Establecimiento del Modelo de Agregación más Apropiado para Ingeniería del Software

Hernán Guillermo Amatriain[1], Eduardo Diez[1], Rodolfo Bertone[2], Enrique Fernandez[1,3]
1. Grupo de Investigación en Sistemas de Información. Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús. Argentina.
2. Instituto de Investigaciones en Informática LIDI. Facultad de Informática. Universidad Nacional de La Plata. Argentina.
3. Cátedra de Sistemas de Programación no Convencional de Robots. Facultad de Ingeniería. Universidad de Buenos Aires. Argentina
hamatriain@unla.edu.ar

***Resumen*—Antecedentes: la síntesis cuantitativa consiste en integrar los resultados de un conjunto de experimentos, previamente identificados, en una medida resumen. Al realizar este tipo de síntesis, se busca hallar un resultado que sea resumen representativo de los resultados de los estudios individuales, y por tanto que signifique una mejora sobre las estimaciones individuales. Este tipo de procedimientos recibe el nombre de Agregación o Meta-Análisis. Existen dos estrategias a la hora de agregar un conjunto de experimentos, la primera parte del supuesto de que las diferencias en los resultados de un experimento a otro obedecen a un error aleatorio propio de la experimentación y de que existe un único resultado o tamaño de efecto que es compartido por toda la población, la segunda estrategia parte del supuesto de que no existe un único tamaño de efecto representativo de toda la población, sino que dependiendo del origen o momento en que se realicen los experimentos los resultados van a modificarse debido a la influencia de variables no controladas, a pesar de esto puede obtenerse un promedio de los distintos resultados para una conclusión general. A la primera de las estrategias se la denominada modelo de efecto fijo y a la segunda se la denominada modelo de efectos aleatorios. Los autores que han comenzado a trabajar en Meta-Análisis, no muestran una línea de trabajo unificada. Este hecho hace que sea necesaria la unificación de criterios para la realización de este tipo de trabajos. Objetivo: establecer un conjunto de recomendaciones o guías que permitan, a los investigadores en Ingeniería del Software, determinar bajo qué condiciones es conveniente desarrollar un Meta-Análisis mediante modelo de efecto fijo y cuando es conveniente utilizar el modelo de efectos aleatorios. Métodos: la estrategia sería la de obtener los resultados de experimentos de características similares mediante el método de Monte Carlo. Todos ellos contarían con un número de sujetos bajo, ya que esa es la característica principal en el campo de la Ingeniería de Software y que genera la necesidad de tener que agregar el resultado de varios experimentos. Luego se agrega el resultado de estos experimentos con el método de Diferencia de Medias Ponderadas aplicada primero con el modelo de efecto fijo, y posteriormente con el modelo de efectos aleatorios. Con las combinaciones realizadas, se analiza y compara la fiabilidad y potencia estadística de ambos modelos de efectos. Resultados: el modelo de efecto fijo se comporta mejor que el modelo de efectos aleatorios, presentando potencia con más de 80 sujetos/experimentos cuando el modelo de efecto aleatorio no posee potencia en ninguno de los casos analizados, y fiabilidad para todos los casos en que la varianza es baja o media. Cuando los efectos poblacionales son altos o muy altos, el modelo de efecto fijo tiende a perder fiabilidad sobre todo cuando se incrementa la cantidad de experimentos y la cantidad de sujetos experimentales. Conclusiones: la baja potencia del modelo de efectos aleatorios dentro del contexto de simulación desarrollado, provoca en la práctica que el resultado final del meta-análisis hecho con este tipo de modelo tienda a dar diferencias no significativas en todo momento, no permitiendo de esta forma poder afirmar que un tratamiento es mejor que otro cuando en realidad lo es.**

***Palabras Claves*—meta-análisis, agregación estadística, modelo efecto fijo, modelo efectos aleatorios, síntesis cuntitativa, diferencia de medias ponderadas**

I. Introducción

A. Meta-Análisis como herramienta para la construcción de conocimiento

A.1. Necesidad del Meta-Análisis

Hoy día es difícil poder pensar que una disciplina que se considere una ciencia no posea un procedimiento Empírico propio para validar la calidad de los conocimientos que se aplican en la misma. En este sentido es difícil pensar que en el mediano plazo la Ingeniería de Software pueda resolver los problemas de escases de sujetos experimentales para el desarrollo de los experimentos con un alto nivel de evidencia.

La cantidad de estudios experimentales en Ingeniería de Software se ha incrementado significativamente en los últimos años en cuanto a cantidad de experimentos [107]. Uno de los principales problemas a los cuales se enfrentan los investigadores en Ingeniería de Software a la hora de generar evidencias empíricas, es la imposibilidad de desarrollar un experimento de gran envergadura (que contenga gran cantidad de sujetos experimentales). Por en ejemplo, en el trabajo de revisión de [13], donde se identifican 21 experimentos, el promedio de sujetos experimentales por experimento ascienda a 6. Si bien las conclusiones de estos trabajos aportan cocimientos interesantes, difícilmente estos trabajos de forma aislada puedan generar las evidencias necesarias para mejorar la calidad de los conocimientos que se aplican en la industria del software, debido a la baja representatividad de las muestras. Esto produce que las nuevas innovaciones en la industria del software se apliquen porque se asumen que serán útiles debido al respeto o fama de las personas que las formulan [67].

Ahora bien, si se deja de ver a los experimentos de manera aislada y los mismos se analiza en forma conjunta mediante su agregación, como se hizo, por ejemplo, en los trabajos de [31, 13], pasamos de tener conocimientos abalados por cuatro o seis sujetos experimentales, a tener evidencias abaladas por más de 100 sujetos, lo cual mejora en gran medida la calidad de la conclusión. No solo porque a la vista de quien analiza los resultados las conclusiones se apoyan en mayor nivel de evidencia, sino también por los métodos estadísticos se

comportan más eficientemente (poseen menor error de tipo I y II) cuando las muestras son mayores.

A.2. Antecedentes

Según [6] la síntesis de resultados consiste en integrar los resultados de un conjunto de experimentos, previamente identificados, en una medida resumen. Al realizar este tipo de síntesis, se busca hallar un resultado que sea resumen representativo de los resultados de los estudios individuales, y por tanto que signifique una mejora sobre las estimaciones individuales. Idealmente, se debe partir de los estudios individuales -con sus virtudes y defectos- y obtener un resultado que sea más fiable que los resultados individuales de los que partíamos. Este tipo de procedimientos recibe el nombre de Agregación, Síntesis de resultados experimentales o Meta-Análisis. Este último término fue definido por [43] y significa análisis después del análisis en alusión a que los elementos con lo que trabajan los Meta-Análisis son los experimentos desarrollados y analizados por sus autores primarios.

En la síntesis de resultados existen dos estrategias bien diferenciadas a la hora de agregar un conjunto de experimentos. La primera de ellas parte del supuesto de que las diferencias en los resultados de un experimento a otro obedecen a un error aleatorio propio de la experimentación y de que existe un único resultado o tamaño de efecto que es compartido por toda la población. La segunda estrategia parte del supuesto de que no existe un único tamaño de efecto representativo de toda la población, sino que dependiendo del origen o momento en que se realicen los experimentos los resultados van a modificarse debido a la influencia de variables no controladas, a pesar de esto puede obtenerse un promedio de los distintos resultados para una conclusión general. A la primera de las estrategias se la denominada modelo de efecto fijo y a la segunda se la denominada modelo de efectos aleatorios [6].

Dentro de este contexto, los autores que han comenzado a trabajar concretamente en Meta-Análisis, no muestran una línea de trabajo unificada, por ejemplo en el Meta-análisis desarrollado por [31], donde se agregan 15 experimentos vinculado a la programación de a pares, se aplican ambos modelos de agregación, en cambio en [13] donde se agregan experimentos vinculados a técnicas de inspección de código en tres grupos de 5, 7 y 9 experimentos respectivamente se aplica únicamente el modelo de efecto fijo. Este hecho hace que sea necesaria la unificación de criterios para la realización de este tipo de trabajos.

A.3. Ejemplos

Tómese como ejemplo el intento de determinar de manera experimental que paradigma de programación es más adecuado para resolver un tipo de problema específico o que lenguaje de desarrollo presenta una curva de aprendizaje más abrupta, se verá que no es sencillo encontrar sujetos experimentales de similares características, es decir, que se presenten homogéneos ante el experimento.

En el primer ejemplo, podría tratarse de determinar si la programación estructurada es más adecuada para el desarrollo de una aplicación de manejo de inventario utilizando un motor de base de datos relacional, o la programación orientada a objetos es una opción más acertada. Para ello, hay que determinar un conjunto de características que deberán ser tenidas en cuenta. Por ejemplo, ¿cuáles van a ser las variables a medirse para determinar que paradigma es más adecuado?, ¿el tiempo de desarrollo?, ¿el tamaño del sistema (que a su vez presenta el inconveniente de decidir cómo se mide el tamaño)?, ¿la performance a la hora de utilizarse? Por otro lado, habría que determinar que lenguaje de programación utilizar para cada paradigma o si se usarán más de un lenguaje por paradigma (lo que dificulta aún más encontrar sujetos experimentales). Estas características hacen que a la hora de realizar el mismo experimento, distintos investigadores en distintos puntos geográficos puedan tomar decisiones distintas, haciendo más dificultosa la agregación. Pero volviendo al tema de discusión, un investigador debería hallar programadores para cada desarrollo. Si pudiera encontrar 10 programadores, tendría que utilizar 5 para el tratamiento de control (programación estructurada) y los otros 5 para el tratamiento experimental (programación orientada a objetos). La cantidad de 10 programadores no es un número sencillo, menos si se requieren desarrolladores ya formados. En ese caso, el investigador se encontrará que no todos los sujetos son egresados de la misma facultad, que no todos tienen la misma edad, y por tanto no todos tienen la misma experiencia, además que la experiencia la habrán obtenido en lugares distintos, cada uno con sus metodologías de trabajo distintas (no solo se encontrará con desarrolladores egresados de distintas facultades, sino que tendrá que los sujetos vienen de distintas experiencias laborales).

Para el segundo ejemplo, si se quiere determinar si es más simple aprender java o VB.NET, habrá que encontrar desarrolladores con conocimientos en programación orientada a objetos, pero que no conozcan estos lenguajes. Otra vez, los sujetos experimentales serán pocos (hay que buscar programadores) y vendrán de distintos centros universitarios. Para un experimento de este tipo, pueden buscarse estudiantes, en cuyo caso, podría suponerse que es más sencillo llegar a un número más adecuado. Sin embargo, si todos los estudiantes vienen de la misma facultad, entonces probablemente manejen los mismos lenguajes, por lo que habría que buscar estudiantes de distintas carreras, lo que hace que los sujetos sean forzosamente heterogéneos.

Por otro lado, si los sujetos experimentales tienen horarios laborales distintos, quizás el experimento deba desarrollarse en distintos horarios, teniendo que desarrollar el experimento con algunos sujetos antes de su horario laboral, y con otros después de su horario laboral, haciendo del cansancio un factor no controlable que influirá en la investigación experimental. En este caso, parecería adecuado utilizar el modelo de efectos aleatorios, pero no hay consenso teórico ni evidencia práctica que lo asevere.

A.4. Aplicabilidad de los Modelos al actual contexto de la Ingeniería de Software

Si se tiene en cuenta la gran diversidad cultural de la personas que trabajan a nivel mundial en la Ingeniería de Software, la evolución de las carreras universitarias en cuanto a la formación de los profesionales del área o la influencia de la experiencia laboral, se puede suponer que lo más apropiado para esta disciplina sería utilizar un modelo de efectos aleatorios, ya que la variedad de factores no controlados de los experimentos es alta y su influencia en los resultados puede asumirse que también lo es.

Ahora bien, que pasa con las advertencia de [6], en Ingeniería de Software no solo los Meta-Análisis poseen pocos experimentos, sino que también los experimentos poseen pocos sujetos (cosa que no sucede en Medicina donde los autores analizan el comportamiento de los métodos), por ello no se puede afirmar a priori que el modelo de efectos aleatorios sea el apropiado para esta rama de la ciencia, por lo tanto ambas

estrategias serán tenidas en cuenta dentro de la presente investigación.

