Diseño de una Nueva Replicación de Experimentos sobre Entrevistas en Elicitación de Requisitos Utilizando Datos de las Amenazas a la Validez

Dante Carrizo[1], Oscar Dieste[2], Marta López[3]

[1]Universidad de Atacama, Dpto. Ingeniería Informática y Ciencias de la Computación
[2] Universidad Politécnica de Madrid, Facultad de Informática
[3] Xunta de Galicia
{dante.carrizo@uda.cl; odieste@fi.upm.es; 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 centrados en estas técnicas y aún menos estudios experimentales. Recientemente hemos experimentado para analizar la efectividad de las entrevistas estructuradas y no estructuradas. Los resultados se combinaron con otros de estudios experimentales realizados en el campo de Sistemas de Información. Para ello se aplicó el meta-análisis, con el objetivo de desarrollar directrices para usar las entrevistas en IR. Sin embargo, se han obtenido pocas debido a la diversidad, en términos de variables respuesta, de los estudios primarios. Aunque los estudios meta-analizados 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 comprender cómo mejorar el diseño de futuras replicaciones, ejecutadas para generar nuevas evidencias y mejorar resultados de los meta-análisis.

Palabras clave: Ingeniería de requisitos; educción de requisitos; entrevistas; experimentación; diseño experimental; meta-análisis; sesgo; validez.

1 Introducción

La Ingeniería de Requisitos (IR) es una disciplina crucial para el desarrollo de software [1]. 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 a aplicar para elicitarlos ([1], [3], [11]). Típicamente las más utilizadas son las entrevistas [8]. Sin embargo, a pesar de su importancia, existe poca investigación sobre la eficiencia de las entrevistas [8]. Destaca especialmente la escasez de estudios empíricos sobre estas técnicas en IR, mientras que en disciplinas como Psicología o Economía las entrevistas son analizadas empíricamente para estudiar su eficiencia, precisión, roles implicados, etc.

Nuestro trabajo en experimentación centrada en la elicitación de requisitos comenzó con una revisión sistemática de las técnicas de elicitación [7]. Algunos de los trabajos analizados resultaron ser estudios experimentales comparativos acerca de

la eficacia de las entrevistas estructuradas y no estructuradas [2] [4] [12] [14]. Sus resultados resaltan que las entrevistas estructuradas funcionan mejor que las no estructuradas. Sin embargo, nuestra opinión es que es preciso disponer de más guías objetivas, como recomienda la Ingeniería del Software Basada en la Evidencia [10], y esto implica el uso de métodos formales, como el meta-análisis, para agregar diversos estudios experimentales. Dado que el meta-análisis no es adecuado si se dispone de un número escaso de experimentos, se diseñaron y se ejecutaron replicaciones de los experimentos base ([2] [4] [12] [14]), tomados como punto de referencia para adaptar los tipos de entrevistas, variables respuesta y proceso experimental a nuestro contexto: experimento de laboratorio con estudiantes de Informática. Se desarrolló un estudio piloto en 2006 y posteriormente se ejecutaron experimentos anualmente, exceptuando 2009 por razones logísticas: 2007 (analizado y publicado [5], 2008 (pendiente de enviar), y 2010 (pendiente de analizar).

El diseño de replicaciones es una tarea desafiante pues los experimentos originales no están descritos, normalmente, con todos los detalles necesarios para su replicación. Además, también es preciso analizar las variables moderadoras para conocer en detalle el 'focus' del experimento que, en nuestro caso, son las entrevistas. Pero estas variables, usualmente, tampoco están identificadas de forma explícita. Hasta ahora, y según nuestro conocimiento, la única estrategia actual que aborda la replicación es la de [9]. Sin embargo, esta aproximación precisa que las replicaciones sean muy similares, 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[1] aspectos que son amenazas a la validez externa del experimento; por ejemplo, las muestras de conveniencia. Estas amenazas son limitaciones reales y no se puede obtener más información sobre ellas. Sin embargo, en algunos casos los autores incluyen razones que expliquen 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 meta-análisis para combinar resultados de los experimentos base ([2] [4] [12] [14]) y aquellos obtenidos por los autores en la experimentación del 2007, descrito en [5]. 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. La estructura del artículo es la siguiente: en la sección 2 se describen los cuatro experimentos base; la sección 3 detalla nuestro experimento; la cuarta presenta el meta-análisis de los resultados obtenidos; la sección 5 describe el análisis de las limitaciones para identificación de las variables moderadoras, que se usarán para desarrollar, en la sección 6, un posible diseño para nuestro futuro experimento. Finalmente, se presentan las conclusiones en la sección 7.

