Uso de Meta-Análisis para Integrar Resultados Experimentales

Mª Esperanza Manso[1], José A. Cruz-Lemus[2], Marcela Genero[2], Mario Piattini[2]

[1]Grupo de investigación GIRO, Departamento de Informática, Universidad de Valladolid.
Campus Miguel Delibes, E.T.I.C., 47011, Valladolid, España.
manso@infor.uva.es
[2]Grupo de investigación ALARCOS, Departamento de Tecnologías y Sistemas de Información.
Universidad de Castilla-La Mancha, Paseo de la Universidad, 4. 13071 Ciudad Real, España.
{JoseAntonio.Cruz, Marcela.Genero, Mario.Piattini}@uclm.es

Resumen. El objetivo principal de este artículo es presentar diferentes técnicas estadísticas, en especial el meta-análisis, que permiten integrar los resultados obtenidos en familias de experimentos y realizar revisiones de estudios con hipótesis comunes de una forma objetiva y sistemática. Aunque estas técnicas son mejores que las revisiones subjetivas de estudios no están exentas de riesgos, ni son aplicables en todas las condiciones. Además veremos cómo se está utilizando el meta-análisis en la Ingeniería del Software empírica, y como caso práctico aplicaremos esta técnica en una familia de cinco experimentos controlados realizados en entornos académicos, cuyas hipótesis investigan por un lado la correlación entre la complejidad estructural y el tamaño de los diagramas de clases UML y su complejidad cognitiva y, por otro, la correlación entre la complejidad cognitiva y la entendibilidad y modificabilidad de dichos diagramas. Los resultados obtenidos tienen implicaciones tanto desde el punto de vista del modelado como el de la enseñanza, ya que muestran evidencia empírica sobre cuáles son los constructores UML que tienen más influencia cuando los modeladores tienen que comprender y modificar los diagramas de clases UML.

Palabras Clave: Meta-análisis, Ingeniería del Software Empírica, experimentación.

1. Introducción

Si consideramos la ciencia como un medio para acumular y refinar tanto la información como el conocimiento que ésta contiene, entonces es esencial establecer líneas directrices objetivas y fiables de cómo integrar y sintetizar los resultados procedentes de diferentes investigaciones, pero con objetivos similares. Existen métodos subjetivos, dependientes del conocimiento y la experiencia particular de los experimentadores, para decidir qué información es relevante incluir y con qué peso se valorará. Frente a ellos existen técnicas cuantitativas que permiten sintetizar objetivamente investigaciones individuales [35]. El problema de la experimentación es que difícilmente los resultados de un solo estudio son definitivos, por lo que es necesario hacer réplicas [1; 23]. Cuando el investigador quiere sintetizar los resultados de experimentos con hipótesis similares, los métodos cuantitativos que proporciona la estadística definen un marco para obtener una estimación del efecto del tratamiento (tamaño del efecto) a través de los experimentos, conociendo el tamaño del efecto de cada uno de ellos, de forma objetiva y rigurosa. Los beneficios que proporcionan estas técnicas es que permiten detectar los efectos del tratamiento aunque sean pequeños. Si nos referimos a la Ingeniería del Software Empírico (ISE), detectar una disminución mínima de tiempo en la realización de una tarea que se repite varias veces al día por muchas personas puede suponer un ahorro importante. Desde ese punto de vista el meta-análisis es una bala de plata para el investigador que le permite obtener conclusiones del esfuerzo realizado en decenas de experimentos por otros investigadores.

Pero estas técnicas no son la panacea, puede haber problemas por no aplicarlas debidamente que invalidan los resultados. Los más críticos con estas técnicas hablan de "comparar manzanas y naranjas", y Glass [13] contesta que tienen parte de razón en eso, pero que efectivamente comparar manzanas con manzanas es trivial. El uso de estas técnicas en ISE está aún pendiente, queda camino por recorrer en cuanto a réplicas de experimentos que aporten información sobre las hipótesis con las que se trabaja en la ingeniería del software [19;25].

El resto del documento está estructurado de la siguiente forma, en la sección 2 veremos cuales son las principales técnicas cuantitativas para sintetizar diferentes experimentos, en la sección 3 se describen las limitaciones de estas técnicas y las amenazas a la validez de los resultados cuando no se utilizan adecuadamente. En la sección 4 revisaremos los estudios de meta-análisis realizados en ISE, y en la 5 aplicaremos el meta-análisis en una familia de 5 experimentos que estudia la relación entre la complejidad estructural y el tamaño de los diagramas de clases UML y su complejidad cognitiva y, entre ésta y la entendibilidad y modificabilidad de dichos diagramas. Para finalizar en la sección 6 expondremos las conclusiones.

