Uso de amenazas a la validez para replicar experimentos en captura de requisitos *software*[1]

Using Threats to Validity to Replicate Experiments in Requirements Elicitation Software[2]

O uso de ameaças à validade para replicar experimentos na captura de requisitos *software*[3]

Dante Carrizo-Moreno[4]
Óscar Dieste-Tubio[5]
Marta López-Fernández[6]

SICI: SICI: 0123-2126(201301)17:1<59:UAVPRE>2.0.TX;2-X

---

[1] Fecha de recepción: 11 de enero de 2012. Fecha de aceptación: 10 de agosto de 2012. Este artículo se deriva del proyecto de investigación TIN2011-23216 desarrollado por el Grupo de Investigación en Ingeniería de Software Experimental (GRISE) de la Universidad Politécnica de Madrid, España. También fue financiado por el Proyecto DIUDA, según resolución 83 de 2011 de la Universidad de Atacama, Chile.

[2] Reception date: January 11th 2012. Admission date: August 10th 2012. This paper originated from the research project TIN2011-23216 carried out by the Experimental Software Engineering research group (GRISE) of the Universidad Politécnica in Madrid, Spain. It was financed by the DIUDA project, according to resolution 83 of 2011 of the Universidad de Atacama, in Chile.

[3] Data de recepção: 11 de janeiro de 2012. Data de aprovação: 10 de agosto de 2012. Este artigo origina-se do projeto de pesquisa TIN2011-23216 desenvolvido pelo *Grupo de Investigación en Ingeniería de Software Experimental* (GRISE) [Grupo de Pesquisa em Engenharia de *Software* Experimental] da Universidad Politécnica de Madrid, Espanha. Foi financiado também pelo *Proyecto DIUDA*, segundo a resolução 83 de 2011 da Universidad de Atacama, Chile.

[4] Ingeniero civil informático, Universidad de Concepción, Chile. Máster en Ingeniería de Software y del Conocimiento, Universidad Politécnica de Madrid, España. Doctor en Ingenieria de Software, Universidad Politécnica de Madrid. Profesor asistente, Universidad de Atacama, Copiapó, Chile. Correo electrónico: dante.carrizo@uda.cl.

[5] Licenciado en Informática, Universidad La Coruña, España. Doctor en Informática, Universidad Castilla La Mancha, España. Profesor, Universidad Politécnica de Madrid. Madrid, España. Correo electrónico: odieste@fi.upm.es.

[6] Licenciada en Informática, Universidad La Coruña, España. Doctora en Informática, Universidad Politécnica de Madrid, España. Xunta de Galicia, La Coruña, España. Correo electrónico: marta.lopez.fernandez@xunta.es.

Resumen

Las entrevistas son las técnicas de elicitación más utilizadas en la ingeniería de requisitos (IR); sin embargo, existen pocos trabajos de investigación experimentales centrados en estas técnicas. Recientemente hemos experimentado para analizar la efectividad de las entrevistas estructuradas y de las no estructuradas. Uno de estos trabajos se ha sumado a otros en el campo de sistemas de información para conformar una base de estudios. Aunque los estudios metanalizados parecen similares según sus diseños, fijándonos en las amenazas a la validez se identifican más diferencias que similitudes. El análisis de estas amenazas puede ser un medio para identificar variables moderadoras y comprender cómo mejorar el diseño de futuras replicaciones para generar nuevas evidencias y mejorar resultados de los metanálisis. Se muestra un procedimiento para lograr este objetivo.

Palabras clave

Ingeniería de requisitos, educción de requisitos, entrevistas, experimentación, diseño experimental, metanálisis.

Abstract

Interviews are the most frequently used elicitation methods in requirements engineering (RE); nevertheless, there are but few experimental research papers focused on such methods. Recently we have carried out experiments in order to analyze the effectiveness of structured and unstructured interviews. One of those papers has joined others in the field of information systems forming a basis for study. Although meta-analyzed studies seem similar in their design, a focus on threats to validity uncovers more differences than similarities. The analysis of such threats can be a tool for identifying moderator variables and understanding how to improve the design of future replications in order to create new evidence and improve the results of meta-analyses. This paper shows a procedure to achieve this.

Keywords

Requirements engineering, requirements elicitation, interviews, experimentation, experimental design, meta-analysis.

Resumo

As entrevistas são as técnicas de elicitação mais utilizadas na engenharia de requisitos (ER); no entanto, existem poucos trabalhos de pesquisa experimentais centrados nessas técnicas. Recentemente, fizemos experimentos para analisar a efetividade das entrevistas estruturadas e das não estruturadas. Um desses trabalhos foi acrescentado a outros no campo de sistemas de informação para formar uma base de estudo. Ainda que os estudos de meta-análise sejam similares em termos de design, ao focar as ameaças à validade, identificam-se mais diferenças do que semelhanças. A análise dessas ameaças pode ser um meio para identificar variáveis moderadoras e compreender como melhorar o design de futuras replicações para, desse modo, gerar novas evidências e melhorar os resultados das meta-análises. Demonstra-se um procedimento para atingir este objetivo.

