Evaluación Empírica de los Mecanismos de Usabilidad: Efectos sobre la Eficiencia y la Eficacia del Usuario

Juan M. Ferreira
Facultad Politécnica, Universidad Nacional de Asunción, CC 1439, San Lorenzo, Paraguay
jmferreira1978@gmail.com

y

Silvia T. Acuña
[2] Universidad Autónoma de Madrid, Calle Tomás y Valiente 11, 28049 Madrid, España
silvia.acunna@uam.es

RESUMEN

Contexto. La usabilidad como atributo de calidad comprende aspectos de eficiencia, eficacia y satisfacción del usuario. Existen varias recomendaciones en la literatura para construir sistemas software usables pero hay escasez de estudios empíricos que proporcionen evidencias de ello. **Objetivo**. En este trabajo se presenta un experimento verdadero con usuarios no informáticos para obtener evidencias del efecto de los mecanismos de usabilidad *Abort Operation* (MU-ABR), *Progress Feedback* (MU-PFB) y *Preferences* (MU-PRF) sobre la eficiencia y eficacia de los usuarios. **Método**. Se realiza un diseño experimental donde se asignan aleatoriamente los sujetos a uno de los 24 grupos establecidos para la realización de una serie de tareas. Se recolectan datos de tiempo, número de clicks y porcentaje de tarea realizada para analizar la eficiencia y la eficacia. **Resultados**. El estudio revela una tendencia a favor de la eficiencia y eficacia del usuario. **Conclusiones**. La presencia de determinados mecanismos de usabilidad evidencian mejoras significativas en la eficiencia y eficacia del usuario. Además, el estudio parece revelar que la mejora en la eficiencia y eficacia depende del dominio del problema. En este sentido, un mecanismo podría tener un efecto distinto en otros contextos, por lo que se necesitan experimentos adicionales para reunir más evidencias y confirmar estos resultados.

Palabras Claves: Usabilidad, mecanismo de usabilidad, estudio empírico, diseño experimental, eficiencia, eficacia, ingeniería del software.

1. INTRODUCCIÓN

En la actualidad, la construcción de sistemas software de calidad es uno de los principales desafíos de la Ingeniería del Software (IS). La calidad es una propiedad inherente de cualquier sistema software que se descompone en diversas características, entre ellas la usabilidad [1]. La usabilidad representa un atributo crítico en sistemas altamente interactivos [2] y va encabezando la lista de factores críticos para el éxito de los sistemas software.

El estándar ISO 25010 [1] define a la usabilidad como la medida en que un producto puede ser usado por usuarios específicos para alcanzar determinados objetivos con eficacia, eficiencia y satisfacción en un contexto específico de uso, sin efectos adversos. En este sentido, con esta definición es posible medir la usabilidad percibida en términos de la eficiencia, eficacia y satisfacción. Además, en la literatura existen varias recomendaciones para alcanzar niveles deseables de usabilidad en los sistemas software [2], [3], sin embargo, hay insuficiencia de estudios empíricos que proporcionen evidencias más allá de la teoría con respecto a las mejoras en usabilidad percibidas por los usuarios.

Tomando la definición del estándar ISO 25010 [1] y las recomendaciones de usabilidad con impacto en el diseño [2], esta investigación trata sobre la evaluación de la usabilidad haciendo uso de la experimentación en IS, donde se busca obtener evidencia empírica sobre el impacto de dotar de usabilidad a un sistema software desde la perspectiva de los usuarios. Concretamente, aborda el estudio empírico del efecto de la presencia/ausencia de los mecanismos de usabilidad *Abort Operation* (MU-ABR), *Progress Feedback* (MU-PFB) y *Preferences* (MU-PRF) en una aplicación web bajo los atributos de eficiencia y eficacia [1], [3].

El resto del artículo se estructura de la siguiente manera. En la sección 2 se describen algunos trabajos relacionados a esta investigación. En la sección 3 se presenta el experimento verdadero junto con el diseño experimental y su ejecución. El análisis de datos y la discusión de resultados se describen en la secciones 4 y 5. En la sección 6 se expone la validez del experimento y finalmente, las conclusiones y líneas futuras se describen en la sección 7.

2. TRABAJOS RELACIONADOS

La usabilidad es un factor de calidad del software que tiene como objetivo proporcionar la respuesta a muchos problemas encontrados en la interacción entre las personas y la tecnología. Al incorporar la usabilidad en aplicaciones web, denominada usabilidad web, es necesario refinar las definiciones generales para capturar la especificidad de esta clase de aplicaciones. En la web, las principales tareas que se realizan son: búsqueda directa de información deseada y servicios, descubrimiento de información a través de la navegación y comprensión de la información presentada. Por tanto, parafraseando la definición de ISO 25010 [1] de la usabilidad, la usabilidad web puede ser considerada como la capacidad de las aplicaciones web para apoyar este tipo de tareas con eficacia, eficiencia y satisfacción.

Considerando estudios experimentales en IS, algunas iniciativas están empezando a considerar la usabilidad desde una perspectiva empírica atendiendo alguno de los atributos de calidad. Sin embargo, estos tipos de estudios empíricos son muy pocos.

Por ejemplo, en la investigación de Panach [4], [5] se ha propuesto la incorporación de propiedades de usabilidad en un

sistema software siguiendo un método de desarrollo dirigido por modelos (*Model-Driven Development* MDD) con el fin de evitar que los sistemas generados por el MDD sean modificados manualmente para incluir características de usabilidad. Panach realiza una validación experimental para comprobar si la usabilidad de las aplicaciones generadas con el nuevo método mejora o no tras el uso del software por determinados usuarios.