2. Técnicas Estadísticas para Sintetizar Experimentos

" El meta-análisis se refiere al análisis de análisis… al análisis estadístico de una colección de resultados procedentes de estudios individuales, y cuyo propósito es integrar dichos resultados. Supone una alternativa rigurosa frente a las discusiones meramente narrativas sobre los resultados de una colección de estudios …" [12]

Nos referiremos aquí a algunos de los métodos estadísticos que permiten acumular e interpretar resultados a través de diferentes experimentos, relacionados por investigar hipótesis similares, acumulando así su conocimiento [15; 12; 28; 35; 33].

Existen fundamentalmente tres modos de abordar la síntesis de resultados a través de experimentos individuales, utilizados en otras áreas como medicina o psicología: meta-análisis, combinación de niveles de significación y recuento de votos. A continuación describiremos estas técnicas, las dos primeras con algo más de detalle pues son las que se utilizan con frecuencia en ISE.

2.1. Meta-Análisis

El Meta-análisis tiene como objetivo obtener un tamaño del efecto global, entendiendo como tamaño del efecto "El grado en que el efecto del tratamiento está presente en la población" [5], concepto relevante cuando se habla de potencia de los test estadísticos. Por eso esta técnica se basa en obtener medidas estandarizadas del tamaño del efecto en cada experimento para después combinarlas, obteniendo así una estimación media del tamaño del efecto global. Existen diferentes modos de realizar esa combinación, el más utilizado consiste en ponderar las medidas estandarizadas por una función del tamaño muestral de cada estudio o bien de las varianzas [15; 25]. Una vez calculada la media del tamaño del efecto se interpreta ese valor bien proporcionando un intervalo de confianza, o un p-valor, que permite decidir sobre las hipótesis del meta-análisis.

El meta-análisis comprende tres pasos principales:

1. Decidir qué estudios incluir en él.
2. Estimar el tamaño del efecto para cada estudio.
3. Combinar los tamaños del efecto de cada estudio para estimar el tamaño del efecto global, que puede tener una expresión genérica como en la ecuación 1, y contrastar con él las hipótesis de interés.

$$efecto\_global = \sum_i w_i efecto_i \quad \textbf{(1)}$$

Los dos primeros pasos pueden ser causa de problemas que revisaremos en la sección 3, y en cuanto al tercero, primero hay que elegir el Modelo del Tamaño del Efecto, si es fijo o aleatorio, pues en este último caso hay que añadir una fuente más de variabilidad [24; 8]. Existen dos tipos fundamentales de estimadores del tamaño del efecto, relacionados con la hipótesis que interesa investigar: diferencias entre dos grupos (diferencias de medias) o grado de asociación entre dos variables (coeficiente de correlación). Desde el punto de vista teórico algunos de estos estimadores son equivalentes conceptualmente, pues hay una forma de obtener cada uno a partir de los otros [15; 12;35].

En el ejemplo de la sección 5 vamos a utilizar la *g de Hedges* definida en [15], que puede adoptar diferentes expresiones, todas con las mismas propiedades asintóticas, pero con pesos (w_i) diferentes para ajustar el sesgo cuando se trabaja con muestras pequeñas.

Uno de los mayores problemas que se pueden encontrar al trabajar con estos estimadores tiene que ver con las suposiciones teóricas que se hacen sobre los tipos de distribución (Normal) y la homogeneidad de las varianzas, aunque existen estimadores robustos para soslayar algunos, hay que tenerlos en consideración. La herramienta con la que hemos trabajado [2] permite elegir las diferentes opciones que hemos mencionado.

2.2. Combinación de Niveles

La Combinación de niveles de significación, como su propio nombre indica, permite obtener una medida o estadístico que resume los diferentes niveles de significación (α) de los experimentos. Según sea la estimación obtenida podemos extraer ciertas conclusiones sobre la significación de la hipótesis a través de todos los experimentos. Existen diferentes estadísticos que combinan los niveles de significación, entre ellos los de Fisher basados en una ji-cuadrado, el de Winer basado en una t-student o el de Stouffer similar al anterior pero basado en una distribución Normal [25; 28; 35].

La ventaja es que en esta técnica no hay dependencia del tipo de datos de cada experimento ni de su distribución, pues considera el tamaño del efecto. A cambio su interpretación puede ser difícil o errónea, pues si el nivel de significación global es significativo eso no implica que se acepta la alternativa en cada estudio, sólo que en alguno sí se acepta. Luego realmente no proporciona información sobre la existencia de efecto a través de los estudios.