Palavras chave

Engenharia de requisitos, estabelecimento de requisitos, entrevistas, experimentação, design experimental, meta-análise.

Introducción

La ingeniería de requisitos (IR) es una disciplina crucial para el desarrollo de *software* (Abran *et al.*, 2004). Dado que el impacto de requisitos incompletos o ambiguos en la calidad del producto *software* es crítico, es preciso centrar la atención en el proceso de elicitación de requisitos y las técnicas que se van a aplicar para educirlos (Abran *et al.*, 2004; Bell y Thayer, 1976; Leuser *et al.*, 2009). Típicamente las más utilizadas son las entrevistas. Sin embargo, a pesar de su importancia, existe poca investigación empírica sobre la eficiencia de las entrevistas (Hickey, Davis y Kaiser, 2003); mientras que en disciplinas como psicología (Willig, 2001) o economía (Bech-Larsen y Nielsen, 1999; Breivik y Supphellen, 2003) las entrevistas se analizan empíricamente para estudiar su eficiencia, precisión, roles implicados, etc.

Nuestro trabajo en experimentación centrada en la captura de requisitos comenzó con una revisión sistemática de las técnicas de elicitación (Dieste y Juristo, 2011). Algunos de los trabajos resultaron ser estudios experimentales comparativos acerca de la eficacia de las entrevistas estructuradas y no estructuradas (Agarwal y Tanniru, 1990; Browne y Rogich, 2001; Marakas y Elam, 1998; Pitts y Browne, 2004). Los resultados de estos experimentos base resaltan que las entrevistas estructuradas funcionan mejor que las no estructuradas; sin embargo, nuestra opinión es que es preciso disponer de más guías objetivas (Kitchenham, Dyban y Jorgensen, 2004), y esto implica el uso de métodos formales, como el metanálisis, para agregar diversos estudios experimentales. Dado que el metanálisis no es adecuado para un número escaso de experimentos, se diseñaron y se ejecutaron replicaciones de los experimentos base, tomados como punto de referencia para adaptar los tipos de entrevistas, variables respuesta y proceso experimental al contexto de experimento de laboratorio con estudiantes de Informática. Se desarrolló un estudio piloto en 2006 y posteriormente experimentos anualmente, exceptuando 2009 por razones logísticas; 2007, analizado y publicado (Carrizo, *et al.*, 2011); 2008, pendiente de enviar, y 2010, pendiente de analizar.

El diseño de replicaciones es difícil, pues normalmente los experimentos originales no están descritos con todos los detalles necesarios para su replicación. Además, también es preciso analizar las variables moderadoras para conocer en detalle las entrevistas. Pero estas variables, usualmente, tampoco están identificadas de forma explícita. Es preciso que las replicaciones sean muy similares (Juristo y Vegas, 2009), de manera que se pueda analizar la posible influencia de las variables moderadoras. Esto no es aplicable si se considera el conjunto de experimentos base. Una alternativa para determinar las variables moderadoras puede ser el conjunto de limitaciones y amenazas a la validez identificadas en los experimentos. En la mayoría de los casos, los autores identifican como limitaciones aspectos que son amenazas a la validez externa del experimento; por ejemplo, las muestras de conveniencia. Estas amenazas son limitaciones reales; sin embargo, en algunos casos los autores incluyen razones que explican los resultados experimentales obtenidos —por ejemplo, la influencia de los entrevistados—. Estas limitaciones apuntan a la existencia de variables moderadoras que se deberían tener en cuenta al diseñar nuevas replicaciones.

En este artículo se presenta una aplicación del metanálisis para combinar resultados de los experimentos base y aquellos obtenidos por los autores en la experimentación de 2007, descrito en (Carrizo *et al.*, 2011). Dada la potencial influencia de las variables moderadoras, también se ha analizado el uso de las limitaciones como fuente para identificar dichas variables, pues no están explícitamente descritas en los experimentos base.

1. Antecedentes

En la literatura de disciplinas de investigación como psicología (Willig, 2001), economía (Bech-Larsen y Nielsen, 1999; Breivik y Supphellen, 2003), etc. se encuentran muchos trabajos acerca de las entrevistas. En el desarrollo de *software* también existen algunos pocos trabajos. Centrándonos en aquellos estudios empíricos sobre las entrevistas, los más representativos son los de Agarwal y Tanniru (1990), Browne y Rogich (2001), Marakas y Elam (1998) y Pitts y Browne (2004). Para cada uno se detallan el ámbito, el objetivo, el diseño, los participantes y el proceso. Todos ellos analizan alguna hipótesis acerca de la efectividad de las entrevistas, que se estudia en función de la experiencia de los entrevistadores.