Otra investigación intenta obtener los beneficios de la usabilidad de manera empírica [6] estudiando el impacto de ciertas características de usabilidad en sistemas software desarrollados para un determinado contexto. El diseño experimental considera también una aplicación web de juguete e intenta evaluar los atributos de eficiencia y satisfacción.

3. EXPERIMENTO VERDADERO

Este apartado incluye la caracterización del experimento realizado, describiendo los objetivos, las variables, las hipótesis, los sujetos y los instrumentos. Así mismo, se detalla el proceso experimental correspondiente a un experimento verdadero.

Objetivo, Pregunta e Hipótesis de Investigación
El objetivo de la investigación, basado en la estructura Goal Question Metric [7], es:

- **Estudiar** empíricamente los mecanismos de usabilidad (MU-ABR, MU-PFB y MU-PRF) en una aplicación web **con el propósito de** evaluar el impacto de la usabilidad **con respecto a** la eficiencia y eficacia **desde la perspectiva de** los usuarios **en el contexto de** usuarios no informáticos.

Conforme al objetivo, el experimento pretende resolver la siguiente pregunta de investigación:

¿La presencia de los mecanismos de usabilidad impacta en la usabilidad de la aplicación?

Más concretamente, se desea responder a las siguientes tres preguntas de investigación:

1. ¿La presencia del mecanismo de usabilidad MU-ABR impacta en la usabilidad de la aplicación?
2. ¿La presencia del mecanismo de usabilidad MU-PFB impacta en la usabilidad de la aplicación?
3. ¿La presencia del mecanismo de usabilidad MU-PRF impacta en la usabilidad de la aplicación?

De la pregunta de investigación, la hipótesis general es:

- H_0 (nula): No existe diferencia significativa en la EFICIENCIA | EFICACIA del usuario al incorporar el MU-ABR | MU-PFB | MU-PRF o al no incorporarlo.

- H_1 (alternativa): Existe diferencia significativa en la EFICIENCIA | EFICACIA del usuario al incorporar el MU-ABR | MU-PFB | MU-PRF o al no incorporarlo.

Concretamente, la hipótesis alternativa intenta revelar que cuando el mecanismo de usabilidad (MU-ABR | MU-PFB | MU-PRF) está presente, se obtienen mejoras en la Eficiencia y Eficacia del usuario.

Como cada mecanismo de usabilidad se evalúa frente a cada atributo de calidad, las hipótesis específicas a ser comprobadas son:

1. MU-ABR
    1.1. EFICIENCIA
        - H.1.1.0 (nula) No existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-ABR o al no incorporarlo.
        - H.1.1.1 (alternativa) Existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-ABR o al no incorporarlo.
    1.2. EFICACIA
        - H.1.2.0 (nula) No existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-ABR o al no incorporarlo.
        - H.1.2.1 (alternativa) Existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-ABR o al no incorporarlo.
2. MU-PFB
    2.1. EFICIENCIA
        - H.2.1.0 (nula) No existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-PFB o al no incorporarlo.
        - H.2.1.1 (alternativa) Existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-PFB o al no incorporarlo.
    2.2. EFICACIA
        - H.2.2.0 (nula) No existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-PFB o al no incorporarlo.
        - H.2.2.1 (alternativa) Existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-PFB o al no incorporarlo.
3. MU-PRF
    3.1. EFICIENCIA
        - H.3.1.0 (nula) No existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-PRF o al no incorporarlo.
        - H.3.1.1 (alternativa) Existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-PRF o al no incorporarlo.
    3.2. EFICACIA
        - H.3.2.0 (nula) No existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-PRF o al no incorporarlo.
        - H.3.2.1 (alternativa) Existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-PRF o al no incorporarlo.

Diseño Experimental
El experimento se diseña siguiendo el proceso experimental propuesto por Wohlin [8]. El diseño del experimento verdadero corresponde a un *Between Subjects (BS)* [9]. En un diseño *BS*, cada sujeto experimental está asignado únicamente a un grupo que corresponde a una combinación de los niveles de los factores. Concretamente, cada sujeto experimental proporciona un único valor para cada variable respuesta que será utilizado en el análisis estadístico. En el caso de este experimento, cada sujeto participa en varios experimentos ejecutados consecutivamente (uno por cada mecanismo de usabilidad

considerado), contrabalanceando el orden de ejecución de las tareas (una por mecanismo) para anular posibles efectos de presentación de las tareas.

Este diseño está caracterizado por tres matrices: tratamientos, orden de exposición y asignación de grupos. En la Tabla 1 se presenta la matriz de tratamientos (los ceros indican que el mecanismo de usabilidad no está presente mientras que, por el contrario, los unos indican que el mecanismo sí está presente).

Tabla 1: Matriz de tratamientos

Tratamiento	ABR	PFB	PRF
A	0	0	1
B	1	0	0
C	0	1	0
D	1	1	1

La segunda matriz es la de órdenes de exposición, donde se establecen las secuencias posibles de realización de las tareas asociadas a cada mecanismo de usabilidad (ver Tabla 2).