[1] Por simplicidad, se usa el término 'limitaciones' en lugar de 'limitacion y/o amenazas a la validez interna y/o externa', aunque los autores son conscientes de sus diferencias. Está pendiente el desarrollo de la terminología adecuada.

2 Antecedentes

En la literatura de disciplinas de investigación como Psicología, Economía, etc., se encuentran muchos trabajos acerca de las entrevistas. En ámbitos relacionados con el desarrollo de software también existen algunos trabajos, sin bien en número mucho menor. Centrándonos en aquellos estudios empíricos sobre las entrevistas, los más representativos son los de Agarwal-Tanniru [2], Browne-Rogich [4], Marakas-Elam [12] y Pitts-Browne [14]. Para cada uno se detallan el ámbito, objetivo, diseño, participantes y proceso. Todos ellos analizan alguna hipótesis acerca de la efectividad de las entrevistas, que se estudia en función de la experiencia de los entrevistadores.

[2] llevaron a cabo un experimento en el ámbito de sistemas expertos para 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 (citado en [2]). El diseño experimental aplicado era un factor único aleatorio, con asignación también aleatoria, con tres niveles de tratamiento: 1) ingenieros de conocimiento novatos aplicando entrevista no estructurada, 2) ingenieros novatos con formación en entrevista estructurada y 3) ingenieros expertos aplicando entrevistas no estructuradas. Los entrevistadores fueron estudiantes graduados (grupos 1 y 2) y profesionales (grupo 3). Los entrevistados fueron 30 expertos. Las sesiones, de 90 minutos máximo, se grabaron. El análisis de contenidos de las transcripciones lo realizaron dos codificadores independientes.

[4] experimentaron en el campo de Sistemas de Información acerca de la utilidad de las técnicas de elaboración de preguntas: 1) "características de las tareas", que incluyen tipo de cuestiones sustantivas (para elicitar tipos específicos de requisitos) y cuestiones procedurales (diseñadas para evitar ciertos obstáculos cognitivos); 2) técnicas sintácticas, que se basan en el uso de cuestiones básicas (qué, quién, dónde, cuándo, porqué y cómo) y 3) técnicas semánticas, en las que las preguntas se clasifican según sean eventos, estados o condiciones, acciones, agentes y objetivos. El diseño experimental fue un estudio completamente aleatorio con asignación aleatoria a 3 grupos de tratamiento, uno por tipo de técnica. Los entrevistados fueron 45 trabajadores universitarios, no docentes y sin experiencia en el desarrollo de software. Todos ellos asumieron el papel de un directivo de un almacén de alimentación que precisa un sistema web de compra de comida. El entrevistador fue una persona que desconocía (ciega) las hipótesis de este estudio. Este sujeto interactuó con cada uno de los 45 entrevistados en un entorno de laboratorio. Todas las sesiones se grabaron y se transcribieron para el análisis y clasificación de los requisitos por un codificador independiente. Otro codificador realizó la misma tarea sobre una muestra aleatoria de un 20% de las transcripciones, para asegurar la fiabilidad.

[12] llevaron a cabo su experimento en el ámbito de los Sistemas de Información, con el objetivo de investigar la efectividad de una técnica de entrevista semántica y la consiguiente correctitud de los DFD (Diagrama de Flujo de Datos) de un sistema de compra de productos. El diseño experimental fue un 2x2 con 10 sujetos en cada grupo. 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 entrevistadores con poca experiencia fueron estudiantes y los de alta experiencia profesionales desarrolladores de software. Todos

fueron asignados aleatoriamente en base a su experiencia. Los entrevistados fueron 4 profesionales externos, asignados de tal manera que cada uno fue entrevistado el mismo número de veces en cada grupo. Se pretendía evitar sesgos derivados de los entrevistados, en relación a sus idiosincrasias personales y forma de responder.

