Pautas para Agregar Estudios Experimentales en Ingeniería del Software

Enrique Fernández, Oscar Dieste, Patricia Pesado, Ramón García-Martínez

Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. UNLP. Argentina
Grupo de Ingeniería de Software Experimental. Facultad de Informática. UPM. España
Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP - CIC
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA. Argentina
Área Ingeniería del Software. Licenciatura en Sistemas. Departamento de Desarrollo Productivo y Tecnológico. UNLa. Argentina

Resumen. El presente trabajo muestra cómo la aplicación de métodos de Meta-Análisis alternativos al método Diferencias Medias Ponderadas, como son el Response Ratio o el Conteo de Votos, puede ayudar a mejorar las piezas de conocimientos generadas en las Revisiones Sistemáticas hechas dentro del campo de la Ingeniería del Software.

1. Introducción

La agregación de estudios experimentales en Ingeniería del Software (IS) fue propuesta originalmente por [1] en 1996 y a partir de la presentación del trabajo de [2] en 2004 el interés por su desarrollo por parte de la comunidad científica se incrementa considerablemente [2; 4; 5; 6; 7]. Según [8] la agregación de experimentos consiste en combinar los resultados de varios estudios experimentales, previamente desarrollados, que analizan el comportamiento de un par de tratamientos específicos, en un contexto determinado con el objeto de generar nuevas piezas de conocimiento. Cuando este tipo de trabajo es realizado mediante métodos estadísticos, se los suele llamar Meta-Análisis (MA), término empleado por primera vez por [9]. Si bien el MA aporta un conjunto de herramientas para poder combinar de manera fiable los resultados de los estudios, para que las conclusiones aportadas sean consideradas representativas no debe existir sesgos en la selección de los estudios que se van a agregar. Esto se logra mediante el uso de las Revisiones Sistemáticas (RS). Una RS es un procedimiento que aplica estrategias científicas para aumentar la fiabilidad del proceso de recopilación, valoración crítica y agregación de los estudios experimentales relevantes sobre un tema [10]. Si bien, en la actualidad se conocen muchas RS en el ámbito de la IS que lograron recopilar experimentos y hacer una valoración crítica de los mismos. Estas, en general, fallan a la hora de agregar los resultados, principalmente, por no poder aplicar el método Diferencias Medias Ponderadas [11] (DMP) propuesto en [2]. Así por ejemplo, en [3] no se pudo desarrollar el MA debido a la escasez de replicaciones y la baja calidad de los informes publicados, o en [4] no se puedo desarrollar el MA debido a la gran diversidad de técnicas y variables respuesta utilizadas. Este hecho motivó que

Actas JISBD 2009, pp.91-102, ISBN: 978-84-692-4211-7

algunos autores viendo la dificultad de desarrollar un MA, debieron recurrir a métodos alternativos no estadísticos. Por ejemplo en [3] se realizó una sumarización de los estudios a favor de cada tratamiento, y proclamó como ganador al tratamiento con mayor cantidad de estudios a favor, lo cual constituye una instancia de un procedimiento bien conocido en Ciencias Sociales y denominado Conteo de Votos no Estadístico [12] (CVNE). Si bien de esta forma es posible derivar nuevas piezas de conocimiento de tipo general, la fiabilidad de las mismas está en entredicho ya que el conteo de votos tiende a producir falsos resultados negativos cuando el poder estadístico de los test es reducido [11]. Ello puede provocar que muchas piezas de conocimiento sean erróneas. Con el objeto de mitigar este problema se han desarrollado nuevos métodos de MA, que sean menos restrictivos en cuanto a sus condiciones de aplicación, pero que mantengan la esencia de fiabilidad que los métodos estadísticos deben aportar.
El objetivo de este trabajo es presentar dos métodos de MA alternativos a DMP, como son el Conteo de Votos Estadístico (CVE) y el Response Ratio (RR), y mostrar como la fiabilidad de las piezas de conocimiento aumenta frente a lo que obtenemos con el CVNE. En la sección 2 de este artículo se describen los métodos de agregación; en la sección 3 se describen un conjunto de pautas respecto de cómo y cuándo utilizar los distintos métodos; en la sección 4 se presenta un caso de estudio; en la sección 5 se presenta un conjunto de puntos de discusión; finalmente, en la sección 6 se describe algunas de las conclusiones obtenidas.

