del método (escala Likert 1 = totalmente en desacuerdo, 5 = totalmente de acuerdo):**
  - A/B Testing permite detectar problemas de usabilidad relevantes.
  - El Recorrido Pluralista permite identificar problemas de usabilidad relevantes.
  - A/B Testing ofrece evidencia objetiva sobre la experiencia del usuario.
  - El Recorrido Pluralista permite comprender mejor el razonamiento del usuario.
  - Los resultados obtenidos con A/B Testing son fácilmente comunicables al equipo de desarrollo.
  - Los resultados del Recorrido Pluralista son útiles para mejorar el diseño del sistema.
  - En mi experiencia, ambos métodos complementan bien el proceso de evaluación de usabilidad.
- **Sección 4 — Satisfacción general y preferencia (escala Likert 1 = totalmente en desacuerdo, 5 = totalmente de acuerdo):**
  - Me sentí cómodo utilizando el método A/B Testing.
  - Me sentí cómodo utilizando el método Recorrido Pluralista.
  - Me resultó más interesante aplicar A/B Testing que el Recorrido Pluralista.
  - Considero que el Recorrido Pluralista fomenta mejor la colaboración entre los evaluadores.
  - Preferiría usar A/B Testing en futuros proyectos de evaluación.
  - Preferiría usar Recorrido Pluralista en futuros proyectos de evaluación.
  - En general, estoy satisfecho con la calidad de los resultados obtenidos con ambos métodos.
- **Sección 5 — Preguntas abiertas (análisis cualitativo):**
  - ¿Qué ventajas destacarías del método A/B Testing en comparación con el Recorrido Pluralista?
  - ¿Qué limitaciones o dificultades encontraste al aplicar cada método?