Antes de realizar las entrevistas los grupos experimentales recibieron formación sobre la técnica semántica que deberían usar. Las entrevistas se realizaron en una única sesión. Si los sujetos experimentales no usaron la técnica que tenían asignada (semántica o no estructurada) en un 70% del tiempo de la entrevista, ésta se anulaba. En estos casos, se recurrió a otros entrevistadores que sí cumpliesen esta condición. Después de la entrevista, cada sujeto elaboró el correspondiente DFD. Las preguntas realizadas durante la entrevista fueron incluidas en una matriz (tipos de entrevistas X 18 categorías semánticas de la taxonomía de Graesser (citado en [12])). Estas matrices fueron desarrolladas por 3 codificadores independientes. A continuación se compararon las matrices de los 3 codificadores, resolviendo los desacuerdos existentes. Otro codificador realizó la comparación de todos los DFD de los sujetos contra un DFD-solución correcto, para obtener una calificación final. El investigador principal participó en la resolución de las discrepancias y desacuerdos. Otro codificador independiente realizó las mismas tareas sobre una muestra de 20 de las 40 calificaciones, con el fin de verificar la fiabilidad de las calificaciones sobre la solución correcta. Un análisis similar se realizó para calificar los DFD de los sujetos pero atendiendo esta vez a los elementos del DFD no existentes pero posibles. El análisis de datos demográficos de cada grupo de tratamiento revelaron que no existían diferencias significativas entre los grupos que pudiesen influir en los resultados.

Por último, [14] 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. Estos autores plantearon el análisis de las entrevistas para estudiar el punto en el que los requisitos se han elicitado y se puede continuar con el siguiente proceso en el desarrollo de software. Para ello identificaron diversas estrategias de parada. 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.

En este experimento participaron 54 entrevistadores, o analistas profesionales con al menos dos años de experiencia. En este caso interesaba que la experiencia fuese mayor pues favorece que el analista haya desarrollado conocimiento heurístico sobre las condiciones de parada. El entrevistado fue una única persona sin relación con el experimento y que, sin conocer las hipótesis del estudio, asumió el papel de un directivo de un almacén de alimentación que precisa un sistema web de compra de comida (mismo problema que [4], aunque el rol es el inverso: en este caso es el entrevistado y en [4] era el entrevistador). Todas las sesiones se grabaron y se transcribieron para el análisis y clasificación de los requisitos por un codificador independiente. Otro codificador realizó la misma tarea sobre una muestra aleatoria de un 10% de las transcripciones, para asegurar la fiabilidad. Para analizar los datos se determinó la estrategia de parada que cada entrevistador aplicó y, en base a eso, se agruparon para aplicar la ANOVA (ANalysis Of VAriance, citado en [14]). La estrategia de parada predominante se obtuvo 1) de las transcripciones y 2) de un cuestionario con el que se valoraba la experiencia del entrevistador.

3 Experimento realizado

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.

3.1 Descripción del experimento

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

1) Entrevistas no estructuradas, caracterizadas por preguntas abiertas y genéricas que no requieren preparación previa.
2) Entrevistas independientes del contexto, que son un tipo de entrevista estructurada que se caracteriza por un conjunto concreto de cuestiones genéricas centradas en el aspecto a analizar que, normalmente, se utilizan en la primera sesión de entrevistas del proceso de elicitación.

Se aplicó un diseño factorial 2x2 con medidas repetidas y con dos factores (tipo de entrevista y tipo de problema) y sólo una única variable respuesta: el número de requisitos identificados por los sujetos experimentales. Éstos fueron 13 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. La Tabla 1 muestra los datos de su experiencia, recopilados en un cuestionario demográfico. Dos de los estudiantes no detallaron sus datos y, por tanto, en esta tabla aparece N con el valor 11, en vez de 13. De todas formas, la *salience* del experimento, esto es, el interés de los sujetos en ejecutar en forma debida la tarea, estaba asegurada ya que el experimento era una de las prácticas evaluables del curso de Requisitos que cursaban en aquel momento.

Tabla 1. Datos descriptivos de los sujetos experimentales.