Según [6], el desarrollo de un Meta-Análisis debe partir de la agregación de los experimentos mediante un modelo de efecto fijo, ya que a priori no se sabe si los resultados de los experimentos muestran incompatibilidades o no. Luego se debe verificar la existencia o no de incompatibilidades entre los experimentos agregados mediando un análisis de heterogeneidad. En caso de no observarse evidencia de heterogeneidad, el proceso de da por concluido, en caso contrario se podrá agregar los experimentos mediante un modelo de efectos aleatorios.

Sin embargo hay estudios que demuestran que los métodos para análisis de heterogeneidad no tienen potencia con pocos experimentos [6]. De acuerdo a [50], cuando se combinan estadísticamente el resultado de varios experimentos a través de los métodos de Meta-Análisis y se estudia la heterogeneidad, si la agregación se realizó con un número bajo de experimentos, entonces el análisis de heterogeneidad no tiene potencia, por lo que no es determinante.

B. Objetivo de la Investigación

Por lo expuesto hasta aquí, no se puede asegurar cual de los dos modelos de efecto de Meta-Análisis se adapta mejor al actual contexto experimental de la Ingeniería de Software.

No hay un argumento teórico que permita determinar dentro del actual contexto de la Ingeniería de Software cual sería el modelo adecuado a emplear en la estrategia de Meta-Análisis, ni evidencia empírica al respecto.

Por otro lado, y debido a la diferencia de opiniones entre los investigadores sobre el modelo a utilizar, se vislumbra la necesidad de zanjar esta cuestión.

Por tanto, el objetivo del presente trabajo de investigación reside en establecer un conjunto de recomendaciones o guías que permitan, a los investigadores en Ingeniería del Software, determinar bajo qué condiciones es conveniente desarrollar un Meta-Análisis mediante modelo de efecto fijo y cuando es conveniente utilizar el modelo de efectos aleatorios, en las actuales condiciones experimentales de la Ingeniería del Software.

II. Estado de la Cuestión

A. Meta-Análisis

A.1. Experimentos y Agregación de Experimentos

Puede observarse en los últimos 20 años el incremento de la cantidad de experimentos realizados dentro del ámbito de la Ingeniería del Software (IS) [107]. Entre los años 1993 y 2003 solo se publicaron 93 experimentos [32] en journals y conferencias de primer nivel, como es por ejemplo IEEE Transactions on Software Engineering. Sin embargo, la cantidad de experimentos desarrollados dentro del campo de la Ingeniería de Software fue mucho mayor, duplicando o incluso triplicando la cifra antes mencionada [26]. Estos experimentos abarcan diversas áreas, tales como el desempeño de las técnicas de testing, la educción de requerimientos, o la performance de los lenguajes de programación, por citar algunos. Si bien los experimentos aportan conocimientos interesantes en cada caso, en general son pequeños (rara vez utilizan más de 20 sujetos experimentales [24]), por ello para que la información que aportan sea valiosa los resultados deben agregarse para poder obtener conclusiones avaladas con la mayor evidencia empírica posible.

La agregación de experimentos consiste en combinar los resultados de varios experimentos, que analizan el comportamiento de un par de tratamientos específico, en un contexto determinado, para obtener un único resultado final. El nuevo resultado será más general y fiable que los resultados individuales, porque el mismo estará sustentado por un mayor nivel de evidencia empírica [15].

Las distintas estrategias de combinación de resultados experimentales se conocen con el nombre genérico de métodos de síntesis [12] o métodos de agregación [16], como típicamente acostumbran a denominarse en Ingeniería del Software.

Para que los resultados de un proceso de agregación sean fiables, se incluyen dentro de una Revisión Sistemática (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 [46]. Para combinar los resultados de los estudios individuales, las RS utilizan el Meta-Análisis como estrategia de agregación.

El Meta-Análisis es un nombre colectivo que hace referencia a un conjunto de métodos estadísticos que intentan hallar un resultado numérico que sea resumen representativo de los resultados de los estudios individuales, y por tanto que signifique una mejora sobre las estimaciones individuales.

Es importante destacar que para poder aplicar Meta-Análisis se deben verificar ciertas restricciones, tales como un número mínimo de experimentos, adecuadamente recopilados y homogéneos [46]. De esta forma se podrá garantizar que la conclusión alcanzada es realmente sólida y fiable.

Si bien la agregación de experimentos no es un tema nuevo para ciencias como la Psicología, la educación o la Medicina, recién a mediados de los 90' comienza a ser propuesta como una alternativa para generar conocimiento en Ingeniería de Software [4]. Desde entonteces varios autores abordaron el tema. Se puede citar al procedimiento de RS desarrollado por [73], el cual surge de una adaptación de los procesos de RS desarrollados en medicina, donde se contemplan diversos niveles de calidad de experimento acorde al actual contexto de la Ingeniería de Software. En dicho procedimiento, de la misma manera de lo que se hace en medicina, se recomienda el uso del método estadístico Diferencia Medias Ponderadas (DMP) [55] para la agregación de los resultados.

En [31] se ha trabajado en la aplicación de las RS, donde se identifican 11 estudios experimentales vinculados al desempeño de los programadores cuando trabajan de a pares y se combinan sus resultados mediante un método de Meta-Análisis estándar (aplicando el método de Diferencias Medias Ponderadas como se sugiere en [73]).

Ahora bien, no todos los trabajos vinculados a la Agregación hechos en el ámbito de la Ingeniería de Software han sido exitosos. En [3, 105, 59, 116, 69, 65], si bien los autores pudieron desarrollar el procedimiento de búsqueda y selección de experimentos, la combinación de los mismos mediante Meta-Análisis resultó impracticable. Los principales motivos que limitaron la aplicación del Meta-análisis, en el actual contexto de la Ingeniería del Software, se vinculan con los siguientes problemas:

- La escasez de experimentos, replicaciones y homogeneidad entre los mismos. [24, 84].
- La carencia de estándares para reportes de experimentos. Por ejemplo, [9] no publican varianzas y [11] ni siquiera reporta las medias de los resultados experimentales.
- La falta de estandarización de las variables respuesta. Los trabajos de [1, 117] utilizan diferentes variables respuesta

para analizar un mismo aspecto, lo cual hace que estos experimentos no puedan ser agregados.

La Ingeniería de Software no es la única ciencia que padece el problema de escasez de experimentos y que posee altos costos para el desarrollo de los mismos. Un ejemplo es la ecología, donde los costos y el tiempo necesario para evaluar el crecimiento de algunas especies son elevados [118]. Tampoco la Ingeniería de Software es la única ciencia que tiene problemas vinculados a la calidad de los estudios desarrollados (en general solo se trabaja con ensayos de laboratorio), los cuales limitan la calidad o fiabilidad de las conclusiones obtenidas, ya que este problema surge en ciencias mucho más comprometidas con la experimentación como lo es la medicina [48, 104].

A.2. Modelos de Meta-Análisis

Según [6] la síntesis de resultados consiste en integrar los resultados de un conjunto de experimentos previamente identificados, en una medida resumen. Al realizar este tipo de síntesis, se busca hallar un valor que sea representativo de los resultados de los estudios individuales, y por tanto que signifique una mejora sobre las estimaciones individuales. Se parte de los estudios individuales -con sus virtudes y defectos- y se obtiene un resultado que es más fiable que los resultados individuales de los que se parte. Este tipo de procedimientos recibe el nombre de Agregación, Síntesis de resultados experimentales o Meta-Análisis. Este último término fue definido por [43] y significa análisis después del análisis en alusión a que los elementos con lo que trabajan los Meta-Análisis son los experimentos desarrollados y analizados por sus autores primarios.

Existen dos estrategias para la agregación de experimentos: modelo de efecto fijo y modelo de efectos aleatorios. La primera de ellas parte del supuesto que las diferencias en los resultados de un experimento a otro obedecen a un error aleatorio propio de la experimentación y que existe un único resultado o tamaño de efecto que es compartido por toda la población. La segunda estrategia parte del supuesto que no existe un único tamaño de efecto representativo de toda la población, sino que dependiendo del origen o momento en que se realicen los experimentos los resultados van a modificarse debido a la influencia de variables no controladas, a pesar de esto puede obtenerse un promedio de los distintos resultados para una conclusión general. A la primera de las estrategias se la denominada modelo de efecto fijo y a la segunda se la denominada modelo de efectos aleatorios [6].

Las Figura 1 y 2 ilustran lo dicho anteriormente, mostrando en la Figura 1 un único tamaño de efecto y un conjunto de resultados que se diferencias del mismo únicamente por un error propio de la experimentación, y en la Figura 2 varios tamaños de efecto, donde los experimentos también poseen error experimental y un promedio general de los tamaños de efecto [103].

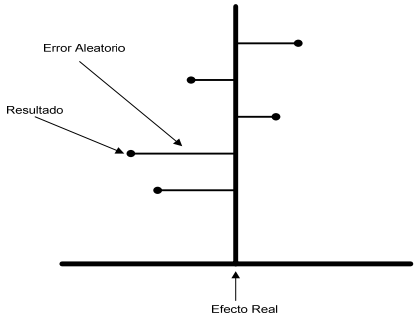

Figura 1: Supuestos del modelo de efecto fijo

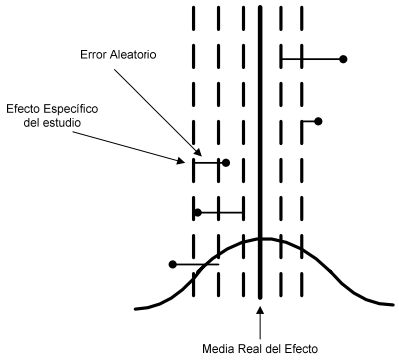

Figura 2: Supuestos del modelo de efectos Aleatorios

En las siguientes subsecciones se explica con más detalle en qué consiste cada una de las estrategias, sus diferencias a la hora de evaluar la ponderación de un experimento dentro de la conclusión y su aplicabilidad en Ingeniería de Software.

A.3. Modelo de efecto fijo

Para el modelo de efecto fijo existe un único tamaño de efecto al cual pertenecen todos los experimentos que van a ser agregados. Por ende, cualquier diferencia en los resultados obedece únicamente a un error experimental aleatorio propio de la experimentación [6]. Por ello, dentro de este enfoque la ponderación de los experimentos se realiza únicamente en base a la inversa de su varianza, asumiendo que cuanto menor sea la varianza más preciso es el experimento. Dado que la varianza es inversamente proporcional al tamaño del experimento (cantidad de sujetos que posee) tendremos que los experimentos de mayor tamaño tendrán una mayor representatividad en la conclusión que los pequeños. Un experimento con 1.000 sujetos experimentales tendrá una ponderación 10 veces mayor que la que tiene una de 100 sujetos experimentales [6]. Esto produce que la conclusión general se vuelque hacia el resultado particular de un estudio cuando este es mucho más grande que los demás.

A.4. Modelo de efectos aleatorios

A diferencia del modelo de efecto fijo, para el modelo de efectos aleatorios existe más de un tamaño de efecto, por ende existen dos tipos de errores, el error propio de cada uno de los experimentos producto de la experimentación (como sucede con el modelo de efectos fijo) y el error producido por la combinación de estudios provenientes de distintos tamaños de efecto [6]. Este se traduce en la estimación de dos tipos de varianzas, la varianza interna de los estudios y la varianza entre estudios. De esta forma los experimentos reciben una "doble ponderación", la cual tiende a mitigar la influencia de los experimentos grandes en la conclusión general haciendo más representativos a los experimentos con menos sujetos, ya que a diferencia del modelo de efecto fijo cada experimento puede estar aportando un tamaño de efecto diferente.

La inclusión de la varianza entre experimentos trae aparejado un nuevo problema, el error asociado a su estimación, el cual se incremente cuando el Meta-Análisis posee pocos experimentos. Por ello, autores como [6], no recomiendan su uso cuando el Meta-Análisis posea pocos experimentos (en la práctica menos de 10).