2.3. Recuento de Votos

El Recuento de votos es una técnica que se basa en obtener una información global sobre los resultados "positivos", "negativos" o "neutros" de cada experimento que se está considerando. Necesitamos conocer, además de las hipótesis, al menos el p-valor o el nivel de significación (α) de los experimentos para poder clasificarlos como positivos si el resultado es significativo, negativo si no los es y neutro en otro caso. La técnica se basa en evaluar la proporción de positivos, rechazando la hipótesis nula, de no efecto del tratamiento, si esa proporción sobrepasa un nivel [15; 25; 33].

3. Amenazas y limitaciones de la síntesis de experimentos

Los resultados obtenidos cuando se sintetiza la información procedente de diferentes experimentos pueden carecer de validez y fiabilidad si no se tienen en cuenta los riesgos y las limitaciones de los mismos, detallados, entre otros, en [21; 35]. Podemos clasificar las amenazas a la validez de estos métodos en cuatro categorías:

1. Los estudios individuales difieren en cuanto a técnicas de medida, definición de variables y tipos de sujetos.
2. Los estudios incluyen experimentos con diseños de alta y baja calidad.

3. Los resultados se pueden sesgar si no se tienen en cuenta estudios cuyos resultados son no significativos, pues éstos no suelen publicarse, lo que afectará sobre todo cuando se hacen revisiones de publicaciones..
4. Cuando del mismo estudio se extraen múltiples resultados, las conclusiones pueden estar sesgadas y parecer más fiables de lo que son porque no hay independencia entre los resultados.

Cuando disponemos de los datos de cada experimento podemos plantearnos trabajar con todos ellos juntos, para así extraer conclusiones "mejores" que estudiándolos por separado, sin embargo eso no es siempre cierto, como lo pone de manifiesto la paradoja de Simpson. Esta paradoja se produce porque la dirección del efecto del tratamiento en cada experimento individual se invierte cuando las muestras se mezclan sin tener en cuenta una variable que influye en los resultados. Más información sobre los problemas del uso y abuso de estas técnicas se pueden encontrar en [4; 13; 24], y en [19; 22; 23] donde se abordan los problemas propios de su aplicación en la ISE. A modo de ejemplo, en [25] abordan el problema de generalizar resultados en ISE, debido a la forma en que usualmente se seleccionan los sujetos y objetos, sin aleatorizar, de la población "general" correspondiente.

4. Uso del Meta-análisis en ISE

En Ingeniería del Software estas técnicas se han utilizado bien para extraer conclusiones en familias de experimentos perfectamente planificados, lo que evita una gran parte de amenazas a la validez de las conclusiones, o bien para realizar revisiones sistemáticas [19] de experimentos publicados sobre ciertos temas:

1. Uso en Familias de experimentos: En [22] se sintetizan los resultados de un estudio sobre el efecto que tenía una herramienta sobre la eficacia y eficiencia de las inspecciones; los tamaños del efecto se obtuvieron a partir de los coeficientes de correlación. Para sacar conclusiones sobre experimentos que evaluaban diferentes técnicas de inspección se utilizó esta técnica en [20;26]. En [14] también se utiliza esta técnica para sintetizar los resultados de un experimento y 4 réplicas que evaluaban técnicas de inspección. En [7] se realiza un meta-análisis para estudiar el efecto de la Programación por Parejas (*Pair Programming*) en la calidad, duración y esfuerzo, el tamaño del efecto se mide con una diferencia de medias, y en [29] se trata de sintetizar diferentes experimentos sobre técnicas de lectura.
2. Uso en revisiones sistemáticas: algunas en [18] se revisan publicaciones de ISE para estudiar el uso que se hace de la potencia de los test estadísticos; en [30] se hace una revisión de 100 proyectos en JAVA para estudiar y sintetizar los resultados individuales sobre el estudio de correlaciones entre diferentes métricas.

5. Aplicación del meta-análisis a una familia de experimentos

Vamos a presentar en esta sección una aplicación del meta-análisis a una familia de experimentos que describiremos en la sección 5.1, en la sección 5.2 presentaremos los resultados de aplicar dicha técnica.

5.1. Una familia de experimentos

Los cinco experimentos controlados se realizaron teniendo en cuenta algunas recomendaciones propuestas en [17, 34]. La Tabla 1 resume las principales características de contexto de los cinco experimentos.