	N	Mínimo	Máximo	Media	Desv. típ.
Experiencia total en años	11	3	16	7,55	3,588
Experiencia como analista en años	11	0	8	3,00	2,449

Todos los 13 estudiantes asumieron el rol de ingenieros de requisitos (o entrevistadores) durante la sesión de elicitación. 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. Estos problemas se plantearon para que uno fuese totalmente desconocido para los estudiantes y el otro aparentemente conocido, si bien los estudiantes no conocen el funcionamiento interno de un departamento universitario.

A cada sujeto se le asignó el tipo de entrevista a aplicar (estructurada o no estructurada) mediante insaculación y, por lo tanto, de modo aleatorio. La asignación fue ciega para los experimentadores y entrevistados hasta el análisis de los datos.

La experimentación comenzó con la planificación de las diversas sessiones, para asignar cada par [estudiante/entrevistador-técnica] a cada [entrevistado-problema]. Las sesiones se planificaron a lo largo de 3 días: los días 1 y 2 para el problema de reciclado y los días 2 y 3 para el otro. Cada sesión duró, como máximo, 30 minutos.

No obstante, dado que algunos estudiantes no acudieron puntualmente a sus citas, fue necesario replanificar y, por tanto, la carga de trabajo, para cada problema, del 1er día fue menor mientras que aumentó la del 3er día. Las entrevistas se realizaron individualmente y fueron grabadas. Cada entrevistador recibió el fichero audio de su entrevista y realizó la transcripción, previo paso al desarrollo de la lista de requisitos software. Estas listas, al igual que lo ficheros audio originales, han sido utilizados para analiar los datos y extraer el número de requisitos de cada problema. Se utilizó el análisis de varianzas (ANOVA) para analizar los resultados de la experimentación.

3.2 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.

Así descrita, parece que la experimentación fue un proceso sencillo de ejecutar y sin complicaciones. Sin embargo, hubo determinados factores que tuvieron gran influencia tanto en el proceso como en los resultados. Uno de los más influyentes fue la aparición del cansancio en los entrevistadores a lo largo del 2º 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 a estudiar", entre la 1ª y la 2ª sesión. En teoría se puede pensar que en la 2ª sesión se puede producir este tipo de efecto pero 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 2ª sesiones de entrevistas y también influido por la experiencia de cada entrevistador en *role-playing*. Este efecto el reanálisis de los datos reaplicando la ANOVA pero teniendo en cuenta sólo las 1ª sesiones de ambos problemas. El resultado de H1 se reafirma pero el de H2 se cambió, pues el análisis inicial indicaba que el tipo de problema no afecta a la eficacia.

Otros resultados muy interesantes se obtuvieron del análisis de los datos demográficos de cada sujeto experimental contrastado con su efectividad. Así, la experiencia del analista correla muy fuertemente con la efectividad del proceso de elicitación de requisitos. Es decir, los analistas experimentados extrajeron más información que los analistas con menor o nula experiencia, teniendo en cuenta la lista de requisitos finalmente entregados.

Respecto a las amenazas a la validez, hemos detectado varias (ver [5] para más detalle) pero sólo una de ellas puede tener influencia en las hipótesis planteadas. Concretamente, se refiere al hecho de que se pueda haber confundido el problema 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.

4 Información proporcionada por el conjunto de las replicaciones experimentales

El conjunto de replicaciones experimentales proporcionan la información mostrada en la Tabla 2.

Tabla 2. Información proporcionada por el conjunto de las replicaciones experimentales.

			[2]	[4]	[5]	[12]	[14]
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	

En lo referente a los factores, el conjunto de replicaciones ha estudiado preferentemente dos de ellos: la técnica de elicitación [2] [4] [5] y la experiencia de los participantes en el proceso [2] [12] [14]. El tipo de problema sólo ha sido explorado en [5].

En lo referente 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 [4]. Otro gran subgrupo de variables respuesta es el referente a los contenidos. Estas variables recogen, por ejemplo, los errores lógicos [12] o la diferencia cualitativa entre la información obtenida por las distintas técnicas de elicitación [4]. Por último, se han estudiado también un conjunto de variables disímiles que hemos agrupado bajo el nombre genérico de *Otras*.

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 mediante meta-análisis, tanto desde la perspectiva de la técnica de elicitación como de la experiencia de los sujetos (zona sombreada en la Tabla 2). También puede estudiase el efecto conjunto de las técnicas y la experiencia de los sujetos, aunque de forma muy limitada (considerando únicamente los experimentos [2] [12]). Las demás variables respuesta no pueden ser agregadas.