Agarwal y Tanniru (1990) llevaron a cabo un experimento sobre sistemas expertos para la toma de decisión. Dentro del proceso de adquisición de conocimiento, el objetivo era comparar la entrevista no estructurada con un tipo concreto de entrevista estructurada basada en el modelo de Duncan. Los entrevistadores

fueron estudiantes graduados y profesionales que sesionaron con treinta expertos. El análisis de las transcripciones lo realizaron dos codificadores independientes. Browne y Rogich (2001) experimentaron en sistemas de información acerca de la utilidad de tres tipos técnicas de elaboración de preguntas. Los entrevistados fueron 45 trabajadores universitarios, no docentes y sin experiencia en el desarrollo de *software*. Todas las sesiones se grabaron y se transcribieron para el análisis y clasificación de los requisitos por un codificador independiente.

Marakas y Elam (1998) llevaron a cabo su experimento en sistemas de información, con el objetivo de investigar la efectividad de una técnica de entrevista semántica por medio de la correctitud de los diagrama de flujo de datos de un sistema de compra de productos. Los dos grupos de control estaban formados por analistas de baja y alta experiencia, respectivamente, que usaban la entrevista no estructurada; los dos grupos experimentales los componían analistas de baja y alta experiencia, respectivamente, utilizando el modelo semántico. Los entrevistados fueron cuatro profesionales externos, asignados de tal manera que cada uno fue entrevistado el mismo número de veces en cada grupo. Por último, Pitts y Browne (2004) diseñaron y ejecutaron un experimento en el ámbito de los sistemas de información para analizar una actividad cognitiva: cómo determinan los analistas que tienen la suficiente información recopilada en el proceso de elicitación de requisitos. Si bien el objetivo no coincide con los anteriores experimentos, se ha incluido, pues entre sus hipótesis se analiza también la eficiencia de la entrevista y la relación con la experiencia del analista. El entrevistado fue una única persona sin relación con el experimento. Más detalles sobre estos experimentos se pueden encontrar en http://www.grise.upm.es/sites/extras/1/.

2. Experimento

En esta descripción se usa el mismo esquema que para los anteriores experimentos: ámbito, objetivo, diseño, participantes y proceso. En comparación con los anteriores, además de la experiencia del entrevistador se añadió el tipo de problema.

2.1. Descripción del experimento

El objetivo del experimento era analizar la eficiencia de dos tipos de entrevistas:

- Entrevistas no estructuradas, caracterizadas por preguntas abiertas y genéricas que no requieren preparación previa.

- Entrevistas independientes del contexto, que son un tipo de entrevista estructurada caracterizada por un conjunto de cuestiones genéricas centradas en el aspecto que se va a analizar que, normalmente, se utilizan en la primera sesión de entrevistas del proceso de elicitación.

Se aplicó un diseño factorial 2×2 con medidas repetidas y con dos factores (tipo de entrevista y tipo de problema) y solo una única variable respuesta: el número de requisitos identificados por los trece sujetos experimentales (alumnos del Máster de Ingeniería de Software, Universidad Politécnica de Madrid). Todos ellos son ingenieros informáticos con experiencia en el desarrollo de *software* y, en alguna medida, en elicitación de requisitos, que asumieron el rol de ingenieros de requisitos durante la sesión de captura. Dos de los autores adoptaron el rol de entrevistados. Uno de ellos, especializado en un sistema de control de máquinas de reciclado de pilas, y otro, en un sistema de gestión de comisiones de un hipotético departamento universitario.

A cada sujeto se le asignó el tipo de entrevista que se iba a aplicar mediante insaculación y, por lo tanto, de modo aleatorio. La asignación fue ciega para los experimentadores y entrevistados. La experimentación comenzó con la planificación de las diversas sesiones, para asignar cada par [estudiante/entrevistador-técnica] a cada [entrevistado-problema]. Cada sesión (30 min) fue grabada y transcrita, previa al desarrollo de la lista de requisitos de *software*. Las listas y los ficheros de audio originales se han utilizado para analizar los datos (Anova) y extraer el número de requisitos de cada problema.

2.2. Resultados del experimento

Las hipótesis analizadas en este trabajo y los resultados finales son los siguientes:

H1. No existen diferencias en efectividad entre la entrevista no estructurada y la entrevista independiente de contexto.

No se puede rechazar H1. Ambas técnicas poseen una eficacia similar. Sin embargo, no puede negarse que la entrevista independiente de contexto ejerce una influencia positiva en el proceso de elicitación de requisitos.

H2. El tipo de problema no afecta a la efectividad de las entrevistas.

No se puede aceptar H2. El tipo de problema tiene influencia en la efectividad de las entrevistas.

Uno de los factores más influyentes fue la aparición del cansancio en los entrevistadores a lo largo del segundo día de sus respectivas sesiones de elicitación. Un análisis de las medias marginales apunta a la existencia de un efecto de *carry-*

over, pues se ha detectado una disminución en la efectividad, tanto respecto al factor "tipo de entrevista" como al factor "problema que se va a estudiar", entre la primera sesión y la segunda.