Tabla 2: Matriz de órdenes de exposición de cada MU

Nro.	Primero	Segundo	Tercero
O1	ABR	PFB	PRF
O2	ABR	PRF	PFB
O3	PFB	PRF	ABR
O4	PFB	ABR	PRF
O5	PRF	ABR	PFB
O6	PRF	PFB	ABR

Finalmente, el diseño se completa con la matriz de asignación de grupos consistente en un producto cartesiano entre la matriz de tratamientos y la matriz de órdenes. Como se muestra en la Tabla 4, en total se tienen 24 grupos elegibles para un sujeto experimental.

Instrumentos, Métricas y Sujetos Experimentales
El instrumento de recolección de datos es una aplicación web de tienda en línea denominada *QuickStore* junto con la interfaz del experimento [10], [11]. Las tareas solicitadas a cada sujeto se detallan en la Tabla 3. Además, se solicita al usuario completar un cuestionario de familiaridad.

Tabla 3: Descripción de tareas solicitadas al sujeto

Tarea	Descripción
MU-ABR	Se pide al sujeto ingresar a su carrito de compra y modificar ciertos datos y, seguidamente, cancelar la operación.
MU-PFB	Se pide al sujeto realizar la búsqueda de un producto en la tienda.
MU-PRF	Esta tarea está dividida en dos partes: Tarea Básica: Se pide al sujeto que intente personalizar la interfaz de la aplicación. Tarea Ficticia: Luego de la tarea básica, se solicita al sujeto buscar cierta información de

Tabla 4: Matriz de asignación de grupos con órdenes de exposición de cada MU

			Primero			Segundo			Tercero		
			ABR	PFB	PRF	ABR	PFB	PRF	ABR	PFB	PRF
A	O1	G1	0				0				1
	O2	G2	0					1		0	
	O3	G3		0				1	0		
	O4	G4		0		0					1
	O5	G5			1	0				0	
	O6	G6			1		0		0		
B	O1	G7	1				0				0
	O2	G8	1					0		0	
	O3	G9		0				0	1		
	O4	G10		0		1					0
	O5	G11			0	1				0	
	O6	G12			0		0		1		
C	O1	G13	0				1				0
	O2	G14	0					0		1	
	O3	G15		1				0	0		
	O4	G16		1		0					0
	O5	G17			0	0				1	
	O6	G18			0		1		0		
D	O1	G19	1				1				1
	O2	G20	1					1		1	
	O3	G21		1				1	1		
	O4	G22		1		1					1
	O5	G23			1	1				1	
	O6	G24			1		1		1		

Se recolectan las métricas en cada tarea, excepto en la tarea básica del MU-PRF porque en ella no se mide el efecto del mecanismo en la eficiencia y eficacia del usuario. Las métricas recolectadas son:

- Eficiencia:
    - Rapidez: el tiempo que un sujeto necesita para completar la tarea [12], [13], medido en segundos.
    - Grado de Interacción: número de clicks que un sujeto necesita para completar la tarea [14], [15].

- Eficacia: porcentaje de tarea resuelta por un sujeto [16].

Téngase en cuenta que el tiempo recolectado por cada sujeto representa el tiempo de lectura e interpretación de la tarea solicitada al sujeto más el tiempo invertido en su realización.

En cuanto a los participantes, o sujetos experimentales, ellos están representados por usuarios finales no relacionados a la informática, es decir, personas pertenecientes a otras áreas del conocimiento distintas a las ciencias de la computación. En este sentido, el experimento ha sido ejecutado en dos contextos: a) académico: con estudiantes de distintas carreras de la Universidad Autónoma de Asunción y b) no académico: con profesionales y no profesionales a través de una carta de invitación.

Operación, Recolección y Validación de Datos
Al momento de la ejecución experimental, los sujetos no tenían conocimiento del objetivo del estudio ni de las hipótesis de la investigación. No se ha mencionado la palabra "Experimento" para evitar cualquier efecto negativo de relacionarlo con ciertos roedores utilizados en laboratorios de biología. Por el contrario, se ha referido al experimento como "Evaluación". También, se ha informado que los resultados de dicha evaluación servirá para mejorar ciertos aspectos en las aplicaciones webs y se garantiza la confidencialidad de los mismos.

Finalmente, se ha notificado que la participación es voluntaria y en caso de que el sujeto acepte se le anima a invertir su mejor esfuerzo en la realización de las tareas. No se ha proporcionado ningún material adicional más que el enlace web en la cual cada participante debe acceder para iniciar la evaluación.

El experimento se ejecutó durante 5 meses, entre marzo y julio del año 2016. Los primeros 4 meses han sido ejecutados en el ámbito académico de la Universidad Autónoma de Asunción utilizando la plataforma de educación a distancia. La ejecución fuera del contexto académico ha sido ejecutado en el último mes. La ejecución llevada a cabo en la Universidad no ha interferido con los objetivos del curso de cada carrera ya que la ejecución del experimento ha sido propuesta como una tarea desafío opcional a la que los estudiantes acceden libre y voluntariamente.

En principio se han recolectado datos para un total de 182 sujetos. Sin embargo se han eliminado los datos de 14 sujetos porque no han completado correctamente las tareas. Concretamente, no se han tenido en cuenta debido a:

- Ausencia de datos de métricas de 2 sujetos.
- 3 sujetos han repetido el experimento.
- 3 sujetos eran perfil informático.
- Datos de 6 sujetos que no han terminado el escenario completo. Es decir, iniciaron con la primera tarea, incluso la segunda, sin completar las siguientes. Recuerde que un escenario es la composición de tres tareas principales como se ha descrito anteriormente y se requiere que las tareas sean realizadas completamente.