A.5. La estadística y el error experimental

Dado que para los investigadores resulta imposible censar o evaluar a todos los individuos de una población, los conocimientos científicos se infieren a través de experimentos realizados sobre muestras que se consideran representativas de

dicha población [41]. Sir Ronald Fisher [38] demostró que los métodos estadísticos son útiles para tratar estos problemas y que los mismos se enfrentan a tres dificultadas principales:

- Error de experimentación, también conocido como ruido, que es causado por factores distorsiónales y, en ciertos casos por problema al realizar la medición.
- Confusión entre correlación y causalidad, que se produce cuando se confunde la influencia entre variables, provocando así que se piense que dos variables son dependientes entre sí, cuando en realidad dependen de una tercera que no se toma en cuenta.
- Complejidad de combinaciones entre varias variables y una tercera, sucede cuando dos o más variables afectan a una tercera de distinta manera, dependiendo de los valores que toma cada una. Entonces no existe una fórmula lineal y directa que prediga los resultados.

Estos problemas pueden ser solucionados con la estadística de la siguiente manera:

- El error de experimentación puede ser reducido por un diseño y análisis apropiado de los experimentos. Al utilizar un análisis estadístico que provea de formas para medir la precisión de la cantidad bajo estudio y juzgar cuando hay fuerte evidencia empírica para atribuir a ciertas razones las diferencias observadas.
- La confusión de correlación y causalidad puede ser solucionado utilizando los principios del diseño experimental y "aleatorización", para generar datos de mayor calidad e inferir las relaciones causales verdaderas.

Bajo estos supuestos, los métodos estadísticos están sometido a dos tipos de errores [17]: α, o error de tipo I, y β, o error de tipo II. Dichos errores se producen por la incertidumbre asociada a estimar parámetros (básicamente medias y desvíos típicos) de una población a partir de una muestra de la misma. Tal y como indica la Tabla I, α es el error asociado a aceptar la hipótesis alternativa (H1) cuando en la población se verifica la hipótesis nula (H0), y β es la probabilidad asociada al evento justamente inverso.

TABLA I. TIPOS DE ERROR DE UN TEST ESTADÍSTICO

	H0 verificada en a población	**H1 verificada en la población**
H0 respuesta del experimento	Decisión correcta (1-α)	β (Tipo II error)
H1 aceptada respuesta del experimento	α (Tipo I error)	Decisión correcta (1-β)

Según la teoría estadística, un test de hipótesis, se basa en 4 factores [34]: α, β, el tamaño del efecto (d) (el cociente entre la diferencia entre las medias y la varianza para el caso del DMP o el cociente de las medias para RR) y el número de sujetos experimentales (n) o, dicho con mayor precisión, el tamaño de la muestra. La relación entre estos factores se muestra en la Fórmula 1 (donde z representa la distribución normal tipificada):

$$z_{1-\beta} = \sqrt{\frac{n}{2}}d - z_{1-\alpha}$$

Fórmula 1: relación entre los factores de un test estadístico

Como se ve, los 4 factores implicados en la fórmula 1 forman un sistema cerrado, haciendo que una disminución o incremento en cualquiera de los factores provoque incrementos o disminuciones en los demás factores, donde solo d depende de las condiciones del experimento y no puede ser controlado o manipulado por el investigador. De los demás factores podemos decir que para un investigador, el error más importante es α. La razón es sencilla: toda la comunidad informática intenta desarrollar nuevos métodos y técnicas que hagan más eficiente el desarrollo del software. Pero necesitamos demostrar que dichos métodos y técnicas son efectivamente mejores. El punto aquí es que si nuestra conclusión es falsa podemos movilizar a la comunidad del software a un cambio innecesario que solo genere gastos. Por este motivo, si el test arroja que H1 es cierto (o sea, acontece la segunda fila de la tabla II que es lo que realmente queremos) esto debe estar acompañado de un error de α lo menor posible (donde generalmente α = 0.05), lo cual hace que el resultado sea fiable. Por ello, al complemento del error α se lo llama fiabilidad y se estima como 1 - α. De los 2 factores restantes, según la teoría estadística la variable de ajuste debería ser el tamaño de la muestra, pero la realidad indica que en Ingeniería de Software (como también sucede en otras ramas de la ciencia) rara vez se cuente con el presupuesto necesario para realizar un experimento que implique gran cantidad de sujetos, por ello, el error de tipo II se verá incrementado o lo que es igual el test perderá potencia estadística, siendo la potencia estadística la capacidad que el test posea para determinar que H1 es verdadera [34]. La relación que existe entre los distintos parámetros, descripta anteriormente, puede apreciarse en la Figura 3:

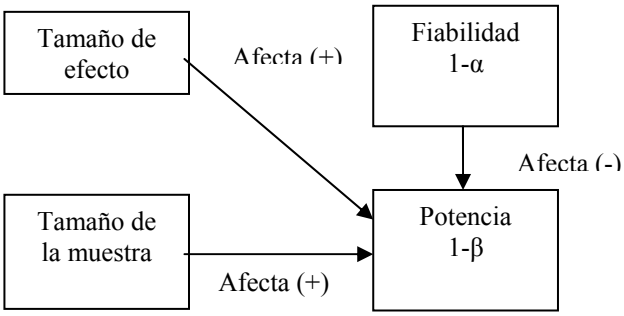

Figura 3: Relación entre factores que afectan la potencia estadística de un test

A.6. Técnicas de Agregación de Experimentos

Existen varias técnicas de agregación cuantitativas, podemos mencionar dentro de las más conocidas [54]:

- Diferencia de medias ponderadas (DMP)
- Response Ratio (RR) paramétricos
- Response Ratio (RR) no paramétricos
- Vote Counting (conteo de votos)

A.6.1. Diferencia de Medias Ponderadas:

La técnica de diferencia de medias ponderadas [55] es la técnica de estimación de tamaño de efecto, o mejora de un tratamiento respecto de otro, más conocida y difundida para el análisis de variables continuas. Esta técnica es conceptualmente sencilla: el estimador de efecto individual (para cada experimento) se estima como el cociente de las diferencias entre las medias y el desvío estándar y el efecto global se calcula como una media ponderada de los estimadores de efecto de los estudios individuales.

La estimación del efecto individual consiste en estimar, para un estudio particular, si el tratamiento Experimental es mejor o no que el tratamiento de control. Estos se hace dividiendo la diferencia de medias de ambos grupos por la varianza conjunta [55]. La estimación del efecto global se realiza como la suma ponderada de los efectos individuales [6, 55]. Donde cada estudio es ponderado en función de su tamaño y la inversa de la varianza, de esta forma los estudios que

incluyan mayor cantidad de sujetos experimentales y posean una menor varianza recibirán una mayor ponderación, por considerar que sus resultados son más fiables, que los estudios más pequeños.

A.6.2. Response Ratio paramétrico:

Para estimar el Response Ratio (RR) de un estudio particular, como se mencionó anteriormente, se debe dividir la media del tratamiento experimental por la media del tratamiento de control [54]. Si bien, realizar en forma directa el cociente de ambas medias permite obtener un índice de mejora para un estudio en particular, para que la combinación de un conjunto de estudios sea más precisa se le incorporó el logaritmo natural [54, 82]. Esto permite linealizar los resultados (mientras que el RR es afectado más por los cambios en el denominador que en el numerador, el Ln (RR), gracias a las propiedades de los logaritmos, afecta de modo parejo al numerador y al denominador y así normalizar su distribución, convirtiéndolo en un método apropiado para estimaciones de conjuntos de experimentos pequeños.

Una vez estimado el tamaño de efecto, podrá estimarse el intervalo de confianza. Para estimar el error típico, esta versión del Response Ratio no requiere conocer las varianzas, como lo hace la versión original. En su lugar hace una estimación en base a la cantidad de sujetos y el response ratio. Una vez estimados el intervalo de confianza, se debe aplicar el anti-logaritmo a los resultados para obtener nuevamente el índice de relación. Es importante destacar que esta situación traer aparejado que el nuevo intervalo de confianza no sea simétrico.

La estimación del efecto global se realiza mediante la suma ponderada de los efectos individuales [63]. Donde, a semejanza de lo que sucede con las diferencias medias ponderadas, cada estudio es ponderado en función de su tamaño y la inversa de la varianza.

A.6.3. Response Ratio no paramétrico:

La estimación del Response Ratio no paramétrico consiste en dividir la media del tratamiento Experimental por la media del tratamiento de Control [54] como en el caso anterior. Para estimar el error típico (necesario para establecer el intervalo de confianza), esta versión del Response Ratio no requiere conocer las varianzas, como lo hace la versión original. En su lugar hace una estimación en base a la cantidad de sujetos y el response ratio [118]

A.6.4. Conteo de Votos:

El Vote Counting es un método que requiere muy poca información para poder ser aplicado, básicamente conocer si existe o no diferencia entre las medias de los tratamientos y la cantidad de sujetos experimentales utilizados en el estudio experimental. Si bien existen varias versiones de esta técnica, en este apartado se describirá la versión desarrollada por [55]. Esta versión permite estimar el tamaño de efecto partiendo del signo de las diferencias de las medias y la cantidad de sujetos experimentales, los cuales se combinan mediante la aplicación de la función de verosimilitud. Esta función que permite establecer, en base al signo de la diferencia de medias y la cantidad de sujetos, cual es el valor de efecto que tiene mayor probabilidad de ocurrencia [55]. Una vez establecido el efecto de mayor probabilidad se podrá determinar el intervalo de confianza para el mismo, el cual es más amplio que el estimado mediante DMP [55]

A.6.5. Técnica de agregación elegida para el desarrollo de la investigación:

Cada técnica antes mencionada tiene sus ventajas y desventajas. Sin embargo, fueron desarrolladas para contextos experimentales maduros, que no es el caso de la Ingeniería de Software. Como se comprobó en [37], en contextos experimentales pocos maduros la técnica DMP ha demostrado mejor potencia y fiabilidad general que los otros métodos. Por ello, se escoge esta técnica para analizarla bajo los dos supuestos de efecto de Meta-Análisis: modelo de efecto fijo y modelo de efectos aleatorios.

A.7. Diferencia de Medias Ponderadas aplicada bajo el supuesto de modelo de efecto fijo

El método Diferencias Medias Ponderadas (DMP) [43], como se mencionó anteriormente, es la técnica de estimación de tamaño de efecto (effect size), o mejora de un tratamiento respecto de otro, más conocida y difundida para el tratamiento de variables continuas. El estimador de efecto individual se estima como el cociente de las diferencias entre las medias y el desvío estándar y el efecto global se calcula como una media ponderada (donde cada estudio es ponderado en base a la inversa de su varianza) de los estimadores de efecto de los estudios individuales. Este método fue desarrollado originalmente por [43] y optimizado por [55], quienes incorporaron a la función de estimación de efecto individual una variable de corrección que permite mejorar la precisión del método cuando los estudios experimentales poseen pocos sujetos. Es un método de tipo paramétrico el cual requiere para ser aplicado normalidad en la distribución y homoesticidad (igualdad de las varianzas). Estos aspectos pueden estimarse por el investigador primario o pueden asumirse en función de las características del fenómeno que se está tratando.

Como los resultados que arroja este método (tamaño de efecto) son poco comprensibles, para facilitar la comprensión de los resultados obtenidos, se ha tabulado un conjunto de valores de corte a partir de los cuales los resultados pueden considerarse: Nulos (no existe diferencia entre los tratamientos analizados), Bajos (si bien uno de los tratamientos es mejor que el otro, la diferencia en el desempeño no es muy importante), Medios (uno de los tratamientos es mejor que el otro, y la diferencia a favor es considerables) o Altos (uno de los tratamientos es mucho mejor que el otro). En la Tabla II, se describe el análisis de resultados realizado por [111], donde se puede observar cual es el nivel de solapamiento o coincidencia entre los resultados en función del tamaño de efecto obtenido.