2. Estado del Arte

2.1 Agregación de Estudios Experimentales

Si todos los estudios a incluir en un proceso de agregación fuesen igualmente precisos, bastaría con promediar los resultados de cada uno de ellos para obtener así una conclusión final. Sin embargo, en la práctica no todos los estudios tienen la misma precisión, por ello cuando se los combine se debe asignar un mayor peso a los estudios que permiten obtener información más fiable. Esto se logra combinando los resultados mediante un promedio ponderado [13]. Además como los resultados de los diferentes estudios pueden medirse en diferentes escalas de la variable respuesta [11], la variable dependiente en un MA debe poder compatibilizar estos aspectos, lo cual se logra mediante la estimación de un "tamaño de efecto", el cual consiste en un estimador estandarizado no escalar de la relación entre una exposición y un efecto. En sentido general, este término se aplica a cualquier medida de la diferencia en el resultado entre los grupos de estudio (por ejemplo, cantidad de líneas de código que posee un programa o cantidad de requisitos educidos). Cuando lo que se desea combinar son estudios que trabajan con variables continuas (como sucede en IS), el método a aplicar por excelencia es Diferencias Medias Ponderadas [1; 2]. Este método es conceptualmente sencillo [9]: el tamaño de efecto de cada estudio se estima como la diferencia de la medias dividida por la varianza conjunta de ambos tratamientos:

$$g = \frac{Y^E - Y^C}{S_P}$$

g es el tamaño de efecto
Y‘s son las medias de los tratamientos
S_p es la varianza conjunta **(1)**

La función (1) fue optimizada por [11], quien incorporó un factor de corrección “J” que se utiliza para aumentar la fiabilidad del método cuando los estudios a agregar cuentan con pocos sujetos experimentales. La nueva función se la conoce como “*d*” y es la recomendada en [2; 8]. Una vez estimado el tamaño de efecto de cada estudio se debe estimar el tamaño de efecto general o global, para ello se utiliza la siguiente función:

$$d^* = \frac{\sum d_i / \sigma^2{}_i(d)}{\sum 1/\sigma^2{}_i(d)}$$

d^* es el tamaño de efecto global
$\sum d_i / \sigma^2{}_i(d)$ es la suma de los efectos individuales
$\sum 1/\sigma^2{}_i(d)$ es la suma de la inversa varianza **(2)**

Para mayores detalles de cómo aplicar las formulas indicadas remitirse a [11]. Es importante destacar que para que este método pueda aplicarse el informes del experimento debe indicar: número de sujetos, medias y varianzas, lo cual en general no ocurre en los artículos presentados en IS [14], por ello muchos autores han recurrido a alternativas de agregación no estadísticas. Es de hacer notar que si bien existen algunas versiones de DMP alternativas a las indicadas anteriormente en las cuales el tamaño de efecto es estimado a partir de los parámetros F o t [11], estas funciones no se encuentran muy difundidas y se desconoce si las mismas arrojan resultados fiables o no.

2.2 Métodos Alternativos de Agregación

2.2.1 Conteo de Votos estadístico

Este método fue propuesto como una alternativa a DMP por [11] para agregar estudios con carencias en los informes o cuando se requiere combinar estudios que utilizan variables respuesta diferentes, pero, compatibles entre sí (por ejemplo estudios que miden el tamaño del código de un programa a través de la cantidad de clases podrían juntarse con otros que miden el tamaño en función de la cantidad de líneas de código). Su principal ventaja radica en que para estimar el tamaño de efecto solo se requiere conocer si existe o no diferencia entre las medias de los tratamientos y la cantidad de sujetos experimentales. Con esta información se realiza un proceso de inferencia que intenta determinar cuál es el tamaño de efecto de mayor probabilidad de ocurrencia en base a los parámetros dados. Su función principal es la siguiente:

$$L(\delta \mid X_1, \ldots, X_i) = \sum_{i=1}^{k} \left\{ \begin{matrix} X_i \ln\left[1 - \phi\left(-\sqrt{\tilde{n}}\delta\right)\right] + \\ (1 - X_i) \ln \phi\left(-\sqrt{\tilde{n}}\delta\right) \end{matrix} \right\}$$

$L(\delta|X_1, \ldots X_n)$ es la probabilidad de tamaño de efecto
δ es el tamaño de efecto a testear
X_i es el valor del voto de cada estudio
$\tilde{n} = (n^E + n^C) / (n^E * n^C)$ donde n’s son las cantidades de sujetos experimentales de cada estudio **(3)**

Para mayores detalles de cómo aplicar las formulas indicadas remitirse a [11].