Finalmente quedaron 168 datos válidos para el análisis estadístico y la interpretación de los resultados.

4. ANÁLISIS DE DATOS Y RESULTADOS

Primeramente se presentan los estadísticos descriptivos de cada mecanismo de usabilidad con respecto a las medidas de tiempo, número de clicks y % de tarea. Posteriormente, se efectúa el contraste de hipótesis para verificar cuáles hipótesis se aceptan y cuáles no. El análisis de datos se realiza utilizando como herramienta software el lenguaje R [17] tanto para los descriptivos como para el contraste de hipótesis.

Estadísticos Descriptivos
En la Tabla 5 se muestran los estadísticos descriptivos de los 168 sujetos experimentales. Los estadísticos descriptivos que se describen son la media, mediana y desviación típica. La eficiencia se mide en tiempo y en número de clicks mientras que la eficacia en % de tarea realizada por cada sujeto. Los resultados obtenidos se comparan para los sujetos con el mecanismo presente y para los sujetos con el mecanismo ausente, denotados por los sufijos _P y _A, respectivamente. Se han construido diagramas de cajas para visualizar los datos obtenidos. Las cajas corresponden a cada variable respuesta haciendo una comparación de los grupos de sujetos que tenían el mecanismo presente contra aquellos que no lo tenían y están denotados por las etiquetas Presente y Ausente. Por ejemplo, se comparan los clicks de los sujetos quienes tenían el mecanismo presente contra los clicks realizados por los sujetos quienes no lo tenían.

Tabla 5: Estadísticos descriptivos de la eficiencia y eficacia de los MUs ABR, PFB y PRF

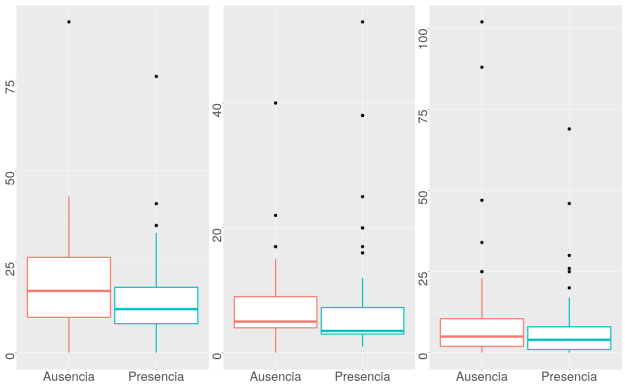
		MU-ABR			MU-PFB			MU-PRF		
Atrib.	Métrica	Media	Mediana	Dev. Est.	Media	Mediana	Dev. Est.	Media	Mediana	Dev. Est.
Eficiencia	Tiempo_P	142,6	110,2	120,1	95,8	72,2	85,5	90,2	66,8	80,9
	Tiempo_A	179,6	157,2	154,7	99,0	65,2	105,3	113,5	76,2	112,2
	Click_P	14,5	12,0	10,9	6,9	3,5	8,4	6,7	4,0	10,2
	Click_A	18,9	17,0	13,2	7,1	5,0	5,5	9,8	5,0	15,9
Eficacia	%Tarea_P	85,4	90,0	19,4	96,1	100	9,8	74,4	90	34,5
	%Tarea_A	61,4	57,5	13,6	95,7	100	11,2	38,6	20	30,8

En la Figura 1 se muestran las cajas de la Eficiencia en clicks de los tres mecanismos de usabilidad. En la misma, se puede observar una leve tendencia de mejora en los clicks cuando el mecanismo de usabilidad está presente, ya que la caja del estado presente esta desplazada más abajo que la ausente. En todas las muestras se observan los valores atípicos.

Con relación a la Eficiencia en tiempo, la Figura 2 muestra un patrón bastante similar que los clicks, con excepción del MU-PFB donde las cajas están prácticamente solapadas. Los valores atípicos se visualizan en las tres muestras.

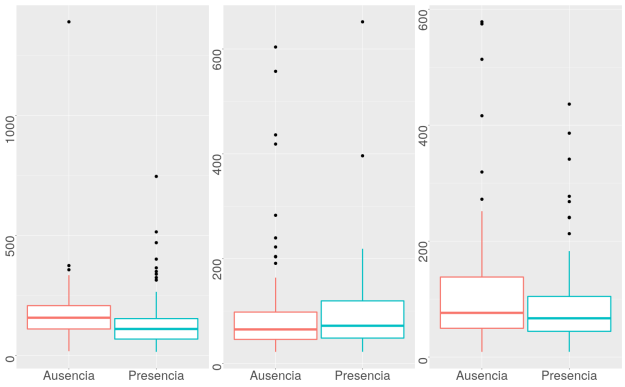
Finalmente, en la Figura 3 se muestran las cajas de la Eficacia de los tres mecanismos de usabilidad. Se observa que la caja del mecanismo presente se encuentra desplazada mas arriba que la caja del mecanismo ausente para el MU-ABR y MU-PRF. Esto da un indicativo que la presencia de estos mecanismos hace que los sujetos avancen en mayor porcentaje de la tarea propuesta. Sin embargo, para el MU-PFB no se forma cajas, es decir, hay una coincidencia casi total de los valores de la mediana y los cuartiles. Esta situación hace notar que la presencia o no de este mecanismo parece no afectar la eficacia del usuario. Concretamente, los sujetos han logrado finalizar la tarea

independientemente de la presencia de los mecanismos de usabilidad. No se observan valores atípicos para el MU-PRF.