TABLA II. INTERPRETACIÓN DEL TAMAÑO DE EFECTO

Nivel de diferencia	Tamaño de efecto (d)	Porcentaje de no solapamiento de los tratamientos
	2.0	81.1%
	1.9	79.4%
	1.8	77.4%
	1.7	75.4%
	1.6	73.1%
	1.5	70.7%
	1.4	68.1%
	1.3	65.3%
	1.2	62.2%
	1.1	58.9%
	1.0	55.4%
	0.9	51.6%
Alto	**0.8**	**47.4%**
	0.7	43.0%
	0.6	38.2%
Medio	**0.5**	**33.0%**
	0.4	27.4%
	0.3	21.3%
Bajo	**0.2**	**14.7%**
	0.1	7.7%
Nulo	**0.0**	**0%**

En las siguientes subsecciones se describen las funciones de estimación del método y las ventajas y desventajas asociadas a su posible uso en Ingeniería de Software

A.7.1. Descripción del método

La aplicación del método consta de dos pasos, primeramente se debe estimar el tamaño de efecto de cada uno de los experimentos, y una vez estimado el mismo, podrá estimarse el tamaño de efecto global. A continuación vamos a presentar las función de estimación del tamaño de efecto para un experimento (o efecto individual) desarrollada por [43] (ver Fórmula 2).

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

g representa el tamaño de efecto
Y's representa a las medias del grupo experimental (E) y de control (C)
Sp representa el desvió estándar conjunto

Fórmula 2: Tamaño de efecto individual

La Fórmula 2 tiende a sobre estimar los efectos de los estudios pequeños (que cuentan con pocos sujetos experimentales), y, como se mencionó anteriormente, fue optimizada por [55], quién incorporó un ponderador que ayuda a reducir el error de los experimentos más pequeños (ver Fórmula 3). Esta nueva versión del método es la más conocida en Ingeniería de Software ([73, 31]).

$$d = J\frac{Y^E - Y^C}{S_P} \qquad J = 1 - \frac{3}{4N - 9}$$

d representa el tamaño de efecto
J representa el factor de corrección
Y's representa a las medias del grupo experimental (E) y de control (C)
Sp representa el desvió estándar conjunto
N representa el total de sujetos experimentales incluidos en el experimento

Fórmula 3: Tamaño de efecto individual con factor de corrección

Luego de estimar el tamaño de efecto, se estima el error típico, y en base a este se establece el intervalo de confianza asociado al efecto para el nivel de fiabilidad deseado, generalmente del 95%, lo que equivale a un error de tipo I del 5% (α = 0,05). Ver Fórmula 4.

$$v = \frac{\tilde{n} + d^2}{2(n^E + n^C)} \qquad \tilde{n} = \frac{n^E + n^C}{n^E * n^C}$$

$$d - Z_{\alpha/2}\sqrt{v} \le \lambda \le d + Z_{\alpha/2}\sqrt{v}$$

v representa el error típico
d representa el tamaño de efecto
n's representa la cantidad de sujetos experimentales del grupo experimental (E) y de control (C)
Z representa la cantidad de desvíos estándar que separan, al nivel de significancia dado, la media del límite. En general es 1,96 (α = 0,05)

Fórmula 4: Error típico e intervalo de confianza

Una vez estimados los tamaños de efectos de los estudios individuales se debe estimar el tamaño de efecto global (ver Fórmula 5).

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

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

Fórmula 5: Tamaño de efecto grupal

Para ello se realiza una suma ponderada de los tamaños de efecto de cada uno de los estudios, donde cada estudio es ponderado en función de la inversa de la varianza [6, 55], de esta forma los estudios con mayor precisión (que en general son los que incluyen mayor cantidad de sujetos experimentales) recibirán una mayor ponderación por considerar que sus resultados son más fiables, o tienen menor posibilidad de incurrir en un error experimental.

Una vez estimado el tamaño de efecto global, se debe estimar el intervalo de confianza asociado al mismo, para ello se utiliza la Fórmula 6.

$$d^* - Z_{\alpha/2}\sqrt{v} \le \lambda \le d^* + Z_{\alpha/2}\sqrt{v}$$

d* representa el tamaño de efecto global
Z representa la cantidad de desvíos estándar que separan, al nivel de significancia dado, la media del límite. En general es 1,96 (α = 0,05)
v representa el error típico

Fórmula 6: Intervalo de confianza grupal

A.7.2. Ventajas y desventajas del método

En esta sección se presentan una serie de ventajas que se espera obtener si un conjunto de experimentos es agregado mediante DMP, y luego un conjunto de desventajas o inconvenientes para su aplicación:

Ventajas

- Existe gran cantidad de reportes que describan su uso en otras ramas de la ciencia, principalmente medicina y educación [16]
- Existe gran variedad de software que soporta su aplicación (por ejemplo [80, 85] entre otros)
- La función corregida por [55] minimiza el error de estimación cuando los estudios son pequeños lo cual es importante en el actual contexto de la Ingeniería de Software.
- Es el método recomendado por [73]
- Se conoce procesos de agregación que han podido aplicar esta técnica en estudios hechos en Ingeniería de Software [83, 31]

Desventajas

- Requiere la publicación de todos los parámetros estadísticos (medias, varianzas y cantidad de sujetos experimentales)
- Se debe verificar o suponer homosticidad y normalidad [55]
- Hay revisiones sistemáticas en Ingeniería de Software que no llegaron a combinar los resultados porque la calidad de los estudios identificados no cumplían con el mínimo de requisitos necesarios para aplicar este método, por ejemplo [27].

A.8. Diferencia de Medias Ponderadas aplicada bajo el supuesto de modelo de efectos aleatorios

La aplicación del método consta de dos pasos, primeramente se debe estimar el tamaño de efecto de cada uno de los experimentos, y una vez estimado el mismo, podrá estimarse el tamaño de efecto global [55]. Dado que la forma de estimar el tamaño de efecto para cada uno de los experimentos es la misma que la que se describió para el método DMP en la sección 2.4, en esta sección solo vamos a describir como se estima el tamaño de efecto global.

Este es un método de tipo paramétrico el cual requiere para ser aplicado normalidad en la distribución y homoesticidad (igualdad de las varianzas). Estos aspectos pueden estimarse por el investigador primario o pueden asumirse en función de las características del fenómeno que se está tratando.

La forma de interpretar los resultados es la misma que se utiliza con el modelo de efecto fijo, a modo de recordatorio.

En la Tabla III, se presenta un resumen de la Tabla II, donde se indican los valores de corte asociados a cada nivel de tamaño de efecto.

TABLA III. INTERPRETACIÓN DEL TAMAÑO DE EFECTO

Tamaño de efecto	Nivel de diferencia
0	Nulo
0.2	Bajo
0.5	Medio
0.8	Alto
1.2	Muy Alto

A.8.1. Descripción del método

Para estimar el tamaño de efecto de un grupo de experimentos, la función de estimación incluye, a demás de la varianza propia de cada experimento, la estimación de la varianza entre experimentos. Esto se debe a que el cálculo final es básicamente un promedio de tamaños de efectos el cual, como todo promedio, tiene una varianza asociada [6, 55]. Ver Fórmula 7.

$$\Delta = \frac{\sum d_i / \gamma^2_i}{\sum 1/\gamma^2_i}$$

Δ representa el tamaño de efecto global

$\sum d_i / \gamma^2_i$ representa la sumatoria de los efectos individuales

$\sum 1/\gamma^2_i$ representa la sumatoria de la inversa de las varianzas entre-estudios e intra-estudios

Fórmula 7: Tamaño de efecto global para el modelo de efectos aleatorios

Una vez estimado el tamaño de efecto global, se debe estimar el intervalo de confianza asociado al mismo, para ello se utiliza la Fórmula 8.

$$\Delta - Z_{\alpha/2}\sqrt{v} \leq \Delta \leq \Delta + Z_{\alpha/2}\sqrt{v}$$

$$v = \frac{1}{\sum 1/\gamma^2_i}$$

Δ representa el tamaño de efecto global

Z representa la cantidad de desvíos estándar que separan, al nivel de significancia dado, la media del límite. En general es 1,96 ($\alpha = 0{,}05$)

v representa el error típico

Fórmula 8: Error típico global e intervalo de confianza para el modelo de efectos aleatorios

A.8.2. Ventajas y desventajas del método

En esta sección se presentan una serie de ventajas que se espera obtener si un conjunto de experimentos es agregado mediante el modelo de efectos aleatorios, y luego un conjunto de desventajas o inconvenientes para su aplicación:

Ventajas

- Permite combinar estudios a pesar de que existen indicios de heterogeneidad entre los mismos
- Existe gran variedad de software que soporta su aplicación (por ejemplo [80, 85] entre otros)
- La función corregida de [55] minimiza el error de estimación cuando los estudios son pequeños lo cual es importante en el actual contexto de la Ingeniería de Software
- Fue aplicada en el procesos de agregación desarrollado por [31]

Desventajas

- Requiere la publicación de todos los parámetros estadísticos (medias, varianzas y cantidad de sujetos experimentales)
- Se debe verificar o suponer homosticidad y normalidad [55]
- Hay autores que afirman que el nivel de error que introduce la varianza entre experimentos, cuando el proceso de agregación contiene menos de 10 experimentos, es muy alto y no debe aplicarse en esos casos [6].

A.9. Heterogeneidad estadística

Es altamente probable que un conjunto de experimentos que analizan el desempeño de un par de tratamientos arrojen resultados diferentes, esto se debe fundamentalmente a la selección y asignación de sujetos experimentales de manera aleatoria. Pero también es esperable que estas diferencias no sean demasiado notorias, ya que si esto sucediera sería esperable que exista algún factor no controlado que está condicionando el resultado del estudio. En cuyo caso se dirá que los experimentos son heterogéneos.

Existen básicamente tres tipo de heterogeneidad: la "heterogeneidad estadística" (diferencias en los efectos reportados), "heterogeneidad metodológica" (diferencias en el diseño de los estudios) y "heterogeneidad de sujetos" (diferencias entre los estudios referidas a características clave de los participantes, nivel de experiencia profesional, formación académica, motivación, entre otros). La heterogeneidad puede deberse a diferencias entre los participantes, la variables respuesta, las métricas y un gran número de otros factores que pueden afectar al experimento y los participantes.

La heterogeneidad estadística puede observarse claramente mediante un diagrama de árboles, gráfico con el que habitualmente se presentan los resultados del meta-análisis. En estos gráficos se representan los tamaños de efectos de los estudios individuales, así como el tamaño de efecto global, juntamente con sus respectivos intervalos de confianza [108]. En la Figura 4, se presenta un ejemplo.

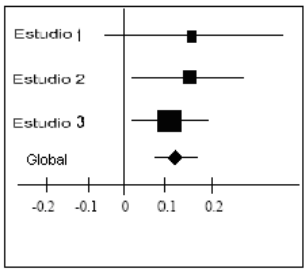

Figura 4: Ejemplo Diagrama de árboles, con experimentos homogéneos

Cuando los estudios de un Meta-Análisis son homogéneos, los intervalos de confianza de los mismos se solapan entre sí. Por el contrario, si el resultado de algún estudio no se solapa con ninguno de los intervalos de confianza de los otros experimentos, es bastante probable que este estudio no es homogéneo (o sea heterogéneo).

Si se analiza el ejemplo de la Figura 4, se puede decir que no existen indicios de heterogeneidad o, lo que es lo mismo, que existe homogeneidad entre los experimentos, ya que los intervalos de confianza de los tres experimentos se solapan entre sí. Por el contrario, si analizamos el ejemplo de la Figura 5, podemos decir que existen indicios de heterogeneidad, ya que el intervalo de confianza del estudio 2 no se solapa con los intervalos de confianza de los otros experimentos. Nótese que si bien estos gráficos son muy fáciles de construir e interpretar, los resultados que aportan son muy dependientes del meta-analista y por ello son cuestionados por algunos autores. Para ser más concretos en esta afirmación, en el ejemplo planteado, en la Figura 5, el estudio 2 no tiene solapamiento con los otros dos experimentos, pero es habitual que exista un leve solapamiento entre los intervalos de confianza de los experimentos, en estos casos dependiendo de quién evalúe los

resultados podrá plantearse que existe o no heterogeneidad. Por eso este tipo de análisis es criticado.