En teoría se puede pensar que en la segunda sesión se puede producir este tipo de efecto; mas al contrario: incrementando la eficiencia, debido a la influencia del efecto de aprendizaje. Nosotros detectamos justo lo contrario. La causa más plausible es el cansancio de los entrevistadores, derivado de la mayor carga de trabajo en los días de las segundas sesiones de entrevistas y también influido por la experiencia de cada entrevistador en *role-playing*. Se reaplicó la Anova pero teniendo en cuenta solo las primeras sesiones de ambos problemas. El resultado de H1 se reafirma y el de H2 se cambió, pues el análisis inicial indicaba que el tipo de problema no afecta a la eficacia.

Otros resultados interesantes se obtuvieron del análisis de los datos demográficos de cada sujeto experimental contrastado con su efectividad. Así, la experiencia del analista se correlaciona muy fuertemente con la efectividad del proceso de elicitación de requisitos. Respecto a las amenazas a la validez, hemos detectado varias (véase Carrizo *et al.*, 2011 para más detalle) pero solo una de ellas puede tener influencia en las hipótesis planteadas. Concretamente, se refiere al hecho de que se pueda haber confundido el problema que se va a estudiar y el entrevistado en su papel de cliente. Esto implica que todos los efectos que se adscriben al tipo de problema pueden deberse en realidad al experimentador.

3. Análisis de los resultados

El conjunto de replicaciones experimentales proporcionan la información de la tabla 1. En lo referente a los factores, el conjunto de replicaciones ha estudiado preferentemente dos de ellos: la técnica de elicitación (Agarwal y Tanniru, 1990; Browne y Rogich, 2001; Carrizo *et al.*, 2011) y la experiencia de los participantes en el proceso (Agarwal y Tanniru, 1990; Marakas y Elam, 1998; Pitts y Browne, 2004). El tipo de problema solo lo han explorado Carrizo *et al.* (2011). Respecto a las variables respuesta, la situación es opuesta: se han estudiado muchas y muy diversas. La eficiencia es la variable respuesta analizada en todos los experimentos, utilizando métricas como el número de requisitos o la cantidad de reglas extraídas. En algún caso se han estudiado también las categorías en las que se subdividen los requisitos (Browne y Rogich, 2001). Otro gran subgrupo de variables respuesta es el referente a los contenidos. Estas variables recogen los errores lógicos (Marakas y Elam, 1998) o la diferencia cualitativa entre la información obtenida por las distintas técnicas de elicitación (Pitts y Browne, 2004).

No es casualidad que la eficiencia sea la variable más explorada en todos los experimentos. Averiguar cómo aumentar la cantidad de información extraída es el objetivo final de todos los experimentadores. La eficiencia puede estudiarse gracias a metanálisis, tanto desde la perspectiva de la técnica de elicitación como de la experiencia de los sujetos (zona sombreada en la tabla 1). También puede estudiarse el efecto conjunto de las técnicas y la experiencia de los sujetos, aunque de forma muy limitada. Las demás variables respuesta no pueden agregarse. Un detalle acerca de la agregación de evidencia se puede encontrar en Dieste y Juristo (2011).

Tabla 1. Información proporcionada por el conjunto de replicaciones experimentales

			Agarwal y Tanniru	Browne y Rogich	Carrizo *et al.*	Marakas y Elam	Pitts y Browne
Factores	Técnica de elicitación		X	X	X	X	
	Novatos c. expertos		X			X	X
	Tipo de problema				X		
Variables respuesta	Eficiencia	Cantidad de información	X	X	X	X	X
		Categorías		X			
	Contenido	Subjetividad	X				
		Errores lógicos				X	
		Diferencias cualitativas		X			
	Otros	*Recall*, anticipación	X				
		Suponer, inferir				X	
		Patrones				X	

Fuente: presentación propia de los autores.

4. ¿Cuál es la contribución de nuestro experimento?

En términos de resultados obtenidos, el experimento revela pocas novedades. Aporta con un conjunto de resultados que amplía el número de factores (aparte de la experiencia se añade el tipo de problema) y que proporciona unas pocas evidencias más sobre la efectividad de las entrevistas en el proceso de elicitación de requisitos. Pero, ahora bien, del análisis de las limitaciones y amenazas a la validez identificadas en los experimentos se pueden extraer variables moderadoras, sin las cuales es difícil diseñar replicaciones de los experimentos. Esto sí constituye una contribución con la que se puede profundizar en esta área de

investigación en la IR. El proceso de identificación de variables moderadoras se puede resumir en tres partes: 1) identificación de limitaciones, 2) análisis de las limitaciones y 3) determinación de variables moderadoras.