2.2.2 Response Ratio

Es el método de agregación recomendado dentro del ámbito de la ecología [15], El mismo consiste en estimar un índice de efecto, o Ratio, entre un tratamiento Experimental y otro de Control mediante el cociente de ambas medias ($RR = Y^E / Y^C$). Este cociente permite estimar la proporción de mejora existente entre ambos tratamientos. Así, por ejemplo, un ratio de 1.3 indicará que el tratamiento experimental es un 30% mejor que el de control, o un ratio de 1 indicará que no hay diferencias en el desempeño de ambos tratamientos.

Para que la combinación de un conjunto de estudios sea más precisa se le incorporó al método el logaritmo natural, que permite linealizar los resultados y normalizar su distribución, convirtiéndolo en un método apropiado para estimaciones de efectos cuando el conjunto de experimentos es pequeño [16]. Es importante destacar que una vez estimados el índice, se debe aplicar el anti-logaritmo al resultado para obtener el índice de efecto final. La aplicación del método consta de dos pasos, primeramente se debe estimar el Ratio de cada uno de los experimentos y luego, en base a éstos, se estima el Ratio o efecto global mediante un promedio ponderado de los ratios individuales, como se muestra en la función 4:

$$L^* = \frac{\sum_{i=1}^{k} W_i^* L_i}{\sum_{i=1}^{k} W_i^*}$$

L^* es el efecto global
L_i es el efecto de cada estudio
W_i es el factor de peso = $1/v$ **(4)**

Es importante destacar que el factor de peso de los estudios puede estimarse aplicando un método del tipo *paramétrico* (RRP) donde el ponderador de los estudios es la inversa de la varianza (como lo indica la función 5) o *no paramétrico* (RRNP), donde el ponderador de los estudios es la cantidad de sujetos experimentales (como lo indica la función 6). Siendo la principal ventaja del método no paramétrico el hecho de que no requiere conocer las varianzas de los estudios y la principal ventaja del método paramétrico su alta precisión aún cuando los estudios incluyen pocos sujetos.

$$v = \frac{S^{2E}}{n^E Y^E} + \frac{S^{2C}}{n^C Y^C}$$

v es el error típico
S^2‘s son las varianzas de los estudios
Y‘s son las medias de los estudios
n‘s son las cantidades de sujetos **(5)**

$$v = \frac{n_C + n_E}{n_E n_C} + \frac{Ln(RR^2)}{2(n_C + n_E)}$$

v es el error típico
n‘s son las cantidades de sujetos
RR es el Ratio **(6)**

Para mayores detalles de cómo aplicar las formulas indicadas remitirse a [15].

3. Aplicación de los Métodos de Agregación

3.1. El problema de la Agregación en la IS

Como se mencionó anteriormente, en la actualidad muchos investigadores han intentado aplicar MA en IS, pero a excepción de [7] que logró agregar 15

experimentos vinculados al rendimiento de los programadores cuando se trabaja de a pares, la mayoría de los trabajos no pudieron aplicarlo. Por ejemplo: en [4] se identificó un conjunto de experimentos vinculados a técnicas de prueba de software, pero no se pudo desarrollar el MA debido a la gran diversidad de técnica encontradas, a la falta de estandarización de variables respuesta y la falta de publicación de las varianzas en los informes; en [17] se identificó un conjunto de estudios experimentales que analizan el comportamiento de los métodos de estimación de software, pero no se pudo desarrollar el MA debido a la gran discrepancia entre los métodos identificados y las variables respuesta utilizadas; en [18] se analizó cuáles son los factores que motivan la adopción de un proceso de mejora basado en CMMI, pero no se pudo desarrollar el MA debido a la falta de estandarización de las variables respuesta y la baja calidad de los informes; en [19] se analizó el desempeño de los métodos de estimación generados a través de la experiencia recabada en una única compañía respecto de los métodos desarrollado en base a la experiencia recabada en varias compañías, pero no se pudo desarrollar el MA por un problema de compatibilidad entre los estudios identificados y la baja calidad de los informes; en [20] se analizó la reutilización del software en la modificación y/o creación de nuevos productos, pero no se pudo desarrollar el MA debido a la gran diversidad en las variables respuesta analizadas en cada estudio.
En resumen, podemos decir que los principales problemas que hoy afronta la comunidad científica en IS cuando decide realizar un MA son: la escasez de experimentos y/o replicaciones, discrepancias en las variables respuesta y la falta de estándares para informes de experimentos.