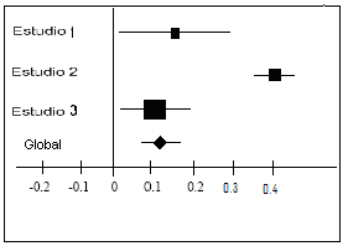

Figura 5: Ejemplo Diagrama de árboles, con experimentos heterogéneos

Los tests estadísticos de heterogeneidad se utilizan para valorar si la variabilidad en los resultados de los estudios (la magnitud de los efectos) es mayor que aquella que se esperaría hubiera ocurrido por azar [57]. La prueba más conocidas para valorarla la heterogeneidad estadística es el método Q propuesto por [25] el cual se basa en el test desarrollado por [14]. Este método en general es recomendado por cuestiones de validez y sencillez computacional [16] (Fórmula 9).

$$Q_T = \sum_{i=1}^{k} w_i E_i^2 - \frac{\left(\sum_{i=1}^{k} w_i E_i\right)^2}{\sum_{i=1}^{k} w_i} = \sum_{i=1}^{k} w_i \left(E_i - \overline{\overline{E}}\right)^2$$

k representa número de experimentos

wi representa el peso del estudio i (se corresponde con la inversa de la varianza del mismo)

Ei representa el tamaño de efecto del experimento

E representa el tamaño de efecto global

Fórmula 9: Estimador Q para análisis de heterogeneidad

Q posee una distribución Chi2 con K – 1 grados de libertad (k, tal y como se indica en la Fórmula 9 representa el número de experimento combinados en el meta-análisis)

Una vez calculado Q, si su valor obtenido es inferior a k-1, se considera que los experimentos son homogéneos. Si el resultado es superior a K-1 es posible que exista heterogeneidad experimental, y la posibilidad de tal existencia se determina utilizando la distribución Chi2 antes citada. El nivel de significación habitual es $\alpha = 0.05$, aunque algunos autores [103] recomiendan utilizar $\alpha = 0.1$ para aumentar la potencia del test.

A pesar de sus ventajas, esta prueba analítica presenta baja potencia estadística cuando se la aplica a un número de estudios experimentales pequeño (como suele suceder en Ingeniería de Software, donde rara vez se pueda superar los 10 experimento). Por eso se considera, como dice [73], que, a pesar de los problemas de interpretación que presenta la técnica Funel Plot [21], hoy día consiste en la alternativa más conveniente para el análisis de la heterogeneidad en Ingeniería de Software.

Si llegara a detectarse que existe heterogeneidad entre los experimentos, existen básicamente dos caminos a seguir, el primero de ellos es intentar combinar los resultados mediante un método de modelo de efectos aleatorios, y el segundo es intentar determinar cuál es el método que genera la heterogeneidad (básicamente mediante la observación del diagrama de árboles) y quitarlo del grupo, para luego re-agregar los experimentos.

B. Análisis de Heterogeneidad en contextos experimentales poco maduros. Limitaciones del estimador Q

Está bien documentada en la literatura la baja potencia de Q cuando el número de experimentos incluidos en el Meta-Análisis es bajo [6]. No obstante, no es este el problema más importante para la Ingeniería de Software. Desde la perspectiva de la Ingeniería de Software, el mayor problema reside en la incapacidad de Q para determinar la heterogeneidad de los experimentos realizados con pocos sujetos experimentales [50]. Esto se debe a que los experimentos pequeños, en general, están asociados a una alta varianza, por su mayor nivel de incertidumbre, lo cual incremente el tamaño del intervalo de confianza (IC). Este hecho hace que para los métodos gráficos los resultados se vean solapados y en lo que respecta al método Q es un fuerte atenuador debido a que cada experimento es ponderado por la inversa de su varianza (Wi = 1/vi). Dificultando de esta forma la posibilidad de detectar diferencia significativa en el test de Q (para $\alpha = 0.05$ o $\alpha = 0.1$).

Para ver esto con más claridad vamos a recurrir a un análisis gráfico, donde los ICs amplios, asociados a los experimentos pequeños, funcionan como una máscara que encubre las diferencias entre los resultados. Esto puede verse claramente en el siguiente ejemplo: supongamos un hipotético caso en el cual se cuenta con 4 experimentos que son agregados mediante el método Diferencia de Medias Ponderadas (DMP) [55] y entre los cuales existe heterogeneidad, dos de ellos dan como resultado un efecto medio (d = 0.5) y los otros dos dan como resultado un efecto muy alto (d = 1). Si estos experimentos son construidos con cien sujetos experimentales se obtiene el resultado que se indica en la Figura 6. Donde el solapamiento de los ICs de los experimentos es nulo y el p asociado al Q (12.626) es 0.00552, por tanto existe claramente heterogeneidad entre los resultados.

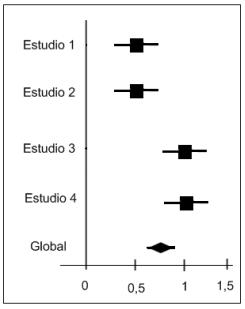

Figura 6: Diagrama de árbol que resultan de agregar los experimentos: E1, E2 con: media 1 = 100, media 2 = 90, desvío std 1 = 10, desvío std 2 = 9, y E3, E4 con: media 1 = 100, media 2 = 95, desvío std 1 = 10, desvío std 2 = 9, incluyendo 100 sujetos experimentales por brazo por cada experimento.

Ahora bien, si en lugar de cien sujetos los experimentos hubieran sido realizados con veinticinco sujetos, los resultados serían los que se indican en la Figura 7.

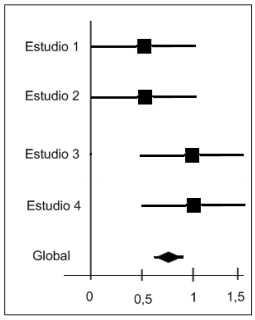

Figura 7: Diagrama de árbol que resultan de agregar los experimentos: E1, E2 con: media 1 = 100, media 2 = 90, desvío std 1 = 10, desvío std 2 = 9, y E3, E4 con: media 1 = 100, media 2 = 95, desvío std 1 = 10, desvío std 2 = 9, incluyendo 25 sujetos experimentales por brazo por cada experimento.

Puede observarse claramente que los ICs son mucho mayores que los del caso anterior, y el solapamiento entre los

mismos es más que evidente. El p asociado al Q (3.087) es 0.378372, lo que indica la no existencia de heterogeneidad. Nótese que la heterogeneidad entre experimentos no ha desaparecido al reducir el número de sujetos (E1 y E2 poseen un efecto de 0.5 y E3 y E4 un efecto de 1). Lo que ha ocurrido es que Q ha perdido su capacidad de detectar dicha heterogeneidad debido al bajo número (relativamente hablando) de sujetos experimentales involucrados.

Es importante destacar que este problema, en general, no puede ser solucionado con la incorporación de más experimentos como podría pensarse. Por ejemplo, supongamos que en lugar de 4 experimentos se contara con 40 (20 de efecto 0.5 y 20 de efecto 1). El resultado del meta-análisis sería el indicado en la Figura 8, donde el solapamiento de ICs se mantiene y el valor Q (30.872) es inferior a la cantidad de experimentos menos 1 (39). Por tanto, Q no arroja evidencia alguna de heterogeneidad.

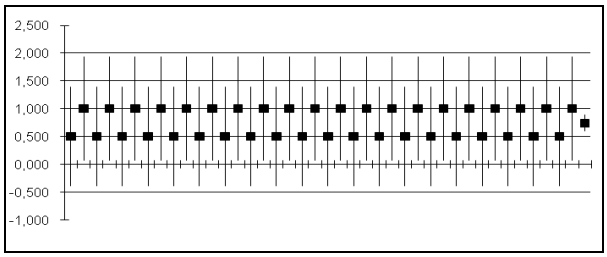

Figura 8: Diagrama de árbol que resultan de agregar los experimentos: E1 a E20 con : media 1 = 100, media 2 = 90, desvío std 1 = 10, desvío std 2 = 9, y E21 a E40 con: media 1 = 100, media 2 = 95, desvío std 1 = 10, desvío std 2 = 9, incluyendo 25 sujetos experimentales por brazo por cada experimento.

Diversos investigadores, tales como [78, 64, 110] han estudiado la potencia del estimador Q. Sin embargo, dichos estudios parten de posiciones muy alejadas a la realidad de la experimentación actual en IS, sobre todo a lo referente al elevado número de sujetos experimentales. Existen, no obstante, dos estudios que se aproximan bastante a la Ingeniería de Software, [72, 50].

En [72] se analiza la potencia de Q mediante una simulación de Monte Carlo variando los siguientes parámetros: cantidad de experimentos, cantidad de sujetos experimentales por experimento y diferencias en el tamaño de efecto. Configurando los mismos de la siguiente forma: la cantidad de experimentos a incluir en los meta-análisis toma los siguientes valores: 5, 10 y 30 experimentos, la cantidad de sujetos por experimentos toma los siguientes valores: 10, 30 y 300 sujetos, la diferencia entre el tamaño de efecto del tratamiento experimental y de control se establece mediante la siguiente relación de uno respecto del otro: 20%, 40% y 60% (por ejemplo, para un tamaño de efecto del tratamiento experimental de 1, el efecto del tratamiento de control se fija en 1.2 o 0.8 en el primer caso, 1.4 o 0.6 en el segundo caso y 1.6 o 0.4 en el tercer caso). Como resultado de este proceso los autores concluyen que el método Q tiene alta potencia (cercana al 100%) cuando los estudios tienen 300 sujetos, independientemente de la cantidad de experimentos que se agreguen o la diferencia de efecto, pero también consideran inaceptable la potencia mostrada cuando los experimentos contienen 10 o 30 sujetos.

Por su parte [50], también analizan la potencia de Q mediante una simulación de Monte Carlo, pero, a diferencia del trabajo anterior, no aplicando directamente la función de Q, sino utilizando la función de potencia estadística de Q. Los parámetros que varían en este trabajo son: la cantidad de experimentos a incluir en los meta-análisis fijado en: 5, 10 y 20 experimentos, y la varianza entre estudios fijada (la cual define la diferencia de efecto entre los mismos) en: 5, 10 y 20. Como resultado de esta simulación los autores concluyen que el método Q posee baja potencia, pero si el peso del estudio heterogéneo es alto, la capacidad del método puede verse beneficiada.

C. Meta-Análisis en Ingeniería de Software Empírica

C.1. Reseña de la utilización del Meta-Análisis en Ingeniería de Software Empírica

La aplicación del Meta-Análisis en Ingeniería de Software es bastante tardía, en consonancia con la también tardía utilización de la experimentación como herramienta metodológica de investigación. El primero en señalar su posible uso en Ingeniería de Software fue [83], que empleó DMP para combinar 5 experimentos que exploraban distintas técnicas de prueba de software. Pero, el Meta-Análisis tardó bastantes años en calar en la comunidad de Ingeniería de Software, en buena parte debido a la deprimente presentación que Miller realizó pues concluyó que el Meta-Análisis (sólo aplicó DMP pero generalizó erróneamente el resultado de su estudio, sin considerar que existen otras técnicas de meta-análisis) no era aplicable porque no se cumplían las condiciones formales que exige. En consecuencia, la síntesis estadística de resultados experimentales quedaba descartada para la Ingeniería de Software Empírica. Las consecuencias de esta conclusión fueron demoledoras, y durante muchos años en los artículos de síntesis [24, 69, 65] hubo una total ausencia de aplicación de Meta-Análisis. En su lugar se realizaban combinaciones de los resultados experimentales mediante técnicas, que podríamos denominar, narrativas por su total ausencia de formalidad.