No obstante, el problema que para nosotros es realmente relevante es lo limitado del área de investigación. Aparte de las técnicas y de la experiencia, sólo un factor adicional ha sido estudiado (el tipo de problema) y sólo en un caso ([5]). Desde nuestra perspectiva, el marco de investigación parece demasiado simple, esto es, no representativo de la complejidad de los problemas reales. Creemos necesario que los experimentos incrementen el número de factores considerados para alinear la investigación y las necesidades prácticas de los analistas.

5 ¿Cuál es la contribución de nuestro experimento?

En términos de resultados obtenidos, el experimento que se llevó a cabo aporta pocas novedades. Es decir, contribuye con un conjunto de resultados que amplían el número de factores (aparte de la experiencia se añade el tipo de problema) y que aportan unas pocas evidencias más sobre la efectividad de las entrevistas en el proceso de elicitación de requisitos en la IR. 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. Este análisis de las limitaciones comienza con la Tabla 3, que muestra las limitaciones identificadas por [2] [4] [5] [12] [14]. Cada experimento identifica entre 4 y 5 limitaciones, por término medio, salvo [4] que sólo detalla dos.

Tabla 3. Número y descripción de *todas las limitaciones* por experimento.

	2	**4**	**5**	**12**	**14**	**Total**
Proceso	3 Sólo un dominio de problema Sin estándar de referencia Sólo una sesión	-	1 Problemas no complejos	2 Entorno de laboratorio Estudio exploratorio	1 Sólo una estrategia de elicitación	7
Muestra	2 Muestra de conveniencia Role playing	1 Experiencia previa de los entrevistadores	2 Entrevistado vs. problema Los 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 Sólo 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 Sólo un método de representación	2 Sólo una medida de las reglas de parada cognitivas Codificación centrada en una taxonomía de requisitos predefinida	6
Total	5	2	4	7	6	24

Las limitaciones se han clasificado en función del aspecto al que hacen referencia. Por ejemplo, cuando [2] indica que "los expertos constituyen una muestra de conveniencia, en vez de una muestra aleatoria, aunque se haya aplicado la asignación aleatoria de 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 puedan ser utilizadas para clasificar las limitaciones en otras áreas, como las pruebas del software.

Analizando las limitaciones por categorías (filas), se pueden obtener las coincidencias, 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, con todo lo que esto implica, como describe [12]. Todas las restantes limitaciones del *Proceso* se centran en aspectos concretos, como el número de sesiones [2] o la complejidad de las tareas experimentales [4].

La categoría más frecuentemente identificada como limitación en los cinco experimentos es la *Muestra*, en la que se incluyen los entrevistadores, entrevistados, 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 [12] o que sean muestras de conveniencia [2]. 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 sólo uno en el entrevistado [5].

El criterio menos relevante, según la Tabla 3, es la *Técnica*. Centrándonos en las técnicas de elicitación aplicadas en las entrevistas, [12] y [5] presentan posturas opuestas, en cuanto si se acepta o no la variabilidad en la aplicación de las técnicas estructuradas. Esta diferencia sólo implica que son dos posibles aproximaciones a la experimentación, dependiendo del control de estas técnicas de elicitación, pero esto no invalida estos experimentos. Son sólo características de los diseños experimentales desarrollados por esos autores. Otro tipo de técnicas son aquellas utilizadas para representar los datos elicitados, como los DFD de [12] o aquellas usadas para codificar los datos extraídos con las técnicas de elicitación [4].

Aparte de estas coincidencias y frecuencias, es más interesante un análisis de cada limitación mostrada en la Tabla 3 para determinar si es una amenaza a la validez (es decir, aspectos que impiden extrapolar los resultados experimentales a poblaciones más amplias) o una limitación en sentido estricto (denominadas, a falta de otra terminología, *limitación efectiva*), que serán tratadas a continuación. Este análisis es crucial para nuestro objetivo, porque las amenazas a la validez son restricciones metodológicas que no pueden ser usadas para identificar las variables moderadoras.