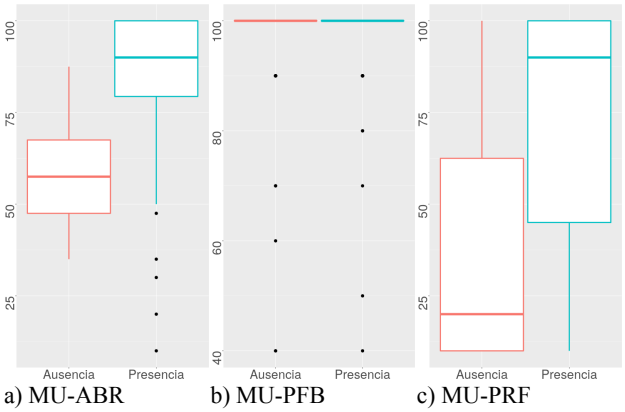
a) MU-ABR b) MU-PFB c) MU-PRF

Figura 1: Boxplot de Eficiencia (clicks)



a) MU-ABR b) MU-PFB c) MU-PRF

Figura 2: Boxplot de Eficiencia (tiempo)



a) MU-ABR b) MU-PFB c) MU-PRF

Figura 3: Boxplot de Eficacia (% tarea)

Antes de pasar a realizar el contraste de hipótesis, se hace un chequeo en busca de valores imposibles. Se consideran valores imposibles aquellos sujetos que terminan la tarea en cero segundos o con cero clicks. Estos valores imposibles se eliminan para el análisis del contraste de hipótesis. No se eliminan los valores atípicos porque se consideran como valores genuinos que reflejan la métrica que se está midiendo. Se opta por usar alguna estrategia de transformación de datos para los contrastes posteriores.

Contraste de Hipótesis

El contraste de hipótesis permite comprobar si se puede aceptar o rechazar la hipótesis nula en cada caso. Este análisis se lleva a cabo a través de un test ANOVA que exige que la muestra sea normal y homogénea. Se utiliza el test de *Kolmogorov-Smirnov (KS)* para la comprobación de Normalidad y el test de *Levene (LE)* para la comprobación de la Homogeneidad. La comprobación de los supuestos de Normalidad y Homogeneidad se incluye para cada mecanismo. Para los casos en las cuales no es posible comprobar la Normalidad o la Homogeneidad o ambos, se utiliza el equivalente no paramétrico del *ANOVA*, la prueba de *KW (Kruskal-Wallis)*. La evaluación de las hipótesis se realizan a un nivel de significación α=0.05. Concretamente, si el resultado del test de hipótesis es un valor (*p-value*) inferior a este nivel de significación quiere decir que hemos obtenido evidencia suficiente para rechazar la hipótesis nula de un determinado mecanismo de usabilidad con respecto a un atributo de calidad.

H.1. Abort Operation **(MU-ABR)**: Se desea comprobar si la presencia del mecanismo de usabilidad *Abort Operation* mejora o no la eficiencia y eficacia del usuario.

En la Tabla 6, mediante la transformación $x \rightarrow \log(x)$, se comprueban los supuestos de Normalidad y Homogeneidad para la Eficiencia. El ANOVA revela que existe un efecto significativo entre la presencia y ausencia del mecanismo MU-ABR, tanto para las medidas de tiempo y número de clicks. Entonces, se rechaza la hipótesis nula (H.1.1.0) y se afirma que "Existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-ABR o al no incorporarlo". En cuanto a la Eficacia, no se logran superar los supuestos del ANOVA incluso aplicando transformaciones. Entonces se aplica KW y se revela la existencia de una diferencia estadísticamente significativa entre la presencia y ausencia del MU-ABR por lo que se rechaza la hipótesis nula (H.1.2.0) y se acepta la hipótesis alternativa de que "Existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-ABR o al no incorporarlo".

Tabla 6: Transformaciones, Normalidad, Homogeneidad y Test de Hipótesis del MU-ABR

Atributo	Métrica	Transf.	Normal.	Homog.	Test	Pr(>F)
Eficiencia	Clicks	$\log(x)$	0.5885	0.9868	ANOVA	0.0489 *
	Tiempo	$\log(x)$	0.4587	0.9444	ANOVA	0.00712 **
Eficacia	%Tarea	Ninguna	0.002406	0.4032	KW	2.2e-16 ***

H.2. Progress Feedback **(MU-PFB)**: De manera análoga al MU-ABR, aquí se busca comprobar si la presencia del mecanismo de usabilidad *Progress Feedback* mejora o no la eficiencia y eficacia del usuario.

Como se muestra en la Tabla 7, mediante las transformaciones de potencia y Box-Cox, se logra Normalidad y Homogeneidad para la Eficiencia. El test ANOVA verifica la existencia de un efecto significativo únicamente para la variable número de clicks, mientras que no se observan influencias significativas de la presencia/ausencia del mecanismo con respecto al tiempo. De lo anterior, se puede concluir un comportamiento distinto en las dos formas de medir la eficiencia. Considerando la eficiencia

como el grado de interacción (número de clicks) se acepta la hipótesis alternativa (H.2.1.1) que afirma que "Existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-PFB o al no incorporarlo". Sin embargo, para la eficiencia medida desde el tiempo, se acepta la hipótesis nula (H.2.1.0) que indica que "No existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-PFB o al no incorporarlo". En cuanto a la Eficacia, no se logran superar los supuestos del ANOVA. Entonces se lleva a cabo el KW y como resultado se acepta la hipótesis nula (H.2.2.0) de que "No existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-PFB o al no incorporarlo".