3.2. Desarrollo de la solución

Si bien existen tres problemas básicos a la hora de desarrollar un MA en IS, el problema de escasez de experimentos no puede ser solucionado por los métodos de agregación, pero contar con métodos de agregación más flexibles puede ayudar a reducir la cantidad de estudios descartados y aumentar así su número. En tal sentido, podemos decir que el RRNP no requiere conocer las varianzas para poder ser aplicado, lo que permita paliar en parte el problema de la baja calidad de los informes experimentales. El método CVE permite combinar los resultados de los experimentos aún cuando las variables respuestas utilizadas en cada experimento no sean las mismas, además no requiere conocer las varianzas ni las medias de los tratamientos. Ahora bien no todos los métodos de agregación presentan resultados con el mismo nivel de fiabilidad y precisión. Según el trabajo de [21] el RRP puede considerarse un método alternativo al DMP para su uso en medicina. En general, el método DMP tiene mayor precisión cuando se agregan muchos experimentos mientras que el RRP tiene mayor precisión cuando se agregan pocos experimentos [16]. El RRNP por su parte tiene un aceptable nivel de error independientemente de la cantidad de estudios que se agreguen pero su potencia es bastante limitada [16], ya que por su esencia no paramétrica requiere que las diferencias entre los tratamientos sean amplias para indicar que la misma es significativa, y el VCE es un interesante método de agregación cuando la cantidad de estudios a favor de cada uno de los tratamientos es similar [11], aumentando su precisión con el incremento de la cantidad de

experimentos [16]. A continuación se presentan un conjunto de pautas para poder aplicar los métodos de agregación presentados de forma conjunta.

3.3. Pautas para el uso de los métodos de agregación

Para determinar cuándo usar un método de MA u otro, se deben analizar sus restricciones de uso (parámetros estadísticos, relación entre las variables respuesta, etc.) y la cantidad de estudios identificados como base para determinar qué método utilizar en caso que pueda aplicarse más de uno. En tal sentido, podemos decir que si se cuenta con un conjunto de 15 o mas experimentos que publican medias, varianza, y cantidad de sujetos experimentales se recomienda utilizar DMP, por ser éste el método mas eficiente [16], pero, en caso de tener menos de 15 experimentos con estas características, se deberían agregar mediante el método RRP [16].
Por otra parte cuando los estudios presentan problemas de informes, se recomienda utilizar el RRNP si los estudios publican solo las medias y la cantidad de sujetos experimentales, y VCE si en los estudios en lugar de detallar las medias se limitan a indicar que un tratamiento es mejor que el otro. En la tabla 1 se describen los criterios antes mencionados.

<table>
<tr><th>El informes publica
Cantidad de estudios</th><th>Medias, varianzas y cantidad de sujetos experimentales</th><th>Medias y cantidad de sujetos experimentales</th><th>Diferencias entre Medias y cantidad de sujetos experimentales</th></tr>
<tr><td>>= 15</td><td>Usar DMP</td><td rowspan="2">Usar RRNP</td><td rowspan="2">Usar VCE</td></tr>
<tr><td>< 15</td><td>Usar RRP</td></tr>
</table>

Tabla 1. Condiciones de aplicación de los métodos.

Se debe tener en cuenta que las condiciones descriptas en la tabla 1, solo se aplican si las variables respuesta publicadas por los experimentos unicamente difieren en la escala de valores utilizada, para el caso de variables respuesta no iguales el único método de agregación aplicable es VCE.
Para aumentar el nivel de evidencia de las conclusiones, se recomienda no utilizar los métodos de forma aislada, sino, por el contrario utilizarlos de forma conjunta, siempre y cuando el uso de alguno de ellos permita incrementar la cantidad de estudios factibles de ser agregados. Así, por ejemplo, si se cuenta con 5 experimentos agregables por RRP y tres por RRNP, la recomendación es agregar los primeros 5 con RRP y los 8 totales por RRNP, de esta forma se podrá generar distintos grupos de agregación que permitan ver como evoluciona el fenómeno con el aumento de la evidencia.
Es de hacer notar que cuando se utiliza más de un método de agregación en forma simultánea, los resultados obtenidos pueden no ser compatibles. Cuando esto sucede, se recomienda hacer un análisis para tratar de identificar variables no controladas que afecten al resultado obtenido por cada método, de forma similar a como se hace con el análisis de heterogeneidad.