Tabla 1. Características de la familia de experimentos

	#Sujetos	Universidad	Fecha	Curso

E1	72	Universidad de Sevilla (España)	Marzo 2003	4º
R1	28		Marzo 2003	
E2	38	Universidad de Castilla-La Mancha (España)	Abril 2003	3º
R21	23	Universidad de Sannio (Italia)	Junio 2003	4º
R22	71	Universidad de Valladolid (España)	Septiembre 2005	3º

5.1.1 Planificación de los experimentos

En esta subsección definiremos las características comunes a todos los experimentos, consistentes en:

Preparación. La familia tiene un doble objetivo, definido como:

- Objetivo 1: analizar la complejidad estructural y el tamaño de los diagramas de clases UML con respecto a su relación con la complejidad cognitiva desde el punto de vista de los modeladores y diseñadores de software en un contexto académico.
- Objetivo 2: analizar la complejidad cognitiva de los diagramas de clases UML con respecto a su relación con la entendibilidad y modificabilidad desde el punto de vista de modeladores o diseñadores software en un contexto académico.

Definición del contexto: en estos estudios hemos utilizado estudiantes como sujetos experimentales. Las tareas a realizar no requerían altos niveles de experiencia industrial, así que creímos que estos sujetos podían considerarse apropiados, como se apunta en varios trabajos [3, 16]. En resumen, trabajar con estudiantes implica una serie de ventajas, como el hecho de que el conocimiento previo es bastante homogéneo, la disponibilidad de un elevado número de sujetos, y la posibilidad de probar diseños experimentales e hipótesis iniciales [31]. Una ventaja más es el uso de principiantes como sujetos en experimentos en entendibilidad es que la complejidad cognitiva del sujeto bajo estudio no es tan alta como la de sujetos con experiencia.

Material: los materiales experimentales consisten en un conjunto de diagramas de clases UML que cubren un amplio rango de valores las medidas presentadas en la Tabla 5 del Apéndice A. Así, obtuvimos tres tipos de diagramas: difíciles de mantener (D), fáciles de mantener (F) y de una dificultad mediana (M). Algunos fueron diseñados específicamente para los experimentos y otros se obtuvieron de aplicaciones reales. Cada diagrama tiene documentación añadida que contiene, además de otras cosas, cuatro tareas de comprensión y cuatro de modificación.

5.1.2 Conducción de los experimentos individuales

Las variables consideradas para medir la complejidad estructural y el tamaño fueron un conjunto de 11 medidas presentadas en la Tabla 5 (Apéndice A). La medida *CompSub* es la percepción subjetiva de los sujetos sobre la complejidad de los diagramas con los que trabajaron durante la tarea experimental. Nosotros consideramos *Compsub* como una medida de la complejidad cognitiva. Los posibles valores de esta variable son: Muy sencillo, Moderadamente sencillo, Medio, Moderadamente complejo y Muy complejo. Para medir la entendibilidad y modificabilidad de los diagramas consideramos el tiempo (en segundos) utilizado por cada sujeto para completar las preguntas de entendibilidad y modificabilidad. Hemos llamado a estas medidas Tiempo de entendibilidad y modificabilidad.

Utilizamos un diseño balanceado inter-sujetos, así cada sujeto trabajó con un único diagrama. Los diagramas se asignaron aleatoriamente y cada diagrama fue asignado al mismo número de sujetos.

Formulamos las siguientes hipótesis, que se derivan de los objetivos de la familia, previamente presentados:

- $H_{0,1}$: la complejidad estructural y el tamaño de los diagramas de clases UML no está correlacionado con la complejidad cognitiva. $H_{1,1}$:$\neg H_{0,1}$
- $H_{0,2}$: La complejidad cognitiva de los diagramas de clases UML no está correlacionada con su entendibilidad y modificabilidad. $H_{1,2}$:$\neg H_{0,2}$

Todos los experimentos fueron supervisados y limitados en el tiempo. Se pueden encontrar más detalles en [9, 10].

Utilizamos SPSS [32] para realizar todos los análisis estadísticos y la herramienta Comprehensive Meta Analysis [2] para realizar el meta-análisis.

5.1.2.1 Experimento 1 (E1) y réplica (R1)

Hemos estudiado las hipótesis utilizando el coeficiente de correlación de Spearman, obteniendo las siguientes conclusiones relacionadas con los dos objetivos de la familia de experimentos:

- Objetivo 1: la complejidad estructural y la complejidad cognitiva presentan una correlación positiva significativa para todas las medidas, con la excepción de NM, NGEN y MAXDIT en R1.
- Objetivo 2: la complejidad cognitiva parece estar correlacionada positivamente con el esfuerzo necesario para comprender los diagramas de clases UML, pero los resultados son significativos sólo para E1. Al mismo tiempo, no hay correlación con el esfuerzo necesario para modificar los diagramas. Una posible explicación para esto puede ser que los sujetos basen su percepción en la dificultad de la primera prueba que realizan, que en este caso es la de comprensión.