4.1. Identificación de limitaciones

El análisis de las limitaciones comienza con la tabla 2, que muestra las limitaciones identificadas por Agarwal y Tanniru (1990), Browne y Rogich (2001), Carrizo *et al.* (2011), Marakas y Elam (1998), Pitts y Browne (2004). Cada experimento identifica entre cuatro y cinco limitaciones, por término medio, salvo Browne y Rogich (2001), que solo detallan dos. Las limitaciones se han clasificado en función del aspecto al que se refieren. Por ejemplo, cuando Agarwal y Tanniru (1990) indican que "los expertos constituyen una muestra de conveniencia, en vez de una muestra aleatoria, aunque se haya asignado aleatoriamente los expertos a los grupos", se está haciendo referencia a la muestra utilizada en la experimentación. Así se han identificado las siguientes categorías:

- Proceso: cómo se ha llevado a cabo el experimento.
- Muestra: características de los entrevistados y entrevistadores.
- Técnicas: cómo se han aplicado las entrevistas.

Estas categorías son genéricas y probablemente se utilicen para clasificar las limitaciones en otras áreas, como las pruebas del *software*.

4.2. Análisis de las limitaciones

Analizando las limitaciones por categorías (filas), se pueden obtener las coincidencias, las frecuencias, etc. Por ejemplo, destaca que la principal limitación identificada en la categoría de *proceso* es que todos los procesos experimentales son experimentos de laboratorio. Todas las restantes limitaciones del *proceso* se centran en aspectos concretos, como el número de sesiones o la complejidad de las tareas experimentales. La categoría más frecuentemente identificada como limitación en los cinco experimentos es la *muestra*, en la que se incluyen los entrevistadores, los entrevistados, los codificadores y cualquier otro rol necesario para llevar a cabo el experimento.

Algunos autores se centran en problemas específicos de las muestras, como su motivación o que sean muestras de conveniencia. Sin embargo, todos los experimentos coinciden en reconocer la experiencia y el *role-playing* como limitaciones. Cuatro de los trabajos centran estos aspectos en el entrevistador y solo uno en el entrevistado. El criterio menos relevante, según la tabla 2, es la *técnica*.

Tabla 2. Número y descripción de todas las limitaciones por experimento

	Agarwal y Tanniru		Browne y Rogich		Carrizo *et al.*		Marakas y Elam		Pitts y Browne		Total
Proceso	3	Solo un dominio de problema Sin estándar de referencia Solo una sesión	-		1	Problemas no complejos	2	Entorno de laboratorio Estudio exploratorio	1	Solo una estrategia de elicitación	7
Muestra	2	Muestra de conveniencia *Role-playing*	1	Experiencia previa de los entrevistadores	2	Entrevistado *vs.* problema Entrevistados no tienen preferencias previas acerca de las técnicas	3	Tamaño de la muestra Motivación de la muestra Compromiso de la muestra	3	Experiencia de los entrevistadores *vs.* dominio del problema Experiencia cuantificada de los entrevistadores Solo un entrevistado	11
Técnicas	-		1	Precisión de la codificación	1	Variabilidad de la técnica aplicada	2	Uso mínimo de la técnica en el 70 % del tiempo de entrevista Solo un método de representación	2	Solo una medida de las reglas de parada cognitivas Codificación centrada en una taxonomía de requisitos predefinida	6
Total	5		2		4		7		6		24

Fuente: presentación propia de los autores.

Centrándonos en las técnicas de elicitación aplicadas en las entrevistas, Carrizo *et al.* (2011) y Marakas y Elam (1998) presentan posturas opuestas, en cuanto a si se acepta o no la variabilidad en la aplicación de las técnicas estructuradas. Son solo características de los diseños experimentales desarrollados por esos autores.

Aparte de estas coincidencias y frecuencias, es más interesante un análisis de cada limitación mostrada en la tabla 2 para determinar si es una amenaza a la validez o una limitación en sentido estricto, que serán tratadas a continuación. Este análisis es crucial para nuestro objetivo, porque las amenazas a la validez son restricciones metodológicas que no pueden usarse para identificar las variables moderadoras. Por ejemplo, Marakas y Elam (1998) identifican como limitación que su estudio es un experimento de laboratorio. Este tipo de experimentos son limitados si nos referimos al tipo de conocimiento que se puede obtener de ellos (entorno estrictamente controlado, condiciones ideales, etc.); pero esto no implica la existencia de una variable moderadora. Análogamente, Agarwal y Tanniru (1990) identifican como limitación el hecho de que los sujetos implicados en el experimento sean una muestra de conveniencia. De todas formas, esto tampoco apunta a la existencia de una variable moderadora.

Las limitaciones relacionadas con el número de sesiones, con el número de problemas, con la complejidad y con el número de técnicas aplicadas están relacionadas con el coste, el esfuerzo y la disponibilidad de las personas involucradas. Claramente son restricciones que afectan a la generalización de los resultados de los experimentos, pero no implican la existencia de aspectos que afecten la efectividad de las entrevistas (es decir, una variable moderadora).

4.3. Determinación de variables moderadoras

Si se eliminan de la tabla 2 los casos mencionados hasta ahora, se obtiene el conjunto de limitaciones efectivas mostradas en la tabla 3, que indican una falta de validez de los resultados experimentales dentro del contexto de cada experimento (es decir, validez interna). Por ejemplo, la experiencia de los entrevistadores podría ser una de las limitaciones efectivas porque Browne y Rogich (2001) señalan que la experiencia podría afectar a la efectividad de las entrevistas. Por lo tanto, si en los experimentos no se controla la experiencia de los sujetos, los resultados podrían invalidarse, es decir, la experiencia es una potencial variable moderadora.