4. Caso de Estudio

4.1 Presentación del caso de estudio

Para mostrar cómo aplicar los métodos de agregación bajo las pautas definidas, se tomó como base los estudios identificados en la RS desarrollado por [3], ya que los estudios están disponibles, y se contrastarán las conclusiones publicadas, en base a dicha RS, en el artículo [22]. Motiva esta elección el hecho de que en esta RS se ha identificado una gran variedad de técnicas de educción de requisitos, tanto tradicionales (como son las Entrevistas o Análisis de Protocolo) como alternativas (como son Card Sort y Ladderin grid entre otras), lo que permite realizar varios MA.

4.2 Comparación de Resultados

Para que el contraste de resultados con los métodos de agregación sea más claro, los resultados obtenidos se muestran mediante los conocidos diagramas de árboles. Es de hacer notar que dichos diagramas en lugar de estar dibujados de forma vertical, como habitualmente lo dibujan los programas especializados en MA, han sido dibujados de forma horizontal mediante Ms EXCEL debido a que no se ha podido identificar ningún programa que aplique los métodos de agregación alternativos. Tener en cuenta que los extremos de las líneas vinculadas al eje Y representan los límites del intervalo de confianza, el punto identifica el tamaño de efecto y el peso del estudio se representa mediante el valor que acompaña el código de experimento debajo del eje X.

4.2.1 Caso 1: El Análisis de Protocolo no aporta más conocimientos que la Entrevista

Esta conclusión fue obtenida en base a cuatro experimentos (E-1 [23], E-2 [24], E-3 [25] y E-4 [25]) que comparan Entrevista vs. Análisis de Protocolo (AP) analizando la variable ganancia (cantidad de cláusulas obtenidas). Dado que solo uno de estos estudios publica todos los parámetros estadísticos y los tres restantes publican medias y cantidad de sujetos experimentales, y no existen problemas con las variables respuesta, los mismos pueden agregarse mediante RRNP.

Como se aprecia en la figura 1, el tamaño de efecto global es apróximadamente 1,5, lo que implicaría que la entrevista es casi un 50 % mejor que el AP, pero, como el intervalo de confianza contiene al valor 1, la evidencia no puede ser considerada significativa al 95%. Por lo tanto, *se reafirma la no existencia de evidencia de que el AP es mejor que la entrevista.*

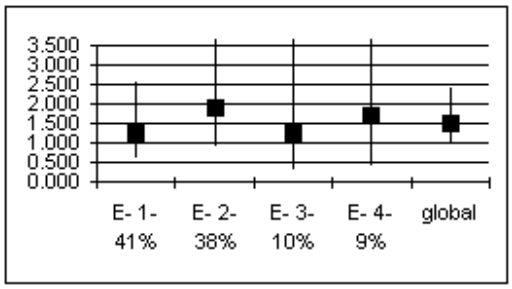

Fig. 1. Entrevista vs. AP (Ganancia).

4.2.2 Caso 2: Las Entrevistas insumen más tiempos que el Análisis de Protocolo
Esta conclusión fue obtenida en base a cuatro experimentos (E-1 [23], E-2 [24], E-3 [25] y E-4 [25]) que comparan Entrevista vs. Análisis de Protocolo (AP) analizando la variable esfuerzo (tiempo necesario para desarrollar la sesión y analizar los datos). Dado que solo uno de estos estudios publica todos los parámetros estadísticos y los tres restantes publican medias y cantidad de sujetos experimentales, y no existen problemas con las variables respuesta, los mismos pueden agregarse mediante RRNP.
Como se aprecia en la figura 2, el tamaño de efecto global es apróximadamente 1, lo que implicaría que no se espera que existan diferencias en el desempeño de ambas técnicas. Por lo tanto, *se contradice la afirmación original, no hay evidencias para afirmar que la entrevista consume más tiempo que el AP.*

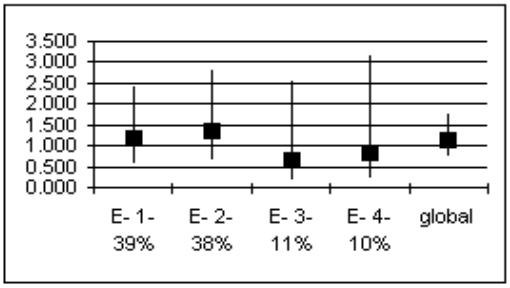

Fig. 2. Entrevista vs. AP (Esfuerzo).