5.1.2.2 Experimento 2 (E2) y sus réplicas (R21 y R22)

En estos estudios, los objetivos y variables fueron los mismos que en los descritos previamente, pero los diagramas utilizados fueron diferentes, y el contexto y diseño fueron mejorados. Se puede encontrar información más detallada en [10].

De nuevo, hemos utilizado a sujetos elegidos por conveniencia, pero en este caso se ha mejorado la selección bloqueando la experiencia de los sujetos. Realizamos un test previo, y con los resultados obtenidos dividimos a los sujetos en dos grupos. Cada diagrama fue asignado al mismo número de sujetos en cada grupo. Se pueden encontrar más detalles de este proceso en [10].

Las medidas de entendibilidad y modificabilidad sólo se incluyeron cuando la realización de las tareas tenía un nivel de calidad mínimo (corrección y completitud). Los sujetos que estuvieron por debajo del 75% en corrección y completitud fueron excluidos del estudio.

Después de comprobar las hipótesis formuladas y relacionándolas con los objetivos de la familia de experimentos, concluimos que:

- Objetivo 1: tenemos resultados favorables que admiten una correlación entre la complejidad estructural y la complejidad cognitiva de los diagramas de clases UML. Muchas de las medidas están significativamente correlacionadas con la complejidad subjetiva en diferentes estudios, especialmente en aquellos relacionados con las jerarquías de herencia (ver Tabla 2).
- Objetivo 2: los resultados también están a favor de la hipótesis que relaciona la complejidad cognitiva con la entendibilidad de los diagramas de clases UML.

Tabla 2. Medidas correlacionadas en E1, R21 y R22

Estudio	Medidas correlacionadas significativamente
E2	NC, NAssoc, NGen, NGenH, MaxDIT (5 sobre 11 medidas)
R21	Todas excepto NM, NGenH y MaxAgg (8 sobre 11 medidas)
R22	Todas excepto NM (10 sobre 11 medidas)

5.2 Estudio de Meta-análisis

Hemos utilizado el meta-análisis para extraer conclusiones globales de la familia de experimentos presentada anteriormente. Como ya hemos mencionado, necesitamos estandarizar los tamaños de los efectos y para ello debemos elegir un estadístico que los mida, en este caso hemos utilizado coeficientes de correlación para, a partir de ellos, obtener la métrica del efecto global medida con la *g de Hedges* adecuada cuando los tamaños muestrales no son grandes. Además en [18] se clasifica el tamaño del efecto global

en pequeño, medio o grande basándose en una revisión de estudios en ISE. En este ejemplo, aunque los diseños experimentales no son exactamente los mismos, consideramos que los 5 estudios cumplen todas las condiciones para incorporarlos en el meta-análisis salvando las amenazas para que los resultados sean válidos, así pues trabajaremos con 5 estimaciones para construir el tamaño del efecto global.

Primeramente, realizamos un meta-análisis para estudiar la correlación entre cada par Medida-*CompSub*. En la Tabla 3 presentamos la estimación global de los coeficientes de correlación, con un intervalo de confianza del 95%, el *p-valor* y el valor de la *g de Hedges*, incluyendo una clasificación del tamaño del efecto como grande (G), mediano (M) o pequeño (P).

Los resultados obtenidos están a favor de la existencia de una correlación positiva entre la complejidad cognitiva y las 11 medidas que miden la complejidad estructural y el tamaño de los diagramas de clases UML. De hecho, muchos de los tamaños del efecto son medios o grandes, exceptuando el de NM que es pequeño. Las medidas de tamaño que tienen más influencia en la complejidad cognitiva son NC y NA, mientras que las medidas de complejidad que tienen más influencia en la complejidad cognitiva son las relacionadas con las asociaciones (NAssoc) y generalizaciones (NGen y MaxDIT). Así que los diagramas de clases UML con mayor número de clases y atributos tienen mayor complejidad cognitiva; además, cuanto mayor número de asociaciones, generalizaciones o profundidad de herencia tenga un diagrama de clases mayor será su complejidad cognitiva.

Respecto de las hipótesis derivadas del objetivo 2, la Tabla 4 muestra que podemos admitir la existencia de correlación entre la complejidad cognitiva y la entendibilidad y la modificabilidad de los diagramas de clases UML. Los tamaños del efecto son medios en ambos casos, pero la estimación de la correlación de la entendibilidad es mayor que la correlación de la modificabilidad. Podemos, por tanto, concluir que cuanta mayor complejidad cognitiva contiene un diagrama, más difícil será comprenderlo o modificarlo.