Otros ejemplos son la "Precisión de la codificación" (Browne y Rogich, 2001) o la "Codificación centrada en una taxonomía de requisitos predefinida" (Pitts y Browne, 2004). Se pueden considerar limitaciones efectivas, pues es posible que sean el origen de un sesgo de medición que podría afectar las hipótesis.

También, las limitaciones agrupadas bajo la categoría *muestra* podrían tomarse como fuentes de sesgos, exceptuando el tamaño de la muestra, que no es considerada una limitación pero sí un factor influyente en la potencia estadística. Una muestra de tamaño mayor solo aumenta la confidencia de una estimación.

En la literatura sobre experimentos en otras áreas científicas —economía (Bech-Larsen y Nielsen, 1999; Breivik y Supphellen, 2003), medicina (Swanson, 1987), etc.— las limitaciones efectivas de la tabla 3 normalmente están relacionadas con tipos de sesgos concretos. En este contexto, el significado de sesgo es el de error sistemático, o aspectos influyentes no deseados, de diverso origen, que es preciso eliminar o minimizar para aumentar la exactitud y precisión de un experimento.

Aparentemente, los autores de los estudios analizados comparten una perspectiva similar acerca de las limitaciones mostradas en la tabla 3. Por ello las incluyen en las secciones *amenazas* a la validez de sus trabajos. En algunos casos, dicho proceder está completamente justificado, si la limitación es un sesgo claramente. Por ejemplo, las limitaciones bajo la categoría *técnicas* de la tabla 3 son instancias de un sesgo de medición (riesgo en la determinación precisa de los valores de variables respuesta). También la motivación es un requisito para ejecutar adecuadamente una tarea, independientemente del área de conocimiento, y no parece que sea un objeto legítimo de investigación. Así, estos dos aspectos quedan fuera de este análisis en la búsqueda de potenciales variables moderadoras.

Sin embargo, en otros casos esto no es cierto, especialmente en la categoría *muestra*, que es la que contiene más elementos en la tabla 3. Lo que podría ser un sesgo en algunas disciplinas (como economía), podría ser un objeto legítimo de investigación en IR. Este es el caso, por citar un ejemplo muy claro, de la experiencia de los sujetos experimentales. No es sorprendente que la mayoría de este tipo de limitaciones proceda de la categoría *muestra*. En elicitación, los participantes son un aspecto de interés del que se deberían estudiar sus particularidades y las relaciones que establecen con el problema objeto de estudio. Por lo tanto, estos aspectos no son sesgos o riesgos, sino aspectos que es necesario considerar para poder comprender adecuadamente cuándo y cómo funcionan bien las entrevistas. Así, este tipo de limitación como potencial variable moderadora.

La siguiente lista muestra una clasificación de las limitaciones de la tabla 3, teniendo en cuenta los sesgos potenciales a los que pueden dar origen. Nótese solo se incluyen los ejemplos más claros:

- Sesgo del artefacto, relacionado con las limitaciones "Entrevistado *vs*. problema" y "Experiencia de los entrevistadores *vs*. dominio del problema".

Tabla 3. Número y descripción de las limitaciones efectivas por experimento

	Agarwal y Tanniru		Browne y Rogich		Carrizo *et al.*		Marakas y Elam		Pitts y Browne		Total
Proceso	-		-		-		-		-		-
Muestra	1	*Role-playing*	1	Experiencia previa de los entrevistadores	2	Entrevistado *vs.* problema Los entrevistados no tienen preferencias previas acerca de las técnicas	2	Motivación de la muestra Compromiso de la muestra	3	Experiencia de los entrevistadores *vs.* dominio del problema Experiencia cuantificada de los entrevistadores Solo un entrevistado	9
Técnicas	-		1	Precisión de la codificación	-		-		1	Codificación centrada en una taxonomía de requisitos predefinida	2
Total	1		2		2		2		4		11

Fuente: presentación propia de los autores.

- Sesgo del entrevistador, o cualquier error sistemático debido a la manera consciente o inconsciente del entrevistador de recopilar los datos. Relacionado con aquellas limitaciones relativas al *role-playing* y la "Experiencia previa de los entrevistadores".
- Sesgo del entrevistado, relacionado con el *role-playing* y con la posible identificación "Entrevistado *vs*. problema" y el sesgo potencial de usar una u otra técnica.