Por ejemplo, [12] identifica como limitación que su estudio es un experimento de laboratorio. Obviamente, este tipo de experimentos son limitados, si nos referimos al tipo de conocimiento que se puede obtener de ellos, por varias razones (entorno estrictamente controlado, condiciones ideales, etc.), pero esto no implica la existencia de una variable moderadora. Análogamente, [2] identifica como limitación el hecho de que los sujetos implicados en el experimento sean una muestra de conveniencia. Esto es también una restricción metodológica. Una muestra que no sea de conveniencia debería usarse en una situación ideal. De todas formas esto tampoco apunta a la existencia de una variable moderadora.

Las limitaciones relacionadas con el número de sesiones, número de problemas, complejidad de los mismos y número de técnicas aplicadas están relacionadas con el coste, esfuerzo y 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 a la efectividad de las entrevistas (es decir, una variable moderadora). En la misma línea de razonamiento, las 'limitaciones técnicas' de [5], [12] y [14] se pueden analizar de la misma forma, pues describen particularidades de dichos experimentos que tampoco restringen la generalización de los resultados.

Si se eliminan de la Tabla 3 todos los casos mencionados hasta ahora, se obtiene el conjunto de limitaciones efectivas, mostradas en la Tabla 4. Son limitaciones en sentido estricto, pues indican una falta de validez de los resultados experimentales dentro del contexto de cada experimento (es decir, validez interna). Sin embargo, no son errores de diseño. Por ejemplo, la experiencia de los entrevistadores podría ser una de las limitaciones efectivas porque [4] señala que la experiencia podría afectar a la efectividad de las entrevistas. Por tanto, si en los experimentos no se controla la experiencia de los sujetos, los resultados podrían invalidarse. En otras palabras, la experiencia de los entrevistadores es una potencial variable moderadora.

Tabla 4. Número y descripción de limitaciones efectivas por experimento.

	2	**4**	**5**	**12**	**14**	**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 Sólo 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

Otros ejemplos son la "*Precisión de la codificación*" [4] o la "*Codificación centrada en una taxonomía de requisitos predefinida*" [14]. Se pueden considerar como limitaciones efectivas porque podrían ser el origen de un sesgo de medición, que podría afectar al análisis de las hipótesis. También, las limitaciones agrupadas bajo la categoría *Muestra* podrían considerarse como potenciales 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 sólo aumenta la confidencia de una estimación. Todas estas limitaciones efectivas se incluyen en el siguiente análisis de las potenciales variables moderadoras.

5.1 Identificación de las variables moderadoras

En la literatura sobre experimentos en otras áreas científicas (Economía, Medicina, etc.) las limitaciones efectivas de la Tabla 4 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 experimentos sobre entrevistas analizados comparten una perspectiva similar acerca de las limitaciones mostradas en la Tabla 4. Por ello las incluyen en las secciones *Amenazas a la validez* de sus trabajos. En algunos casos, este proceder está completamente justificado, si la limitación es un sesgo claramente. Por ejemplo, las limitaciones bajo la categoría *Técnicas* de la Tabla

4 son instancias de un *sesgo de medición* (riesgo en la determinación precisa de los valores de variables respuesta). Otro ejemplo es la *Motivación de la muestra*, que es una instancia de un *sesgo de motivación*. La motivación es un prerequisito 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 del alcance de este análisis en la búsqueda de potenciales variables moderadoras.

Sin embargo, en muchos de los otros casos esto no es cierto, especialmente en la categoría *Muestra*, que es la que contiene más elementos en la Tabla 4. 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 procedan de la categoría *Muestra*. En la IR, particularmente en el uso de las entrevistas 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 tanto, estos aspectos no son sesgos o riesgos, sino más bien aspectos que hay que considerar para poder comprender adecuadamente cuándo y cómo funcionan bien las entrevistas. Ésta es la razón por la que se ha identificado este tipo de limitación como potencial variable moderadora.