4.2.3 Caso 3: Laddering Grid aporta más conocimiento que Card Sort
Esta conclusión fue obtenida en base a cuatro experimentos (E-1 [23], E-2 [24], E-3 [25] y E-4 [25]) que comparan las técnicas Laddering Grid (LD) vs. Card Sort (CS) analizando la variable ganancia (cantidad de cláusulas obtenidas). Dado que solo uno de estos estudios publica todos los parámetros estadísticos y los tres restantes publican medias y cantidad de sujetos experimentales, y no existen problemas con las variables respuesta, los mismos pueden agregarse mediante RRNP.
Como se aprecia en la figura 3, el tamaño de efecto global es apróximadamente 2, lo que implicaría que LG duplica el rendimiento de CS, mostrando además evidencias

significativas al 95%. Por lo tanto, *se reafirma el mejor desempeño de LG y se destaca que las diferencias son significativas*

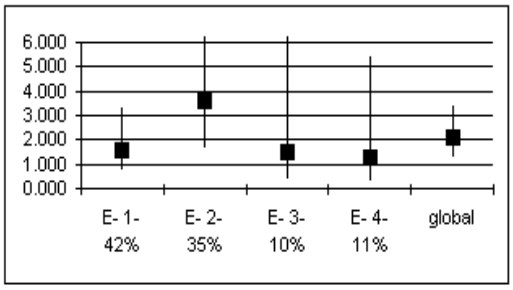

Fig. 3. LD vs. CS (Ganancia).

4.2.4 Caso 4: Laddering Grid consume más tiempo que Card Sort

Esta conclusión fue obtenida en base a cuatro experimentos (E-1 [23], E-2 [24], E-3 [25] y E-4 [25]) que comparan las técnicas Laddering Grid (LD) vs. Card Sort (CS) analizando la variable esfuerzo (tiempo necesario para desarrollar la sesión y analizar los datos). Dado que solo uno de estos estudios publica todos los parámetros estadísticos y los tres restantes publican medias y cantidad de sujetos experimentales, y no existen problemas con las variables respuesta, los mismos pueden agregarse mediante RRNP.

Como se aprecia en la figura 4, el tamaño de efecto global es apróximadamente 1,3, lo que implicaría que LG mejora un 30 % el desempeño de CS, pero la evidencia no es significativa al 95%, dado que el intervalo de confianza contiene claramente al valor 1. Por lo tanto, *se contradice la afirmación original, no hay evidencias para afirmar que la LG consume más tiempo que el CS.*

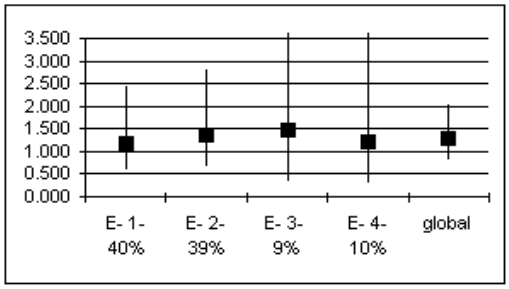

Fig. 4. LD vs. CS (esfuerzo).

4.2.5 Caso 5: La entrevista es más eficiente que CS

Esta conclusión fue obtenida en base a cinco experimentos (E-1 [23], E-2 [24], E-3 [25], E-4 [25] y E-5 [26]) que comparan las Entrevistas vs. Card Sort (CS) analizando la variable ganancia estimada como: la cantidad de cláusulas, en los primeros 4 experimentos, y la cantidad de reglas en el último. Dado que solo uno de estos estudios publica todos los parámetros estadísticos y los cuatro restantes publican medias y cantidad de sujetos experimentales, se realizarán dos agregaciones, la

primera incluirá cuatro experimentos mediante RRNP y la segunda incluirá a los cinco experimentos mediante VCE.
Como se aprecia en las figuras 5 y 6, las Entrevistas muestran un mejor comportamiento que CS. En la primer agregación el tamaño de efecto global es apróximadamente 1,35, lo que implicaría que la mejora alcanza el 35%, pero no es significativa al 95% da que el intervalo de confianza contiene al valor 1, en cambio en la segunda agregación, que incorpora un nuevo experimento, muestra resultados significativos y un tamaño de efecto medio similar al estimado anteriormente. Por lo tanto, *se reafirma el mejor desempeño de la entrevista y se destaca que las diferencias son significativas.*