Tabla 7: Transformaciones, Normalidad, Homogeneidad y Test de Hipótesis del MU-PFB

Atributo	Métrica	Transf.	Normal.	Homog.	Test	Pr(>F)
Eficiencia	Clicks	$x^{-(\frac{1}{2})}$	0.0576	0.5147	ANOVA	0.0197 *
	Tiempo	Box-Cox (λ= -0.5)	0.6121	0.9404	ANOVA	0.708
Eficacia	%Tarea	Ninguna	2.2e-16	0.8444	KW	0.9844

***H.3. Preferences* (MU-PRF)**: Aquí se busca comprobar si la presencia del mecanismo de usabilidad *Preferences* mejora o no la eficiencia y eficacia del usuario.

En la Tabla 8 se muestran las comprobaciones de los supuestos de Normalidad y Homogeneidad para la Eficiencia. Se aplica el test ANOVA y la misma revela que no existe efecto significativo del mecanismo MU-PRF en la eficiencia medida en tiempo ni en el número de clicks. Por tanto, se acepta la hipótesis nula (H.3.1.0) que sostiene que "No existe diferencia significativa en la EFICIENCIA del usuario al incorporar el MU-PRF o al no incorporarlo". Se aplica KW para la Eficacia porque no se logra superar los supuestos del ANOVA. El KW demuestra un efecto estadísticamente significativo entre la presencia y ausencia del MU-PRF. En consecuencia, se acepta la hipótesis alternativa (H.3.2.1) de que "Existe diferencia significativa en la EFICACIA del usuario al incorporar el MU-PRF o al no incorporarlo".

Tabla 8: Transformaciones, Normalidad, Homogeneidad y Test de Hipótesis del MU-PRF

Atributo	Métrica	Transf.	Normal.	Homog.	Test	Pr(>F)
Eficiencia	Clicks	$\log(x)$	0.2937	0.8495	ANOVA	0,1282
	Tiempo	$\log(x)$	0.427	0.9922	ANOVA	0,169
Eficacia	%Tarea	Ninguna	7.085e-07	0.8667	KW	1.6e-10 ***

5. DISCUSIÓN

El análisis de los datos obtenidos en los diferentes test, presentados en el apartado anterior, facilitan entender las conclusiones de la investigación. Así, con respecto al mecanismo de usabilidad *Abort Operation* (MU-ABR), la presencia de este mecanismo influye positivamente en la **Eficiencia**. En este sentido, se han obtenido evidencias de una mejora en el tiempo y en la cantidad de clicks invertidos por el sujeto durante la realización de la tarea. Por un lado, cuando el mecanismo está activo se espera que el usuario utilice la opción cancelar directamente, mientras que, cuando no lo está, invierta algún tiempo adicional buscando la manera de hacerlo. Por otro lado, la ausencia de una opción de cancelación produce un incremento en la cantidad de clicks, entendiéndose como mayor esfuerzo invertido por el usuario. Para la **Eficacia**, la mejora se refleja en que un mayor número de usuarios ha logrado cumplir con el desafío de la tarea gracias al mecanismo presente

Para el mecanismo de usabilidad *Progress Feedback* (MU-PFB), los resultados son muy dispares dependiendo de la variable medida para la **Eficiencia**. Se han obtenido evidencias significativas a favor de una mejora en el nivel de interacción del usuario pero ninguna evidencia a favor de una mejora en la rapidez del usuario. En tal sentido, resulta lógico pensar que es razonable que no mejore la rapidez del usuario porque el mecanismo despliega información emergente acerca del progreso de la tarea haciendo que los tiempos de ejecución de cada usuario tienda hacia los mismos valores. Sin embargo, se hace evidente una mejora en la interacción ya que, al no estar presente el mecanismo, no se proporciona indicación alguna del estado del sistema causando un probable incremento en el número de clicks del usuario. Este incremento se interpretaría como una pérdida temporal del control del sistema o una desorientación del usuario debido a la sensación de inactividad o retardo derivado de la tarea de búsqueda. En términos de **Eficacia**, las evidencias obtenidas para el MU-PFB suponen que la presencia o no del mecanismo no ha sido determinante para la finalización exitosa de la tarea debido a la simplicidad de la tarea y al poco retardo en su realización haciendo que los efectos del mecanismo sean ignorados por los usuarios.

Por último, para el mecanismo de usabilidad *Preferences* (MU-PRF) no se observan mejoras con respecto a la **Eficiencia**, ni en tiempo ni en el número de clicks. En este sentido, recuerde que la tarea que involucra al MU-PRF es la tarea ficticia y consiste en buscar visualmente información en la pantalla (sin implicar procesos complejos) después de haber personalizado la interfaz de la aplicación. Con esta premisa, resulta lógico pensar que no existan diferencias significativas en el tiempo y en el número de clicks debido a la propia sencillez de la tarea, que consiste en pasearse visualmente por la tienda buscando solamente información de plazos y localizar dicha información requiere hacer, a lo más, un solo click. Considerando la **Eficacia**, las evidencias son extremadamente fuertes dando un indicativo que un mayor número de usuarios han logrado cumplir con la tarea cuando contaban con el mecanismo activado.