Tabla 3. Meta-análisis para las correlaciones entre Medidas-*CompSub*

H0: ρ<=0	Correlación (ρ) Tamaño del Efecto Global	Límite inferior	Límite superior	p-valor	g de Hedges
NC	0.566	0.464	0.653	0.0000	1.322(G)
NA	0.541	0.435	0.632	0.000	1.219(G)
NM	0.177	0.040	0.307	0.012	0.339(P)
NAssoc	0.566	0.465	0.653	0.000	1.318(G)
NAgg	0.481	0.368	0.581	0.000	1.051(M)
NDep	0.484	0.371	0.584	0.000	1.060(M)
NGen	0.484	0.371	0.584	0.000	1.018 (G)
NGenH	0.422	0.302	0.529	0.000	0.903 (M)
NAggH	0.393	0.270	0.504	0.000	0.814 (M)
MaxDIT	0.492	0.379	0.590	0.000	1.080 (G)
MaxHAgg	0.360	0.233	0.474	0.000	0.734 (M)

Tabla 4. Meta-análisis para las correlaciones *CompSub*-Tiempo de Entendibilidad y Modificabilidad

H0: ρ<=0	Correlación (ρ) Tamaño del Efecto Global	Límite inferior	Límite superior	p-valor	g de Hedges
Tiempo de Entendibilidad	0.330	0.200	0.449	0.000	0.684 (M)
Tiempo de Modificabilidad	0.186	0.044	0.320	0.011	0.368(M)

En la Figura 2 aparece el meta-análisis de la correlación entre un par de medidas y la medida *CompSub*, y entre su entendibilidad y su complejidad cognitiva, obtenido con la herramienta utilizada [4].

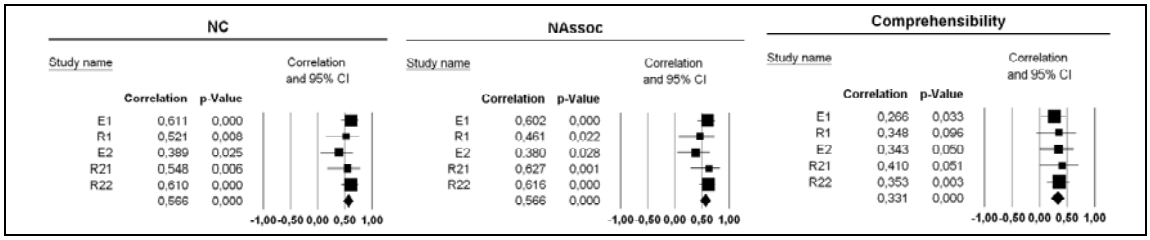

Fig. 2. Meta-análisis para NC-*CompSub*, NAssoc-*CompSub* y *CompSub*-Tiempo de entendibilidad

Los resultados del meta-análisis son relevantes porque podrán utilizarse como medios para controlar el nivel de ciertos atributos de calidad de los diagramas de clases UML durante la fase de modelado. Además, tienen implicaciones tanto prácticas como pedagógicas, proporcionando información sobre cuales son los constructores de UML que tienen más influencia sobre el esfuerzo para comprender y modificar los diagramas de clases UML. Por ello cuando existen diseños alternativos de un diagrama de clases UML, este conocimiento puede ser útil para seleccionar el que minimice estos constructores.

6 Conclusiones

Hemos presentado algunas de las técnicas que permiten sintetizar los resultados de varios experimentos de forma objetiva, facilitando así la acumulación y el refinamiento del conocimiento que se quiere conseguir cuando se propone una hipótesis de estudio. En áreas como la medicina y la psicología la experimentación y la síntesis de resultados se está utilizando desde hace décadas, luego en esas áreas ya se han enfrentado a problemas que los investigadores de ISE podemos aprovechar [4;24]. En [25] se muestra uno de los riesgos propio de ISE: la generalización de resultados es más que conflictiva, pues la elección de sujetos, en general, es por conveniencia, y los objetos proceden sólo de ciertos dominios.

En Ingeniería del Software el meta-análisis no se ha utilizado demasiado, y sobre todo el uso ha sido para comparar diferentes técnicas de revisión de software y para hacer revisiones sistemáticas de algunos temas [20; 22; 23; 14; 29], por lo que aún quedan ámbitos e hipótesis por explorar con esta técnica: efecto del uso de herramientas y nuevas tecnologías en atributos de calidad del software, o sobre la productividad en ciertas fases de desarrollo del software, entre otros. Nuestra intención es seguir utilizando esta técnica en experimentos realizados con diagramas de estado [6] y las expresiones OCL [27], pues estamos convencidos de que el meta-análisis es una herramienta útil si se realiza en determinadas condiciones.