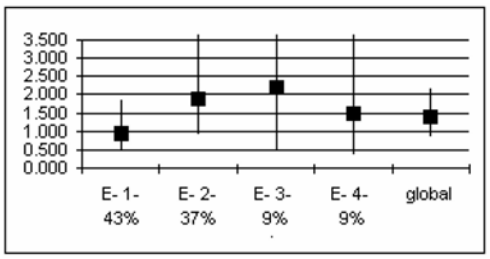

Fig. 5. Entrevista vs. CS (Ganancia) - RRNP

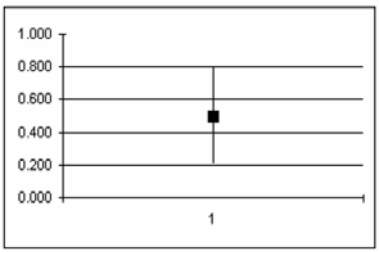

Fig. 6. Entrevista vs. CS (Ganancia) - VCE.

5. Discusión

Como se puede apreciar en los resultados obtenidos, en general, cuando todos los estudios apoyan a un mismo tratamiento, el resultado obtenido en el MA estadístico es compatible con el resultado original mediante CVNE, pero esto no puede tomarse como una regla universal, ya que en el caso de la combinación de los estudios vinculados a la Ganancia de LD vs. CS, a pesar de que todos los estudios apoyaban a la técnica LG, el resultado final no fue significativo. Esto refuerza la importancia de utilizar métodos estadísticos cuando se realiza un MA para evitar caer en errores de interpretación de los resultados.
También es importante el uso de MA cuando no existe un claro ganador entre los tratamientos, ya que en muchos casos pueden existir diferencias significativas a pesar de que no se advierta un claro ganador mediante el CVNE.
Por otra parte, el hecho de haber desarrollado dos agregaciones en la comparación de las Entrevistas vs Card Sort permitió, por un lado, afirmar que las diferencias son significativas, gracias al resultado del VCE, y por otro lado, ver como es la influencia de cada estudio en la conclusión final, gracias al RRNP.

6. Conclusiones

Se considera que la incorporación de los métodos de agregación RRNP y VCE ha sido altamente positiva. Si bien la mayoría de las nuevas conclusiones son compatibles con las tomadas originalmente en el artículo [22], el hecho de poder estimar un tamaño de efecto junto con su intervalo de confianza con una fiabilidad del 95% permite a los investigadores expresar sus conclusiones con mayor seguridad. Además mostrar los resultados a través de gráficas como el diagrama de árboles permite hacer un chequeo de heterogeneidad, un aspecto más que relevante a la hora de generar un resultado mediante la agregación de experimentos.
Dado que los métodos de agregación utilizadas han sido desarrolladas por investigadores de otras ramas de la ciencia, como son la medicina y la ecología, se considera de utilidad desarrollar nuevos trabajos tendientes a mostrar como es el comportamiento de estos métodos en un contexto experimental como el que hoy presenta la IS.

7. Financiamiento

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología del Gobierno de España, en el marco del proyecto TIN2008-00555.