En 2004, Kitchenham publica su bien conocido reporte acerca de revisión sistemática de experimentos [73]. En la actualidad, se estima que se han publicado unas 100 revisiones sistemáticas en Ingeniería de Software [20]. Sería esperable, por lo tanto, que el número de meta-análisis hubiera aumentado en concordancia (nótese que revisión sistemática y Meta-Análisis son conceptos relacionados, pero distintos. Una revisión sistemática es una revisión sistemática comprende todo el proceso de revisión, mientras que el Meta-Análisis se circunscribe al proceso de síntesis [73]). Sin embargo, la realidad muestra que, a pesar del gran número de revisiones sistemáticas realizadas, sólo se ha aplicado meta-análisis en el 2% de casos. Pero aún, de las citadas 100 revisiones sólo se aplicó alguna técnica de síntesis (estadística o no) en unos cinco casos como máximo [20], tratándose el 95% de los casos restantes de lo que recientemente se ha dado en denominar mapping studies, para diferenciarlos de las revisiones sistemáticas puras donde la agregación de resultados es una parte consustancial.

Si bien el trabajo de [84] fue pionero en cuanto a la aplicación de un Meta-Análisis en Ingeniería del Software, solo combino los resultados de cuatro experimentos, por cuanto la aplicación real de los conocimientos generados fue escaza. Más recientemente, en el trabajo de [31], se identifican 20 experimentos sobre programación de a pares, los cuales son agregados en tres grupos, el primero de 11 experimentos, el segundo de 11 experimentos y el tercero de 10 experimentos, donde el experimento más pequeño contiene 4 sujetos por tratamiento, el mayor 35 sujetos por tratamiento, mientras que el promedio asciende a 13 sujetos por tratamiento.

Así como se menciona el trabajo de [31], hay una gran cantidad de autores que, si bien pudieron desarrollas las tareas

de búsqueda y selección de estudios experimentales, no pudieron agregar los resultados para generar una conclusión basada en un mayor nivel de evidencia utilizando el método DMP. Por ejemplo: en [27] se analizaron un conjunto de experimentos vinculados a las técnicas de educción de requisitos y, debido a que la gran mayoría de los reportes no publicaba las varianzas, se generan un conjunto de recomendaciones respecto del uso de las mismas mediante un conteo de votos (el tratamiento que tenía mayor cantidad de estudios que indicaban que era mejor era proclamado como el más adecuado), en [86] se analizó la reutilización del software en la modificación y/o creación de nuevos productos, como había problemas de compatibilidad en las variables respuesta analizadas en cada estudios y falencias en los reportes, se recurrió a un conteo de votos como estrategia para generar las conclusiones.

C.2. Meta-Análisis realizados en Ingeniería de Software Empírica

Hasta el presente se han realizado 2 Meta-Análisis en IS, los cuales ya han sido mencionados, pero a continuación se los describe con mayor detalle:

- [31], en este trabajo se identifican 20 experimentos sobre programación de a pares, los cuales son agregados en tres grupos, el primero de 11 experimentos, el segundo de 11 experimentos y el tercero de 10 experimentos, donde el experimento más pequeño contiene 4 sujetos por tratamiento, el mayor 35 sujetos por tratamiento, mientras que el promedio asciende a 13 sujetos por tratamiento. Los resultados son agregados mediante DMP en sus dos versiones modelo de efecto fijo y aleatorio.
- [13], en este trabajo se identifican 21 experimentos sobre técnicas de inspección, los cuales son agregados en tres grupos, el primero conteniendo 7 experimentos, el segundo 9 y el tercero 5, dichos experimentos poseen tamaños variados, conteniendo el menor 3 sujetos experimentales por tratamiento, el mayor 45 sujetos por tratamiento, mientras que el promedio asciende a 6 sujetos por tratamiento. Los resultados son agregados mediante una variante del método DMP para modelo de efecto fijo.

III. Descripción Del Problema

A. La agregación estadística de experimentos en Ingeniería de Software como solución a un contexto experimental poco maduro

La Ingeniería en Software, de acuerdo a la norma 610.12 de la IEEE, debe aplicar conocimiento científico para el desarrollo, operación y mantenimiento de los sistemas software. Esto implica poder determinar dentro de una serie de métodos, técnicas y herramientas cual se debe utilizar en cada actividad de acuerdo a las condiciones del proyecto [67]. Para ello, es necesario contar con un marco que permita a los ingenieros de software poder conocer como es el comportamiento de los métodos y herramientas mediante un enfoque científico y por lo tanto objetivo. El marco experimental a fin a todas las ingenierías, permite brindar información objetiva sobre que conviene aplicar en cada etapa de un proyecto software según las circunstancias. Entonces, como afirma [92], de esta forma se "permita ganar más entendimiento de cómo hacer un software bueno y cómo hacerlo mejor".

Varios autores [24, 84, 15, 32] han señalado que la Ingeniería de Software presenta un contexto experimental poco maduro, donde existe escasez de experimentos, y la cantidad de sujetos experimentales es también pobre [24]. En este contexto, la mejor opción que se presenta es la de combinar el resultado de diversos experimentos por medio de un proceso de agregación como propone [15].

En el capítulo anterior se exploraron las técnicas de agregación más utilizadas y mejor documentadas. Hay estudios [37] que establecieron que la Diferencia de Medias Ponderadas (DMP) se adapta mejor al contexto poco madura que presenta en la actualidad la Ingeniería de Software. No obstante ello, queda otra cuestión que zanjar: esta técnica de agregación estadística puede ser utilizada en dos modelos de Meta-Análisis bien diferentes entre sí: (a) el modelo de efecto fijo supone la existencia de un único resultado poblacional, el cual se irá estabilizando a medida que se incorporen experimentos al meta-análisis, y (b) el modelo de efecto aleatorio supone que existe un conjunto de variables no controladas que influyen en los resultados de los experimentos provocando que los resultados cambien a medida que se incorporan experimentos de distintas vertientes, ambos analizados.

Entonces se presenta el dilema de cuál es el modelo de efecto de Meta-Análisis que mejor se adapta a las necesidades de la experimentación en el campo de la Ingeniería de Software.

B. Los modelos de efecto de Meta-Análisis en el actual contexto experimental de la Ingeniería de Software

El Meta-Análisis es hoy día una práctica conocida dentro de la Ingeniería de Software Empírica. Desde que fue propuesto por [5] hubo varios trabajos abordados desde esta óptica [3, 105, 59, 116, 69, 65, 73, 31], sin embargo no existe un criterio unificado de qué modelo de Meta-Análisis se debe aplicar hoy día en Ingeniería de Software.

Si se remite a la teoría estadística, esto debería hacerse en base al análisis de la heterogeneidad entre los experimentos, lo cual no es factible realizar en la práctica en el actual contexto de la Ingeniería de Software debido a que las muestras de los experimentos son pequeñas y los Meta-Análisis agregan pocos experimentos. En este contexto, como se indica en [50, 72], la heterogeneidad no es medible ya que los test no tienen potencia. Esto hace que autores como Tore Dyba aplique en sus trabajos [31] ambos métodos de Meta-Análisis, o que [13] se pase por alto el tema y se agreguen los experimentos solo con el modelo de efecto fijo.

Fuera del contexto de la Ingeniería de Software, a nivel general [103] proponen utilizar el modelo de efectos aleatorio en todos los casos, ya que por las limitaciones de las técnicas de heterogeneidad no es factible determinar si existe o no la misma, y según el modelo teórico que ellos abalan, si un grupo de experimentos sin heterogeneidad se agrega mediante el modelo de efectos aleatorios, el resultado será similar al del modelo de efecto fijo.

Por otra parte, se define en [6] que para el Meta-Análisis donde el conjunto de experimentos es menor a 10 el error de la varianza entre experimentos es demasiado alto y la única alternativa es el modelo de efecto fijo. Es decir, no hay desde la teoría consenso respecto sobre qué modelo utilizar, como así tampoco, evidencia práctica sobre cuál de las posturas es la correcta dentro del ámbito de la Ingeniería de Software.

C. Elección del modelo de efecto de Meta-Análisis: alcance y límites del problema

Cuando se intenta determinar de manera experimental si una metodología o técnica nueva presenta una mejor performance que una ya existente, se plantea un test de hipótesis para establecer que tratamiento es superior. En el

actual contexto de la Ingeniería de Software Empírica, poder realizar una prueba de hipótesis con el número adecuado de sujetos experimentales [24] es sumamente difícil, por las características intrínsecas de la misma experimentación. Por ello se planteó [4] como solución la utilización del Meta-Análisis. Sin embargo, existen dos modelos bien definidos y debe determinarse cuál es el más adecuado. Esto constituiría una herramienta para que el investigador pudiera determinar de manera aproximada a la realidad (errores de Tipo I y II bajos [17], establecidos en 0,05 y 0,2) si un tratamiento experimental es superior al tratamiento de control, y cuál es la medida de esta mejoría (tamaño de efecto).

Para determinar la utilidad e importancia del desarrollo de una herramienta de este tipo debería plantearse la siguiente pregunta: ¿cuál es el riesgo de no poder o saber escoger entre un modelo u otro? El riesgo principal consiste en terminar eligiendo un modelo de Meta-Análisis no adecuado, y que haga que el resultado obtenido de la agregación no tenga suficiente potencia o carezca de la fiabilidad adecuada. O incluso, que no se sepa que potencia y fiabilidad tiene el método de agregación seleccionado con el modelo de Meta-Análisis utilizado.

No tener la fiabilidad deseada, implica caer en un error de Tipo I, y la falta de potencia lleva a caer en un error de Tipo II. En el primer caso, se estará afirmando que el tratamiento experimental supera al de control, cuando en realidad no lo hace, llegando a un resultado equivocado. En el segundo caso, se tiene un tratamiento experimental que supera al de control (que es lo que se quiere determinar) pero la experimentación no permite aseverarlo. Por último, si no se conoce la fiabilidad o potencia del modelo usado, entonces se puede cometer un error (de Tipo I y/o de Tipo II), y no se conocería esta circunstancia, en cuyo caso, las consecuencias serían las de tener un resultado con un grado de certeza desconocida (incertidumbre del resultado obtenido).

El problema que se presenta entonces, es determinar qué modelo de efecto de Meta-Análisis es más adecuado para aplicar la técnica de Diferencia de Medias Ponderadas para agregar estadísticamente los resultados de investigaciones experimentales en el campo de la Ingeniería e Software. Como se mencionara anteriormente, y al no existir un consenso sobre qué modelo de Meta-Análisis utilizar, se deberá recurrir a la experiencia práctica, para lo que deberán probarse ambos modelos y analizar su comportamiento. Sin embargo, por tratarse de un contexto experimental poco maduro, no es posible contar con los experimentos necesarios para realizar el estudio, por lo que se recurrirá a realizar un proceso de simulación que permita validar el modelo adecuado a utilizar en el actual contexto de la Ingeniería de Software Empírica.

Lo que se propone, entonces, es determinar cuál es el modelo de Meta-Análisis más adecuado para utilizar en una agregación de experimentos por el método de Diferencia de Medias Ponderadas. Para ello, se van a realizar agregaciones con experimentos de igual tamaño y con experimentos de tamaños distintos (que es un contexto más representativo de la realidad experimental en la Ingeniería de Software Empírica). Esto lleva a considerar la existencia de dos escenarios bien definidos: (a) la agregación de experimentos realizados con distinta cantidad de sujetos (experimentos de distinto tamaño), y (b) la agregación de experimentos realizados con la misma cantidad de sujetos (experimentos de igual tamaño).

D. Resumen de investigación

Al realizar una agregación de experimentos, hay un número de parámetros que influyen directamente en las características de los mismos, por lo que impactan en los resultados obtenidos. Por ello, debe tenerse en cuenta que determinar cuál es el modelo de Meta-Análisis que mejor se adapta al actual contexto de la Ingeniería de Software Empírica o se limita a señalar uno solo de los modelos en forma determinante, sino que se deberá determinar qué modelo es más adecuado de acuerdo a las características que presenta la agregación (dependiendo, es decir, de los conjuntos de valores que tomen los diversos parámetros).

También se mencionó que hay dos escenarios bien definidos en la agregación de experimentos: la combinación de experimentos de igual tamaño (todos con la misma cantidad de sujetos experimentales) y la combinación de experimentos de distinto tamaño. Éste último escenario es más representativo de la realidad experimental de la Ingeniería de Software Empírica, pero no puede elegirse a priori sin determinar la fiabilidad y potencia que presenta cada modelo, o si alguno tiene un error (sea de Tipo I o de Tipo II) no aceptable (Error Tipo I mayor a 0,05 y/o Error Tipo II mayor a 0,2 [17]).