6. VALIDEZ EXPERIMENTAL

La validez interna se centra en la supervisión del proceso para establecer las relaciones de asociación entre las variables, independientes y dependientes, de manera que confiemos que los resultados del experimento se puedan interpretar y se consideren válidos. La validez externa está relacionada con el establecimiento de las condiciones que permiten la generalización de los resultados al ámbito natural en el que aparecen los procesos investigados.

Validez Interna

Esta validez se relaciona con la calidad del experimento. Las amenazas a la validez interna son influencias que pueden afectar la variable independiente con respecto a la causalidad fuera del conocimiento del investigador. En este experimento se consideran potenciales amenazas a la validez interna lo relacionado a:

1. Conocimiento de la tecnología: aunque todos los participantes parten de un mismo nivel de experiencia hacia

este tipo de experimentos (nivel novato), no todos tienen el mismo nivel de conocimiento de la actividad que se va a desarrollar.
2. Orden de realización de las tareas: podría producir un sesgo por efecto de aprendizaje, ya que las tareas asociadas a cada mecanismo se ejecutan secuencialmente.
3. Baja experiencia de los usuarios: todos los sujetos son novatos en la aplicación, por tanto existe la amenaza de que no inviertan el esfuerzo necesario para entender las indicaciones, no entiendan cómo proceder, etc.
4. Motivación: es de esperarse que cada participante tenga una reacción diferente al experimento y pueden existir sujetos que se vean afectados negativamente sobre todo cuando alternan la realización del experimento con otras actividades.

Para mitigar cualquier posible influencia de las dos primeras amenazas los sujetos son asignados aleatoriamente dentro de un grupo manteniendo el balance de participantes entre grupos. Este procedimiento de aleatorización, según Suresh [18], representa un seguro experimental ya que las interferencias pueden o no ocurrir, independientemente de su impacto. Generalmente, se recomienda tomarse el trabajo de aleatorizar a los efectos de eliminar cualquier sesgo potencial.

La tercera amenaza se suple introduciendo el orden como un factor dentro del diseño tal y como se detalla en la Tabla 2.

Finalmente, si bien la cuarta amenaza no se puede evitar [19], y a los efectos de atenuar su impacto se han realizado entrevistas aleatorias a los sujetos para averiguar si fueron afectados por algún factor de cansancio, pereza o similar, que deban ser tenidos en cuenta durante el análisis y la interpretación de los resultados.

Validez Externa

La validez externa hace referencia a la extensión y forma en que los resultados pueden generalizarse en otros contextos. En este sentido, los resultados del experimento no son generalizables a todos los usuarios. Se han seleccionado participantes no informáticos para evitar el sesgo por la familiaridad con la tecnología, sin embargo ellos son asiduos usuarios de Internet. Adicionalmente, estos sujetos son personas que conforman una parte importante de la población de usuarios que utilizan poco las aplicaciones web para comprar productos, esto permite obtener evidencias empíricas con buen nivel de confianza sobre el impacto de los mecanismos de usabilidad analizados a nivel de usuarios no informáticos.

Finalmente, una probable amenaza o limitación del estudio es la generalización a otras aplicaciones diferentes al dominio seleccionado. Esta amenaza puede ser eliminada en implementaciones del experimento usando otro dominio de aplicación.

7. CONCLUSIONES Y LÍNEAS FUTURAS

En este artículo se ha presentado un estudio empírico de los mecanismos de usabilidad en una aplicación web con el propósito de evaluar el impacto de la usabilidad desde la perspectiva de los usuarios en el contexto del usuarios no informáticos. Los mecanismos sometidos al estudio empírico han sido: *Abort Operation* (MU-ABR), *Progress Feedback* (MU-PFB) y *Preferences* (MU-PRF).

A nivel global, las evidencias obtenidas presentan una tendencia a favor de una mejora en la usabilidad del sistema. No se han obtenido evidencias de que la usabilidad vaya en detrimento de la interacción del usuario con el sistema. En este sentido, se puede extraer las siguientes conclusiones:

- El único mecanismo de usabilidad que introduce mejoras en ambos atributos de usabilidad es el MU-ABR. Concretamente, los usuarios, quienes cuentan con la presencia del mecanismo realizan la tarea con más eficiencia y eficacia que aquellos quienes no lo tienen.

- Se obtienen evidencias de que la usabilidad no mejora la eficiencia del usuario en términos de rapidez. En el caso de MU-PRF, resulta razonable que no mejore en la rapidez del usuario al tratarse de una tarea sencilla, es decir, se solicita localizar visualmente una información en la pantalla y no requiere hacer complejas interacciones con el sistema. En el caso del MU-PFB, cuya responsabilidad es la de proporcionar retroalimentación sobre lo que hace el sistema, no representa una interferencia o condición bloqueante para el usuario en la realización de la tarea. Solo se intenta mantener la visibilidad del estado del sistema durante potenciales esperas del usuario producidas por el proceso de búsqueda. Por este motivo, los tiempos invertidos por los usuarios tienden a ser los mismos, independientemente de la presencia/ausencia del mecanismo.