La siguiente lista muestra una clasificación de las limitaciones de la Tabla 4, teniendo en cuenta los sesgos potenciales a los que pueden dar origen (según otras disciplinas como Economía, Medicina, etc.). No es una lista exhaustiva, pues sólo se incluyen los ejemplos más claros, pero es útil para un análisis inicial:

- Sesgo del artefacto, relacionado con las limitaciones "*Entrevistado vs. Problema*" y "*Experiencia de los entrevistadores vs. dominio del problema*".
- 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 "E*xperiencia previa de los entrevistadores*".
- Sesgo del entrevistado, relacionado con el "*Role playing*" y, según la Tabla 4, con la posible identificación "*Entrevistado vs. Problema*" y el sesgo potencial derivado de la preferencia 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 1) que necesitamos conocer para explicar las razones de la efectividad de las entrevistas y 2) que necesitamos tener en cuenta para llevar a cabo la elicitación en la práctica. Por lo tanto, del anterior conjunto de limitaciones (o sesgos) podemos identificar las siguientes variables moderadoras: Problema, Experiencia y Características personales. Y en base a estas variables moderadoras, se pueden deducir las siguientes recomendaciones:

1ª. Ejecutar entrevistas acerca de diferentes tipos de problemas, de diferente tamaño y complejidad y, preferiblemente, de diferentes dominios.
2ª. 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, y siempre que sea posible y apropiado, además de los años, indagar por el número y tamaño de los proyectos en los que ha participado un sujeto.
3ª. Análisis de los roles asignados a cada sujeto, en base a su experiencia, aptitudes, conocimiento del dominio, etc. Podría recopilarse mediante mediciones concretas

o tests psicológicos para conocer la personalidad del sujeto. No hay que olvidar que existen más roles que los entrevistadores y entrevistados y que también ellos pueden ejercer influencia sobre los resultados finales del experimento.

La aplicación de estas recomendaciones en la práctica 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 tenerse en cuenta en los experimentos sobre entrevistas. En otras áreas de la Ingeniería del Software se podrán encontrar otras variables moderadoras (quizá no aquellas relativas a la experiencia del sujeto). La aplicación de un análisis similar al aquí presentado en aquellos experimentos sobre entrevistas podría ser útil para identificar otras variables moderadoras.

6 Posible Diseño de un Experimento Futuro

Un posible diseño experimental que incluya las variables moderadoras identificadas es el que se presenta a continuación.

- *Ámbito*. Aplicando la 1ª recomendación, la descripción del problema contiene 4 aspectos: tipo, tamaño, complejidad y dominio. En vez de aplicar todas estas posibilidades conjuntamente, que implicaría 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 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 utilizará el mismo 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 2ª y 3ª, en función del tipo de sujeto experimental.
  - Entrevistador: sólo se considera aplicable la 2ª, pues dado que seguirá siendo un experimento de laboratorio con estudiantes. En posteriores experimentos se podría plantear, aplicando la 3ª recomendación, que algunos de ellos asuman el rol de entrevistados. En cuanto a la experiencia de los entrevistadores, se plantea la realización de una encuesta demográfica con el objetivo de conseguir, antes de la realización del experimento, datos sobre la experiencia. Sería una tarea análoga a la ya realizada pero ampliando la información a recabar. No sólo si tiene o no experiencia y, en caso afirmativo, cuánto tiempo. También incluir preguntas sobre el número de proyectos en general y calificación de los tamaños: *x* de tamaño pequeño, *y* de tamaño medio, *z* de tamaño grande.
  - Entrevistado: se pueden aplicar las recomendaciones 2ª y 3ª. Mediante encuesta o entrevista se debe averiguar el tipo de conocimiento que tienen del dominio. Respecto a este rol se puede plantear los siguientes aspectos que se podrían combinar (de menor a mayor dificultad) en sucesivos experimentos:

- Conocimiento del dominio: mismo o dispar.
- Experiencia previa: docente, profesional del desarrollo de software, sin relación alguna con el desarrollo de software.
- Relación con el experimento: ciego, con conocimiento.
- Implicación en el papel: en función de la personalidad y el tipo de respuestas (bombardeo de información vs. escueto).

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 sólo codifica un porcentaje para verificación de la fiabilidad). En el siguiente experimento se aplicará: [entrevistador-total] y [profesional-parcial al50%].

- *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.

7 Conclusiones

En las publicaciones empíricas se tiende a no distinguir entre amenazas a la validez y otras limitaciones efectivas, considerando ambas como restricciones a la validez externa. Pero, probablemente, esto sea una equivocación. Las restricciones metodológicas como la muestra, el tipo o número de sujetos afectan a la validez externa, pero otras limitaciones como la experiencia de los entrevistadores no (en el caso de experimentos acerca de entrevistas).