Finalmente, se pueden formular los siguientes objetivos y preguntas de investigación:

Objetivo 1: determinar cuál es el modelo de efecto que mejor se adapta al Meta-Análisis en Ingeniería de Software Empírica cuando los experimentos a agregar tienen el mismo tamaño.

Pregunta de investigación: ¿Es posible establecer un conjunto de recomendaciones para escoger el modelo de Meta-Análisis más adecuado en la agregación de experimentos en Ingeniería de Software de igual tamaño con la utilización del método de Diferencia de Medias Ponderadas?

Objetivo 2: determinar cuál es el modelo de efecto que mejor se adapta al Meta-Análisis en Ingeniería de Software Empírica cuando los experimentos a agregar tienen el distinto tamaño.

Pregunta investigación: ¿Es posible establecer un conjunto de recomendaciones para escoger el modelo de Meta-Análisis más adecuado en la agregación de experimentos en Ingeniería de Software de distinto tamaño con la utilización del método de Diferencia de Medias Ponderadas?

IV. Materiales y Métodos

A. Propuesta de simulación como medio para establecer el modelo de efectos de Meta-Análisis a aplicar en Ingeniería de Software

El modelo de efecto en el cual aplicar el método de Diferencia de Medias Ponderadas [37] no está determinado aún, y hay posturas encontradas al respecto [13, 103].

La estrategia para decidir si se debe utilizar un modelo de efecto fijo o un modelo de efecto aleatorio, se basa en determinar si hay heterogeneidad o no entre los experimentos, sin embargo hay estudios que demuestran que los métodos para análisis de heterogeneidad no tienen potencia con pocos experimentos [6, 50, 72]. Por ello, es imposible determinar metodológicamente cual sería el modelo a utilizar, surgiendo la necesidad de obtener una solución de forma empírica.

Para resolver el problema pragmáticamente, bastaría con agregar los resultados de diferentes experimentos en cada uno de los escenarios propuestos, y luego analizar cuál se adecúa más a las características que presenta la Ingeniería de Software Empírica.

La estrategia sería la de obtener los resultados de experimentos de características similares. Todos ellos contarían con un número de sujetos bajo, ya que esa es la característica principal en este campo [24] y que genera la necesidad de tener

que agregar el resultado de varios experimentos. Luego se agregaría el resultado de estos experimentos con el método de Diferencia de Medias Ponderadas [55] aplicada primero con el modelo de efecto fijo, y posteriormente con el modelo de efectos aleatorios. Con las combinaciones así realizadas, se debería analizar y comparar la fiabilidad y potencia estadística de ambos modelos de efectos. De esta manera, se podrá determinar el modelo de efectos que mejor se adapta al propósito señalado.

Por simulación de Monte Carlo se obtendrá la cantidad de estudios experimentales necesarios para poder llegar a una conclusión apoyada por una evidencia empírica adecuada (potencia y fiabilidad estadística de 80% y 95% respectivamente [17]).

B. Simulación por el Método de Monte Carlo

Como herramienta de investigación, la simulación de Monte Carlo es un procedimiento sólidamente establecido que proporciona soluciones aproximadas a una gran variedad de problemas.

El método de Monte Carlo (o simulación de Monte Carlo, como también se denomina) [81] es un tipo de algoritmo probabilístico que permite encontrar soluciones a problemas que no poseen una formulación explícita pero pueden plantearse en términos de experimentos aleatorios. Un ejemplo bien conocido es el cálculo de π mediante la Buffon's Leedle [7].

Un uso bastante corriente de las simulaciones de Monte Carlo es comprobar el comportamiento de estimadores estadísticos (en este caso, los distintos modelos de meta-análisis) en situaciones no asintóticas [101], como la ausencia de normalidad o pequeñas muestras, que es el caso que aborda esta investigación. De hecho, la simulación de Monte Carlo ha sido utilizada en todos las investigaciones similares a la presente [51, 77, 39, 110, 78, 64] realizadas hasta la fecha.

Para una técnica de meta-análisis, la simulación de Monte Carlo se realizaría del modo siguiente:

Paso 1. Se definen los parámetros de la(s) población(es) sobre las que se desea probar la con exactitud del estimador. Entre estos parámetros se encuentra el tipo de distribución de probabilidad (por regla general se escoge la distribución normal), así como el tamaño de efecto poblacional δ. Otros parámetros poblacionales (medias y varianzas) pueden definirse si la simulación así lo exige.

Paso 2. Se extraen muestras de dicha(s) población(es), utilizando para ello una tabla o generador de números aleatorios. El número de muestras extraídas depende de los parámetros de la simulación.

Paso 3. Se calculan los valores del estimador estadístico correspondiente (por ejemplo: tamaño de efecto global d* calculado mediante el modelo de efectos fijos, intervalos de confianza de d*). Nótese que estos valores son calculados utilizando las fórmulas asintóticas (basadas en la Ley de Grandes Números) del estimador bajo estudio.

Paso 4. Se comparan los valores del estimador con los valores poblacionales. Por ejemplo, para obtener la exactitud del estimador, habría que comprobar si el intervalo de confianza de d* contiene el tamaño de efecto poblacional δ. En caso afirmativo, se incrementaría el valor de una variable (número de aciertos). Para la potencia empírica se procedería de modo análogo.

Paso 5. Se repiten los pasos 2-4 un número de veces que asegure la convergencia al resultado, ya que la precisión de una simulación de Monte Carlo es directamente proporcional al número de veces que se ejecuta [7].

C. Simulador

La herramienta esencial para la realización del proceso de simulación, es un software de simulación. Se procederá, entonces, al diseño y desarrollo de un simulador para los propósitos de la presente investigación.

Para diseñar y desarrollar el simulador habrá que tener en cuenta las herramientas de desarrollo disponibles y el hardware disponible. En este caso, se utiliza VB.NET su generador de números aleatorios. Como el generador de VB.NET genera números aleatorios de distribución uniforme, el conjunto de números generados debe convertirse a otro conjunto de números de distribución normal, fijando la media y desvío estándar deseado y utilizando una función de conversión (fórmula 10)

$$X = \mu + \sigma * c$$

$$c = \sqrt{-2 * In(r_1)} * \cos(2\pi r_2)$$

Fórmula 10: Función de conversión de valores de distribución uniforme a valores de distribución normal.

Donde:
μ es la media
σ es el desvío estándar
r_1 y r_2 son números aleatorios uniformemente distribuidos

Estos puntos, así como el diseño y desarrollo del simulador escapan a los propósitos de esta investigación, y su complejidad se corresponde con un desarrollo de grado de nivel medio, por lo que los detalles del mismo se obvian en la Investigación.

V. EXPERIMENTACIÓN

A. Introducción

El objetivo que se persigue con la experimentación es determinar a través de simulaciones cuales son las condiciones bajo las cuales es conveniente agregar experimentos bajo el modelo de efecto fijo y cuando es mejor utilizar el modelo de efectos aleatorios, basándose siempre en el método de agregación Diferencia de Medias Ponderadas [37].

Utilizando los modelos de efecto fijo y efectos aleatorios, y con la ayuda de un simulador que se desarrollará con la finalidad de realizar las experimentaciones (sección V.D) éstas se llevaron a cabo, y a partir del análisis de los resultados, se ha podido desarrollar una serie de recomendaciones para el investigador a la hora de utilizar el Meta-Análisis como herramienta de estudio.

B. Diseño experimental

Se realizan simulaciones bajo distintas condiciones para determinar qué modelo se comporta con mejor fiabilidad y potencia estadística en cada circunstancia. Las condiciones consideradas son cada una de las combinaciones posibles de los valores que toman las distintas variables independientes (sección V.C.1) del experimento.

Las simulaciones consisten en recrear las condiciones experimentales que pueden darse en una investigación de

Ingeniería de Software Empírica. A través de un tratamiento de control, busca determinarse si un tratamiento experimental tiene mejor performance. Para ello debe establecerse el parámetro a medir en cada experimento y hacer una comparación de las medias obtenidas entre los tratamiento de control y experimental.

Se realiza la simulación de varios experimentos, los cuales se agregan con la técnica Diferencia de Medias Ponderadas, bajo dos supuestos:

1) Existe un único tamaño de efecto poblacional, y las diferencias observadas son propias del error asociado directamente a la experimentación. En este caso se utilizará el modelo de Meta-Análisis de efecto fijo.
2) No existe un único tamaño de efecto poblacional, y las diferencias se deben a circunstancias ajenas a los errores experimentales. En este caso se utilizará el modelo de Meta-Análisis de efectos aleatorios.

En cada corrida, se generan números aleatorios que representan las medias poblacionales de un tratamiento y otro (tratamiento de control y experimental). Para ello, se realiza la simulación suponiendo que el tratamiento experimental es mejor que el tratamiento de control en una medida igual al tamaño de efecto teórico a analizar.

Posteriormente se calculan los tamaños de efecto según el caso que se trate (modelo de efecto fijo o aleatorio), y luego se observa la fiabilidad y potencia estadística que se obtiene al aplicar cada modelo, para poder saber en qué condiciones es aconsejable utilizar uno u otro modelo.

C. Variables de estudio

C.1. Variables independientes

Se considera el siguiente conjunto de variables independientes:

- Cantidad de sujetos por experimentos.
- Cantidad de experimentos a agregar/combinar.
- Media poblacional del tratamiento de control.
- Desvío estándar teórico.
- Tamaño de efecto poblacional teórico.

A continuación se detallará cada una de estas variables junto a los valores que les fueron asignados durante las simulaciones.

C.1.1. Cantidad de sujetos experimentales

La cantidad de sujetos experimentales es que la cantidad de sujetos que intervienen en cada experimento.

Para conocer los valores límites de esta variable, se ha tenido en cuenta que ese número, es muy bajo [24]. Como se señala en [36] el concepto de pequeña muestra es bastante impreciso. Hoyle proporciona la cifra general de 150 sujetos como representativa de las pequeñas muestras [58], si bien existen estudios que rebajan esta cifra a 50 [47, 94]. En meta-análisis las cifras manejadas son incluso menores, en el orden de los 10-20 sujetos por estudio [52].

Debido a que existe una falta de acuerdo, se ha decidido tomar una postura intermedia, y estudiar experimentos que posean unos 40 sujetos totales (20 para el grupo del tratamiento de control y 20 para el grupo del tratamiento experimental). Con valores mayores de 20 por grupo, se estaría muy cerca ya de los 30 sujetos que habitualmente se consideran suficientes para asumir normalidad [41]. Por otro lado, y como se indica en [36] la mayoría de estudios en IS están por debajo de los valores anteriores, a modo de ejemplo [107] reporta que la mediana de la distribución del número de sujetos por experimentos es de 30, estando la media muy por debajo de este valor. Así mismo, los Meta-Análisis realizados hasta el momento en IS muestran los siguientes valores: en [31] el promedio de sujetos por experimento asciende a 13 sujetos por brazo y en [13] el promedio asciende a solo 6 sujetos por grupo.

Por tanto, la decisión tomada con respecto al tamaño de experimentos (cantidad de sujetos experimentales) es de tomar como cota mínima 4 y cota máxima 20. Asimismo, se hará referencia a estas cantidades como baja o pocos (4-9), media (10-12) y alta o muchos (13-20).

C.1.2. Cantidad de experimentos a combinar

El contexto experimental de la Ingeniería de Software no aporta hoy día muchos experimentos potencialmente combinables en un proceso de agregación, debido a diversos motivos, a saber: escasez de experimentos, replicaciones y homogeneidad entre los mismos [24, 84], carencia de estándares para reportes de experimentos. Por ejemplo, [9] no publican varianzas y [11] ni siquiera reporta las medias de los resultados experimentales, y la falta de estandarización de las variables respuesta, por ejemplo, los trabajos de [1, 117] utilizan diferentes variables respuesta para analizar un mismo aspecto, lo cual hace que estos experimentos no puedan ser agregados. Esta limitación hace que la cantidad de experimentos que simulemos para agregar no sea elevada. Algunos autores coinciden en señalar [6] que si el Meta-Análisis posee menos de 10 experimentos los riesgos de caer en un error son altos, al punto que no se recomienda utilizar el modelo de efectos aleatorios si la cantidad de experimentos es inferior a 10. Los Meta-Análisis realizados hasta el momento en IS muestran los siguientes valores: en [31] se realizaron dos agregaciones de 11 experimentos y una de 10 y en [13] se realizaron 3 agregación con 5, 7 y 9 experimentos. Por ello, la cantidad de experimentos a agregar en cada meta-análisis oscilará entre 2 y 10, incrementándose de dos en dos. Se hará referencia a estos valores como bajo o pocos (2 y 4), medio (6) y alto o muchos (8 y 10).