Agradecimientos

Esta investigación forma parte del proyecto ESFINGE (TIN2006-15175-C05-05) financiado por el Ministerio de Educación y Ciencia (España), y del proyecto IDONEO (PAC08-0160-6141) financiado por la Consejería de Ciencia y Tecnología de la Junta de Comunidades de Castilla-La Mancha (España).

Referencias

1. Basili, V., Shull, F. y Lanubile, F.: Building Knowledge through Families of Experiments. IEEE Transactions on Software Engineering, 25, pp. 456-473. 1999.
2. Biostat, Comprehensive Meta-Analysis v2, 2006. http://www.meta-analysis.com/.
3. Briand, L., Bunse, C. y Daly, J.: A Controlled Experiment for Evaluating Quality Guidelines on the Maintainability of Object-Oriented Designs. IEEE Transactions on Software Engineering, 27(6), pp. 513-530. 2001.
4. Brooks, A. Meta Analysis- A silver Bullet- for Meta-Analysts. Empirical Software Engineering, 2 333-338. 1997.
5. Cohen J. Statistical power analysis for the behavioural sciences second ed. Lawrence Erlbaum Associates, Publishers London. 1988.
6. Cruz-Lemus, J. A.., Genero, M. y Piattini, M.: Metrics for UML Statechart Diagrams. In: Metrics for Software Conceptual Models, Imperial College Press, UK., 2005.
7. Dybå, T., Arisholm, E., Sjøberg, D. I. K., Hannay, J. E. y Shull, F.: Are Two Heads Better than One? On the Effectiveness of Pair Programming. IEEE Software, 24(6), pp. 10-13. 2007.
8. Field, A.P. Meta-Analysis of Correlation Coefficients. Phsychological Methods, 6 (2) 161-180. 2001.
9. Genero, M., Manso, M. E. y Piattini, M.: Early Indicators of UML Class Diagrams Understandability and Modifiability. ACM-IEEE International Symposium on Empirical Software Engineering, pp. 207-216. 2004.
10. Genero, M., Manso, M. E., Visaggio, A., Canfora, G. y Piattini, M.: Building Measure-Based Prediction Models for UML Class Diagram Maintainability. Empirical Software Engineering, 12, pp. 517-549. 2007.
11. Genero, M., Poels, G., Manso, M. E. y Piattini, M.: Defining and Validating Metrics for UML Class Diagrams. in Metrics for Software Conceptual Models, Imperial College Press, UK., 2005.
12. Glass, G. V., McGaw, B., and Smith, M. L. Meta-Analysis in Social Research. Sage Publications. 1981.
13. Glass, G.V. http://glass.ed.asu.edu/gene/papers/meta25.html. (16-7-2008)
14. Hayes, W.: Research in Software Engineering: a Case for Meta-Analysis. 6th IEEE International Symposium on Software Metrics (METRICS'99), pp. 143-151. 1999.
15. Hedges, L. V. y Olkin, I.: Statistical Methods for Meta-Analysis, Academia Press, 1985.
16. Höst, M., Regnell, B. y Wohlin, C.: Using Students as Subjects - a Comparative Study of Students & Proffesionals in Lead-Time Impact Assessment. 4th Conference on Empirical Assessment & Evaluation in Software Engineering (EASE 2000), pp. 201-214. 2000.
17. Juristo, N., y Moreno, S.: Basics of Software Engineering Experimentation. Kluwer Academic Publishers, 2001.
18. Kampenes, V., Dybå, T., Hannay, J. E. y Sjoberg, D. I. K.: A Systematic Review of Effect Size in Software Engineering Experiments. Information and Software Technology, 49(11-12), pp. 1073-1086. 2007.
19. Kitchenham, Barbara. Procedures for Performing Systematic Reviews, Joint Technical Rreport, Keele University TR/SE-0401 and NICTA 0400011T.1, July 2004.
20. Laitenberger, O., El-Emam, K. y Harbich, T.: An Internally Replicated Quasy-Experimental Comparison of Checklist and Perspective-based Reading of Code Documents. IESE, 006.99/e, 1999.
21. Lipsey, M. y Wison, D.: Practical Meta-Analysis, Sage, 2001.
22. Miller, J. y McDonald, F.: Statistical Analysis of Two Experimental Studies. University of Strathclyde, EFoCS-31-98, 1998.
23. Miller, J. Applying Meta-Analytical Procedures to Software Engineering Experiments. Journal of Systems and Software, 54: 29-39. 2000.
24. Pai, Madhukar, McCullovch, Michael, Gorman, Jennifer D., Pai, Nitika, Enanoria, Wayne, Kennedy, Gail, Tharyan, Prathap, Colford, John M. Jnr. Systematic reviews and meta-analysis: An illustrated, step-by-step guide. The National medical Journal of India, 17(2) 2004, pp 86-95.
25. Pickard, M.: Combining Empirical Results in Software Engineering. University of Keele, T-R V1, 2004.
26. Porter, A. y Johnson, M.: Assessing Software Review Measurement: Necessary and Sufficient Properties for Software Measures. Information and Software Technology, 42(1), pp. 35-46. 1997.
27. Reynoso, L., Genero, M. y Piattini, M.: Measuring OCL Expressions: An approach based on Cognitive Techniques. in Metrics for Software Conceptual Models, Imperial College Press, UK., 2005.