Como se ha mencionado, los tres puntos anteriores no pueden ser considerados sesgos en el contexto de las entrevistas en IR, porque son aspectos que necesitamos conocer para explicar las razones de la efectividad de las entrevistas y para llevar a cabo la elicitación en la práctica. Por lo tanto, del anterior conjunto de limitaciones podemos identificar las siguientes variables moderadoras: problema, experiencia y características personales. Y con base en estas variables moderadoras, se pueden deducir las siguientes recomendaciones:

1. Ejecutar entrevistas acerca de diferentes tipos de problemas, de diferente tamaño y complejidad y, preferiblemente, de diferentes dominios.
2. Análisis de la experiencia de los sujetos. Un mayor detalle de esta experiencia facilitará el control del experimento y la obtención de datos de mayor calidad. Por ejemplo, indagar por los años, el número y el tamaño de los proyectos en los que ha participado un sujeto.
3. Análisis de los roles asignados a cada sujeto, según su experiencia, aptitudes, conocimiento del dominio, etc. Podría recopilarse mediante test psicológicos para conocer la personalidad del sujeto. No hay que olvidar que existen más roles que los entrevistadores y entrevistados y que pueden ejercer influencia sobre los resultados del experimento.

En la práctica, la aplicación de estas recomendaciones es compleja, debido a características específicas del experimento, medidas, falta de sujetos experimentales adecuados, etc. No obstante, estas variables moderadoras podrían tener influencia y deberían considerarse en los experimentos sobre entrevistas.

5. Posible diseño de un experimento futuro

Un posible diseño experimental que incluya las variables moderadoras identificadas sería:

- Ámbito. Según la primera recomendación, la descripción del problema contiene cuatro aspectos: tipo, tamaño, complejidad y dominio. En vez de aplicar todas estas posibilidades conjuntamente, que conllevaría un diseño

con elevada complejidad, se opta por incluir poco a poco cada aspecto en sucesivos experimentos. En nuestro experimento se consideraron problemas de diferente tipo, pero de similar tamaño y complejidad y mismo dominio. Por lo tanto, en el próximo se incluirán diferentes tipos y tamaños, pero de similar complejidad y mismo dominio.

- Objetivo. Las variables moderadoras no influyen en este aspecto: se utilizarán las mismas que en el experimento ya realizado.
- Diseño. Al igual que el anterior, tampoco es un aspecto en el que influyan las variables moderadoras. La elección del diseño dependerá de los participantes, factores, variables, etc., concretas y no es posible determinarlo en este momento.
- Participantes. Es el aspecto en el que más influyen las variables moderadoras, pues se pueden aplicar la segunda y tercera, en función del tipo de sujeto experimental.
- Entrevistador. Solo se considera aplicable la segunda recomendación, pues seguirá siendo un experimento de laboratorio con estudiantes. En posteriores experimentos se podría plantear, aplicando la tercera recomendación, que algunos de ellos asuman el rol de entrevistados. Se plantea la realización de una encuesta demográfica con el objetivo de conseguir, antes del experimento, datos amplios sobre la experiencia, incluyendo número de proyectos, tiempo y calificación de los tamaños.
- Entrevistado. Se pueden aplicar las recomendaciones 2 y 3. Mediante una encuesta o una entrevista se debe averiguar el tipo de conocimiento que tienen del dominio. Los siguientes aspectos se podrían combinar en sucesivos experimentos:
  - ✓ Conocimiento del dominio: mismo o dispar.
  - ✓ Experiencia previa: docente, profesional del *software* o sin relación alguna.
  - ✓ Relación con el experimento: ciego, con conocimiento.
  - ✓ Implicación en el papel: en función de la personalidad y el tipo de respuestas.

  Así, en el siguiente experimento, se aplicaría la cuaterna (conocimiento parejo-todos docentes-con conocimiento del objetivo-misma implicación con aportación de la información justa).
- Codificador. Los valores pueden ser el mismo entrevistador, profesional externo ciego. Su implicación puede ser total o parcial (si solo codifica un porcentaje para verificación de la fiabilidad). En el siguiente experimento se aplicará: [entrevistador-total] y [profesional-parcial al 50 %].

- Proceso. No influyen las recomendaciones en este aspecto. Únicamente que se precisará planificar para intentar balancear problemas como el cansancio de los entrevistados, duración de la(s) sesión(es) de la(s) entrevista(s), etc. Todo este conjunto de aspectos se puede avanzar pero deberá concretarse posteriormente.

Conclusiones

En las publicaciones empíricas se tiende a no distinguir entre amenazas a la validez y otras limitaciones efectivas, considerando ambas restricciones a la validez externa. Pero, probablemente, esto sea una equivocación. Las restricciones metodológicas como la muestra, el tipo o el número de sujetos afectan la validez externa, pero otras limitaciones —por ejemplo, la experiencia de los entrevistadores— no. El análisis de las limitaciones identificadas en los experimentos puede ser una estrategia útil para encontrar variables moderadoras. De esta manera, estas se pueden incluir explícitamente en los diseños de nuevas replicaciones. En este artículo se han aplicado estas ideas a un conjunto de experimentos sobre entrevistas y finalmente se ha obtenido un posible diseño de futuros experimentos, que será mejorado para incorporar una formulación y terminología más rigurosa y sistemática.