El análisis de las limitaciones identificadas en los experimentos puede ser una estrategia útil para encontrar variables moderadoras. De esta manera, éstas 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 obteniendo finalmente un posible diseño de futuras 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, influídos por la Ingeniería del Software Empírica, es más fácil considerarlas como posibles variables moderadoras que podrían tener influencia en los resultados de los experimentos. Como ésta, se pueden identificar otros aspectos interesantes, a menudo basados en opiniones expertas expresadas como relaciones causales sin verificar experimentalmente, que pueden representar limitaciones a tener en cuenta en futuras experimentaciones [6]. 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 [15]. Un caso típico es la relación causal informal 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 en no sólo 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 éstas se aplican.

Referencias

1 Abran, A., Moore, J.W., Bourque, P., Dupuis, R., Tripp, L.L.: Guide to the Software Engineering Body of Knowledge (SWEBOK). IEEE (2004)

2 Agarwal, R., Tanniru, M. R.: Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation. Journal of Mng. Inf. Systems, 7, 1, pp. 123-140 (1990)

3 Bell, T.E. Thayer, T.A.: Software Requirements: Are they really a problem?. In Proc. of the 2nd Int. Conference on Software Engineering (ICSE'76) pp. 61-68 (1976)

4 Browne, G. J., Rogich, M. B.: An Empirical Investigation of User Requirements Elicitation: Comparing the Effectiveness of Prompting Techniques. Journal of Manag. Inf. Systems, 17, 4, pp. 223-250 (2001)

5 Carrizo, D., Dieste, O., Juristo, N., López, M. Estudio experimental de la efectividad de la entrevista abierta frente a la entrevista independiente de contexto. In: Lencastre, M. (Ed.). 14th Workshop on Requirements Engineering, pp. 297-308. Rio de Janeiro, Brasil (2011)

6 Carrizo, D., Dieste, O., Lopez, M.: Identifying Moderator Variables Through Requirements Elicitation Experiments Limitations. In Proceedings of the 12th Int. Conference on Product Focused Software Development and Process Improvement. Italy, June 20-22 (2011)

7 Dieste, O., Juristo, N.: Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques. IEEE Transac. on Software Engineering, 37, 2, pp. 283-304 (2011)

8 Hickey, A., Davis, A., Kaiser, D.: Requirements Elicitation Techniques: Analyzing the Gap Between Technology Availability and Technology Use. Comparative Technology Transfer and Society 1, 3, pp. 279-302 (2003)

9 Juristo, N., Vegas, S.: Using Differences among Replications of Software Engineering Experiments to Gain Knowledge. Third Int. Symposium on Empirical Software Engineering and Measurement (ESEM'09) (Florida, Octubre 15-16). IEEE. pp. 356-366 (2009)

10 Kitchenham, B., Dybå, T., Jørgensen, M.: Evidence-based Software Engineering. In Proceedings of the 26th Int. Conference on Software Engineering (ICSE'04) (Edinburgh, UK, May 23-28, 2004). IEEE Computer Society, Washington DC, USA, pp. 273-281 (2004)

11 Leuser, J., Porta, N., Bolz, A., Raschke, A.: Empirical Validation of a Requirements Engineering Process Guide. In Proceedings of the 13th Int. Conference on Evaluation and Assessment in Software Engineering (EASE'09) (UK, April 20-21), pp. 1-10 (2009)

12 Marakas, G.M. Elam, J.J.: Semantic Structuring in Analyst Acquisition and Representation of Facts in Requirements Analysis. Information Systems Research, 9, 1, pp. 37-63 (1998)

13 Miller, J.: Applying Meta-Analytical Procedures to Software Engineering Experiments. Journal of Systems and Software, 54, pp. 29-39 (2000)

14 Pitts, M.G., Browne, G.J.: Stopping Behavior of Systems Analysts During Information Requirements Elicitation. Journal of Mng. Inf. Systems, 21,1, pp. 203-226 (2004)

15 Swanson, D.R.: Two Medical Literatures that are Logically but not Bibliographically Connected. American Society for Information Science, 38, 4, pp. 228-233 (1987)