8. Referencias

1. Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sörumgård, S., Zelkowitz, M.; 1996; *The empirical investigation of perspective-based reading*, International Journal on Empirical Software Engineering, Vol. 1, No. 2; pp. 133–164
2. Kitchenham, B. A.; 2004; *Procedures for performing systematic reviews*. Keele University; TR/SE-0401. Keele University Technical Report.
3. Davis, A.; Dieste o.; Hickey, A.; Juristo, N.; Moreno, A.; 2006; *Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review*; 14th IEEE Int. Requirements Engineering Conference (RE'06); pp. 179-188
4. Juristo N., Moreno A., Vegas S.; 2004; *Towards building a solid empirical body of knowledge in testing techniques*. ACM SIGSOFT Software Engineering Notes (SIGSOFT) 29(5); pp. 1-4
5. Jedlitschka, A. and Ciolkowski, M.; 2006; *Reporting Experiments in Software Engineering*; Franhofer Institute for Experimental Software Engineering.
6. Dyba, T., Kampenes, V., & Sjoberg, D.; 2006; *A systematic review of statistical power in software engineering experiments*. Information and Software Technology, 48(8); pp. 745-755.
7. Dyba, T., Aricholm, E.; Sjoberg, D.; Hannay J.; Shull, F.; 2007; *Are two heads better than one? On the effectiveness of pair programming*. IEEE Software; pp. 12-15.
8. Miller, J.; 2000; *Applying Meta-analytical Procedures to Software Engineering Experiments*. Journal of Systems and Software. (54): 1; pp. 29-39.
9. Glass, G; 1976; *Primary, secondary, and meta-analysis of research.* Educational Researcher 5; pp. 3-8

10. Goodman C.; 1996*; Literature Searching and Evidence Interpretation for Assessing Health Care Practices;* SBU; Stockholm.
11. Hedges, L.; Olkin, I.; 1985; *Statistical methods for meta-analysis*. Academic Press.
12. Mohagheghi, P., & Conradi , R.; 2004; *Vote-Counting for Combining Quantitative Evidence from Empirical Studies - An Example*. Proceedings of the International Symposium on Empirical Software Engineering (ISESE'04) .
13. Borenstein, M.; Hedges, L; Rothstein, H.; 2007; *Meta-Analysis Fixed Effect vs. random effect*; www.Meta-Analysis.com.
14. Sjoberg, D.; 2005; *A survey of controlled Experiments in Software Engineering*; IEEE Transactions on Software Engineering; Vol 31 Nro. 9
15. Gurevitch, J. and Hedges, L.; 2001; *Meta-analysis: Combining results of independent experiments.* Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch); pp. 347–369. Oxford University Press, Oxford.
16. Lajeunesse, M & Forbes, M.; 2003; *Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques*. Ecology Letters, 6; pp. 448-454.
17. Jřrgensen, M.; 2004; A *Review of Studies on Expert Estimation of Software Development Effort*. Journal of Systems and Software. (70): 1-2; pp. 37-60.
18. Staples, M; Niazi,M; Ross Jeffery, D; Abrahams, A; Byatt, P; Murphy, R; 2007: *An exploratory study of why organizations do not adopt CMMI*. Journal of Systems and Software 80(6); pp. 883-895
19. Kitchenham, B.; 2007; *Cross versus Within-Company Cost Estimation Studies: A Systematic Review*; IEEE Transactions on Software Engineering; Vol 33 Nro. 5.
20. Mohagheghi, P., & Conradi , R.; 2004; *Vote-Counting for Combining Quantitative Evidence from Empirical Studies - An Example*. Proceedings of the International Symposium on Empirical Software Engineering (ISESE'04) .
21. Friedrich, J, Adhikari, N; Beyene, J; 2008; *The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study*; BMC Medical Research Methodology
22. Dieste, O.; Juristo, N.; Shull, F.; *Understanding the Customer: What Do We Know about Requirements Elicitation? Software*, IEEE; Volume 25, Issue 2, March-April 2008; pp. 11 – 13.
23. Burton, A., Shadbolt, N., Hedgecock, A. y Rugg, G.; 1988; *A Formal Evaluation of Knowledge Elicitation Techniques for Expert Systems: Domain 1*. Proceedings of Expert Systems '87 on Research and Development in Expert Systems IV**.** Pág. 136-145.
24. Corbridge, C., Rugg, G., Major, P., Shadbolt, N. y Burton, A. 1994. *Laddering: Technical and Tool in Knowledge Acquisition.* Department of Psychology, University of Nottingham.
25. Burton, A., Shadbolt, N., Rugg, G. y Hedgecock, A.; 1990. *The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise*. Knowledge Acquisition 2(2): 167-178.
26. Schweickert, R., Burton, A., Taylor, N., Corlett, E., Shadbolt, N., Rugg, G. y Hedgecock, A.; 1987. *Comparing Knowledge Elicitation Techniques: A Case Study*. Artificial Intelligence Review (1): 245-253.