Las limitaciones efectivas son, en realidad, aspectos del conocimiento teórico de la correspondiente área científica. Por ejemplo, la experiencia del entrevistador podría tener influencia en la efectividad, como sugiere el sentido común. Sin embargo, influidos por la ingeniería del *software* empírica, es más fácil considerarlas posibles variables moderadoras que podrían influir en los resultados de los experimentos. Como esta, se pueden identificar otros aspectos basados en opiniones expertas enunciadas como relaciones causales sin verificar experimentalmente, que pueden representar limitaciones para tener en cuenta en futuras experimentaciones.

Por ejemplo, en medicina existen estudios que demuestran la existencia de relaciones lógicas informales que, potencialmente, pueden revelar nuevo conocimiento o transformarse en hipótesis interesantes. Un caso típico es la relación causal entre la deficiencia de magnesio y las migrañas en revistas médicas, detectable mediante el uso de técnicas de minería de datos.

Siguiendo en esta línea, nuestro trabajo futuro se centrará no solo en aplicar estas ideas al proceso experimental, sino también a las propias técnicas de educción. Con ello se pretende extraer hipótesis y conocimientos, potencialmente interesantes, hasta ahora ocultos en la literatura sobre técnicas de educción en los diversos campos de conocimiento en donde estas se aplican.

Referencias

ABRAN, A.; MOORE, J. W.; BOURQUE, P.; DUPUIS, R. y TRIPP, L. L. *Guide to the Software Engineering Body of Knowledge (SWEBOK)*. California: IEEE, 2004.

AGARWAL, R. y TANNIRU, M. R. Knowledge acquisition using structured interviewing: an empirical investigation. *Journal of Management Information Systems*. 1990, vol. 7, núm. 1, pp. 123-140.

BECH-LARSEN, T. y NIELSEN, N. A. A comparison of five elicitation techniques for elicitation of attributes of low involvement products. *Journal of Economic Psychology*. 1999, núm. 20, pp. 315-341.

BELL, T. E. y THAYER, T. A. Software requirements: are they really a problem? *Proceedings 2nd International Conference on Software Engineering (ICSE'76)*. 1976, pp. 61-68.

BREIVIK, E. y SUPPHELLEN, M. Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques. *Journal of Economic Psychology*. 2003, vol. 24, pp. 77-98.

BROWNE, G. J. y ROGICH, M. B. An empirical investigation of user requirements elicitation: Comparing the effectiveness of prompting techniques. *Journal of Management Information Systems*. 2001, vol. 17, núm. 4, pp. 223-250.

CARRIZO, D.; DIESTE, O.; JURISTO, N. y LÓPEZ, M. Estudio experimental de la efectividad de la entrevista abierta frente a la entrevista independiente de contexto. *Actas del 14th Workshop on Requirements Engineering*. Río de Janeiro, Brasil. 2011, pp. 297-308.

DIESTE, O. y JURISTO, N. Systematic review and aggregation of empirical studies on elicitation techniques. *IEEE Transaction on Software Engineering*. 2011, vol. 37, núm. 2, pp. 283-304.

HICKEY, A.; DAVIS, A. y KAISER, D. Requirements elicitation techniques: Analyzing the gap between technology availability and technology use. *Comparative Technology Transfer and Society*. 2003, vol. 1, núm. 3, pp. 279-302.

JURISTO, N. y VEGAS, S. Using differences among replications of software engineering experiments to gain knowledge. *Third International Symposium on Empirical Software Engineering and Measurement*, Florida, EE. UU. 2009, vol. 15-16, pp. 356-366.

KITCHENHAM, B.; DYBA, T. y JORGENSEN, M. Evidence-based software engineering. *Proceedings 26th International Conference on Software Engineering (ICSE'04)*. Edinburgh, UK, May 23-28 2004.Washington: IEEE Computer Society, pp. 273-281.

LEUSER, J.; PORTA, N.; BOLZ, A. y RASCHKE, A. Empirical validation of a requirements engineering process guide. *Proceedings 13th International Conference on Evaluation and Assessment in Software Engineering (EASE'09)*. UK, April 20-21, 2009, pp. 1-10.

MARAKAS, G. M. y ELAM, J. J. Semantic structuring in analyst acquisition and representation of facts in requirements analysis. *Information Systems Research*. 1998, vol. 9, núm. 1, pp. 37-63.

MILLER, J. Applying meta-analytical procedures to software engineering experiments. *Journal of Systems and Software*. 2000, núm. 54, pp. 29-39.

PITTS, M. G. y BROWNE, G. J. Stopping behavior of systems analysts during information requirements elicitation. *Journal of Management Information Systems*. 2004, vol. 21, núm. 1, pp. 203-226.

SWANSON, D. R. Two medical literatures that are logically but not bibliographically connected. *American Society for Information Science*. 1987, vol. 38, núm. 4, pp. 228-233.

WILLIG, C. *Introducing qualitative research in psychology: adventures in theory and method*. London: Open University Press, 2001.