28. Rosenthal, R. Meta-Analytic Procedures for Social Research. Sage Publications. 1986..
29. Shull, F., Carver, J., Maldonado, J., Conradi, R., Basili, V. Replicated Studies: Building a Body of Knowledge about Software Tecniques. In N. Juristo y A. Moreno (Eds). Lecture Notes on Empirical Software Engineering (pp 39-84). Singapore. World Scientific. 2003.
30. Succi, G. , Spasojevic, R., Hayes, J.J., Smith, M.R., Pedrycz, W. Application of Statistical Meta-Analysis to Software Engineering Metrics Data. Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, vol 1 709-714. 2000.
31. Sjoberg, D. I. K. , Hannay, J. E., Hansen, O., Kampenes, V., Karahasanovic, A., Liborg, N. K. y Rekdal, A. C.: A Survey of Controlled Experiments in Software Engineering. IEEE Transactions on Software Engineering, 31(9), pp. 733-753. 2005.
32. SPSS, SPSS 12.0, Syntax Reference Guide, 2003.
33. Sutton, J. A., Abrams, R. K., Jones, R. D., Sheldon, A. T., and Song, F. Methods for Meta-Analysis in Medical Research. John-Wiley & Sons. 2001.
34. Wohlin, C., Runeson, P., Hast, M., Ohlsson, M. C., Regnell, B. y Wesslen, A.: Experimentation in Software Engineering: an Introduction. Kluwer Academic Publisher, 2000.
35. Wolf, F. M. Meta-Analysis: Quantitative Methods for Research Synthesis. Sage Publications. 1986.

Apéndice A

Después de estudiar el metamodelo UML y revisar la literatura existente sobre medidas, propusimos un conjunto de medidas para la complejidad estructural de los diagramas de clases UML [11] (ver Tabla 5). Las medidas propuestas están relacionadas con el uso de las relaciones de UML, como asociaciones, dependencias, agregaciones y generalizaciones.

Tabla 5. Medidas para diagramas de clases UML

	Nombre	Descripción
Medidas de tamaño	Número de clases (NC)	El número total de clases en un diagrama de clases.
	Número de atributos (NA)	El número de atributos definidos en todas las clases en un diagrama de clases (no incluye el número de atributos heredados o atributos definidos en métodos). Esto incluye atributos definidos en una instancia.
	Número de métodos (NM)	El número total de métodos definidos en todas las clases de un diagrama de clases, no incluyendo métodos inherentes (porque podrían ser contados doblemente). Esto incluye métodos definidos en una clase instanciada.
Medidas de complejidad estructural	Número de asociaciones (NAssoc)	El número total de relaciones de asociación en un diagrama de clases.
	Número de agregaciones (NAgg)	El número total de relaciones de agregación (cada pareja de elementos en una relación de agregación).
	Número de dependencias (NDep)	El número total de relaciones de dependencia.
	Número de generalizaciones (NGen)	El número total de relaciones de generalización (cada pareja padre-hijo en una relación de generalización).
	Número de jerarquías de generalización (NGenH)	El número total de jerarquías de generalización, esto es, contar el número total de estructuras con relaciones de generalización.
	Número de jerarquías de agregación (NAggH)	El número total de jerarquías de agregación, esto es, contar el número total de estructuras todo-partes en un diagrama de clases.
	Máximo DIT (MaxDIT)	El máximo valor DIT obtenido para cada clase en un diagrama de clases. El valor DIT para una clase en una jerarquía de generalización es el mayor camino desde la raíz de la jerarquía (Chidamber y Kemerer, 1994).
	Máximo HAgg (MaxHAgg)	El máximo valor HAgg obtenido para cada clase en un diagrama de clases. El valor HAgg de una clase en una jerarquía de agregación es el mayor camino de la clase a las hojas.

Pai, Madhukar, McCullovch, Michael, Gorman, Jennifer D., Pai, Nitika, Enanoria, Wayne, Kennedy, Gail, Tharyan, Prathap, Colford, John M. Jnr. Systematic reviews and meta-analysis: An illustrated, step-by-step guide. The National medical Journal of India, 17(2) 2004, pp 86-95.