- Es posible afirmar que la ausencia de ciertos mecanismos de usabilidad resulta casi determinante para la Eficacia del usuario como el caso del MU-ABR y el MU-PRF. Por un lado, la ausencia del MU-ABR ha requerido esfuerzo adicional de los usuarios para deshacer sus cambios de manera manual al no existir una opción directa de cancelación. Por otro lado, la ausencia del MU-PRF ha ocasionado que los usuarios se vean imposibilitados de hacer la tarea debido a la dificultad visual ocasionada por la apariencia poco legible de la interfaz.

De todo ello, el estudio revela la importancia de la presencia de los MU-ABR y MU-PFB para obtener una mejora en la eficiencia y del MU-ABR y MU-PRF para la eficacia de los usuarios. Asimismo, parece mostrar que la mejora en la eficiencia y eficacia depende del dominio del problema. En este sentido, puede que un mecanismo de usabilidad, implementado en tareas más complejas o en otro dominio, tenga un efecto distinto del obtenido en este estudio.

Como trabajo futuro se extenderá el análisis de impacto de los tres mecanismos de usabilidad sobre el atributo de Satisfacción. Con esto, se logrará reunir mayores evidencias para justificar el esfuerzo adicional invertido por mejorar los niveles de usabilidad de los sistemas software y para establecer un estudio acerca de los beneficios que aportan la incorporación de ciertas características de usabilidad y que son percibidas por los usuarios en términos de eficiencia, eficacia y satisfacción.

Finalmente, los estudios empíricos van tomando cada vez más importancia en IS generando evidencias que sustentan la evolución del conocimiento a fin de ir mejorando las teorías como consecuencia de los resultados empíricos. Esta investigación se encamina en esta dirección.

Agradecimientos. Esta investigación ha sido apoyada por el Grupo de Investigación en Ingeniería del Software Empírica (GrISE) y por la Universidad Autónoma de Asunción (UAA). También ha sido financiada por el Ministerio de Educación, Cultura y Deportes de España, los proyectos FLEXOR (TIN2014-52129-R) y TIN2014-60490-P, y el proyecto eMadrid-CM (S2013/ICE-2715).

8. REFERENCIAS

[1] ISO/IEC-25010, "Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - System and Software Quality Models," 2010.

[2] N. Juristo, A. M. Moreno, and M.-I. Sanchez-Segura, "Analysing the Impact of Usability on Software Design," *Journal of Systems and Software*, vol. 80, no. 9, pp. 1506–1516, Sep. 2007.

[3] A. Issa and C. H. Bong, "Measuring Software Quality in Use: State-of-the-Art and Research Challenges," *ASQ. Software Quality Professional*, vol. 17, no. 2, pp. 4–15, 2015.

[4] J. I. Panach, "Incorporacion de Mecanismos de Usabilidad en un Entorno de Producción de Software Dirigido por Modelos," Supervisors: Pastor, O. and Juristo, N. Universidad Politécnica de Valencia, 2010.

[5] J. I. Panach, N. J. Juzgado, and O. Pastor, "Including functional usability features in a model-driven development method," *Computer Science and Information Systems*, vol. 10, no. 3, pp. 999–1024, 2013.

[6] M. Aveledo, "Identificación Empírica de Beneficios de Usabilidad," Supervisor: Sánchez-Capuchino, A.M. Universidad Politécnica de Madrid, 2014.

[7] V. R. Basili, G. Caldiera, and H. D. Rombach, "The goal question metric approach," *Encyclopedia of Software Engineering*, vol. 2, pp. 528–532, 1994.

[8] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, 2000.

[9] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental Methods: Between-Subject and Within-Subject Design," *Journal of Economic Behavior and Organization*, vol. 81, no. 1, pp. 1–8, 2012.

[10] J. M. Ferreira, "QuickStore (Experimental Application)," 2016. [Online]. Available: http://webadm.senado.gov.py/tesisweb/.

[11] J. M. Ferreira and S. T. Acuña, "A Software Application for Collecting Usability Empirical Data about User Efficiency , Effectiveness and Satisfaction," in *XII Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento JIISIC'2017*, 2017, p. 11.

[12] ISO 9241-11, "Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)–Part II Guidance on Usability," 1998.

[13] ISO/IEC 25022, "Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of quality in use," ISO/IEC, 2012.

[14] K. Hornbæk, "Current Practice in Measuring Usability: Challenges to Usability Studies and Research," *International Journal of Human Computer Studies*, vol. 64, no. 2, pp. 79–102, 2006.

[15] A. Seffah, M. Donyaee, R. B. Kline, and H. K. Padda, "Usability Measurement and Metrics: A Consolidated Model," *Software Quality Journal*, vol. 14, no. 2, pp. 159–178, 2006.

[16] D. C. McFarlane, "Coordinating the Interruption of People in Human-Computer Interaction," in *Human-computer interaction, INTERACT'99: IFIP TC. 13 International Conference on Human-Computer Interaction*, 1999, pp. 295–303.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2009.

[18] K. Suresh, "An Overview of Randomization Techniques: An Unbiased Assessment of Outcome in Clinical Research," *Journal of Human Reproductive Sciences*, vol. 4, no. 1, pp. 8–11, 2011.

[19] S. España, N. Condori, R. Wieringa, A. González, and Ó. Pastor, "Model-Driven System Development: Experimental Design and Report of the Pilot Experiment," *Computer Research Repository*, vol. abs/1111.0, p. 83, 2011.