Sin embargo, puede suceder que haya que combinar experimentos de distintos tamaños por no poder disponer del ideal recién expresado. En este caso, estaríamos combinando el resultado de experimentos con distinta cantidad de sujetos experimentales. Con este punto planteado, se consideró que era necesario realizar simulaciones que agreguen experimentos que difieran en cantidad de sujetos marcadamente, por lo que se descartó combinar experimentos con una cantidad de sujetos media, y se tomaron como tamaño de experimentos los valores 4, 14 y 20. Para este caso, se tomarán las combinaciones de sujetos y experimentos que se muestran en la Tabla IV.

De esta manera, se realizarán cuatro tipos de simulaciones:

- Simulaciones con modelo de efecto fijo e igual tamaños de experimentos a agregar (todos con la técnica DMP)
- Simulaciones con modelo de efecto aleatorio e igual tamaños de experimentos a agregar (todos con la técnica DMP)
- Simulaciones con modelo de efecto fijo y distintos tamaños de experimentos a agregar (todos con la técnica DMP)
- Simulaciones con modelo de efecto aleatorio y distintos tamaños de experimentos a agregar (todos con la técnica DMP)

C.1.3. Media poblacional del tratamiento de control

Es la media estadística del valor medido en cada experimento aplicado al tratamiento de control. La media poblacional del tratamiento de control (μ^c) es fijada en 100 a efectos de cálculo. Este valor es tomado de manera arbitraria,

sin necesidad de seguir ningún lineamiento. Por ello, se elige este número (100), ya que se considera que facilitará los cálculos, la lectura de resultados y el posterior análisis comparativo.

TABLA IV. COMBINACIÓN DE SUJETOS Y EXPERIMENTOS A AGREGAR

Sujetos por experimentos	Experimentos a agregar
4-4-14	3
4-4-20	3
4-14-14	3
4-14-20	3
4-20-20	3
14-14-20	3
14-20-20	3
4-4-4-4-14-14	6
4-4-4-4-20-20	6
4-4-14-14-14-14	6
4-4-14-14-20-20	6
4-4-20-20-20-20	6
14-14-14-14-20-20	6
14-14-20-20-20-20	6
4-4-4-4-4-4-14-14-14	9
4-4-4-4-4-4-20-20-20	9
4-4-4-14-14-14-14-14-14	9
4-4-4-14-14-14-20-20-20	9
4-4-4-20-20-20-20-20-20	9
14-14-14-14-14-14-20-20-20	9
14-14-14-20-20-20-20-20-20	9

C.1.4. Desvío estándar teórico

Es la desviación estándar estadística de la media poblacional. Se supone igual para ambos tratamientos. Posteriormente, se calculará el desvío estándar real que es el que se observa en los experimentos simulados.

De acuerdo al actual contexto experimental de la Ingeniería de Software, los valores que pueden alcanzar el desvío estándar en distintos experimentos [39] son del 10% (desvío estándar bajo), del 40% (desvío estándar medio) y del 70% (desvío estándar alto) de la media poblacional.

C.1.5. Tamaño de efecto poblacional teórico

El tamaño de efecto indica la mejoría de un tratamiento experimental sobre uno de control. El efecto teórico, se refiere a la suposición que se realiza sobre la medida en que el tratamiento experimental supera al de control.

Los tamaño de efecto poblacional (d) a analizar, de acuerdo al actual contexto experimental de la Ingeniería de Software [17, 71], son: bajo (0,2), medio (0,5), alto (0,8) y muy alto (1,2).

C.2. Variables dependientes e intermedias

Se utiliza el término de variables intermedias, a aquellas que no son independientes, ni son las variables de interés directo en el estudio, pero son necesarias para calcular las variables dependientes. Se han tenido en cuenta el siguiente conjunto de variables:

- Los tamaños de efecto poblacional medido o real, desvío estándar medido e intervalo de confianza (intermedias)
- La media poblacional del tratamiento experimental (intermedia)
- Potencia y fiabilidad deseadas (errores de tipo I y II, variables dependientes).

C.2.1. Tamaño de efecto poblacional medido o real, desvío estándar e intervalo de confianza

El tamaño de efecto indica la mejoría de un tratamiento experimental sobre uno de control, y se calcula de acuerdo al método de agregación que se esté utilizando. En este caso, se usará la fórmula de la Diferencia de Medias Ponderadas, la cual varía teniendo en cuenta que se trate del modelo de efecto fijo o el modelo de efecto aleatorio. El desvío estándar medido es la desviación estándar estadística de la media poblacional, observado en la simulación de los experimentos

Para el modelo de efecto fijo, estas variables se calculan con ecuaciones vistas en el capítulo dos y que aquí se repiten:

$$d = J\frac{Y^E - Y^C}{S_P} \qquad J = 1 - \frac{3}{4N-9}$$

d representa el tamaño de efecto
J representa el factor de corrección
Y's representa a las medias del grupo experimental (E) y de control (C)
Sp representa el desvió estándar conjunto
N representa el total de sujetos experimentales incluidos en el experimento

Fórmula 11: tamaño de efecto individual por el método de Diferencia de Medias Ponderadas para el modelo de efecto fijo

$$v = \frac{\tilde{n} + d^2}{2(n^E + n^C)} \qquad \tilde{n} = \frac{n^E + n^C}{n^E * n^C}$$

$$d - Z_{\alpha/2}\sqrt{v} \le \lambda \le d + Z_{\alpha/2}\sqrt{v}$$

v representa el error típico
d representa el tamaño de efecto
n's representa la cantidad de sujetos experimentales del grupo experimental (E) y de control (C)
Z representa la cantidad de desvíos estándar que separan, al nivel de significancia dado, la media del límite. En general es 1,96 (α = 0,05)

Fórmula 12: error típico e intervalo de confianza por el método de Diferencia de Medias Ponderadas para el modelo de efecto fijo

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

$$v = (1/\sum 1/\sigma^2{}_i(d))$$

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

Fórmula 13: tamaño de efecto grupal y desvío estándar por el método de Diferencia de Medias Ponderadas para el modelo de efecto fijo

$$d^* - Z_{\alpha/2}\sqrt{v} \le \lambda \le d^* + Z_{\alpha/2}\sqrt{v}$$

d* representa el tamaño de efecto global
Z representa la cantidad de desvíos estándar que separan, al nivel de significancia dado, la media del límite. En general es 1,96 (α = 0,05)
v representa el error típico

Fórmula 14: intervalo de confianza grupal por el método de Diferencia de Medias Ponderadas para el modelo de efecto fijo

Y para el modelo de efectos aleatorios:

$$\Delta = \frac{\sum d_i / \gamma^2{}_i}{\sum 1/\gamma^2{}_i}$$

Δ representa el tamaño de efecto global
$\sum d_i / \gamma^2{}_i$ representa la sumatoria de los efectos individuales
$\sum 1/\gamma^2{}_i$ representa la sumatoria de la inversa de las varianzas entre-estudios e intra-estudios

Fórmula 15: tamaño de efecto global por el método de Diferencia de Medias Ponderadas para el modelo de efectos aleatorios

$$\Delta - Z_{\alpha/2}\sqrt{v} \leq \Delta \leq \Delta + Z_{\alpha/2}\sqrt{v}$$

$$v = \frac{1}{\sum 1/\gamma^2{}_i}$$

Δ representa el tamaño de efecto global

Z representa la cantidad de desvíos estándar que separan, al nivel de significancia dado, la media del límite. En general es 1,96 (α = 0,05)

v representa el error típico

Fórmula 16: desvío estándar grupal e intervalo de confianza por el método de Diferencia de Medias Ponderadas para el modelo de efectos aleatorios

C.2.2. Media poblacional del tratamiento experimental

El valor de la media poblacional del tratamiento experimental deberá estimarse, y esto se hará de la siguiente forma $\mu^E = 100 + \delta * \sigma$.

C.2.3. Potencia y fiabilidad deseadas (errores de tipo I y II)

El error α, o error de tipo I, y β, o error de tipo II, fueron tratados en el capítulo dos. Aquí se explica cómo se calculan en la simulación y los valores que se tomarán como cota mínima aceptable. También se repite la tabla que muestra estos errores (Tabla V).

TABLA V. TIPOS DE ERROR DE UN TEST ESTADÍSTICO

	H0 verificada en a población	**H1 verificada en la población**
H0 respuesta del experimento	Decisión correcta (1-α)	β (Tipo II error)
H1 aceptada respuesta del experimento	α (Tipo I error)	Decisión correcta (1-β)

Los resultados vinculados a la fiabilidad indican el porcentaje de veces que el intervalo de confianza estimado contuvo el valor del tamaño de efecto poblacional, mientras que los resultados vinculados a la potencia estadística indican el porcentaje de veces que dicho intervalo de confianza no contuvo el valor 0 (cero). En cada corrida de la simulación, entonces, la forma de proceder será calcular la media del tratamiento experimental, los tamaños de efecto poblacionales, error estándar e intervalos de confianza. Luego, si el tamaño de efecto poblacional calculado queda dentro del intervalo de confianza, entonces esa corrida suma a la cantidad de veces que el intervalo de confianza contuvo al efecto (aciertos). Al final de las simulaciones, se calcula el porcentaje de aciertos, y ese es el valor de la fiabilidad. De la misma manera, se calcula la cantidad de veces que el intervalo de confianza no contuvo el valor 0, y el porcentaje calculado al final del proceso de simulación, indicará la potencia del método.

Con el objeto de determinar en qué condiciones los métodos de agregación son fiables y tienen buena potencia estadística, se fijarán los valores deseados en 95% (error de tipo I = 0,05) y 80% (error de tipo II = 0,2) respectivamente, ya que estos valores son los habitualmente recomendados [17].

C.3. Cantidad de simulaciones

La Ley de los Grandes Números [36] permite establece una cota de error a los resultados de una simulación. Esta cota depende del problema particular a resolver, pero en general su magnitud es proporcional a $\sqrt{}(1/n)$, siendo n la cantidad de muestreos realizados [81]. En otras palabras, a medida que el número de muestreos se incrementa, la precisión de la simulación aumenta en consecuencia. Con 300 muestreos se alcanza una precisión razonable [87]. En el caso de esta investigación, en cada simulación se realizarán 1000 muestreos, los cuales aseguran la validez de los resultados obtenidos. Para cada simulación todas las combinaciones de parámetros posibles, según los valores fijados, de esta manera, se tendrán las siguientes combinaciones para igual tamaño de experimento:

- Cantidad de sujetos experimentales: 4, 8, 10, 14 y 20.
- Cantidad de experimentos a agregar: 2, 4, 6, 8 y 10.
- Cantidad de efectos poblacionales distintos: 0.2, 0.4, 0.8 y 1.2.
- Cantidad de desvíos estándares a considerar: 10%, 40% y 70%.
- Modelos de efecto de Meta-Análisis: modelo de efecto fijo y modelo de efecto aleatorio.

Por tanto, se tienen 600 (5x5x4x3x2) combinaciones posibles. Como se dijo que se realizarían 1000 simulaciones por cada combinación de parámetros posibles, se tendrá un total de 600.000 simulaciones.

Por otro lado, las simulaciones a realizar con distinto tamaño de experimentos será:

- Combinación de experimentos y sujetos: 21
- Cantidad de efectos poblacionales distintos: 0.2, 0.4, 0.8 y 1.2.
- Cantidad de desvíos estándares a considerar: 10%, 40% y 70%.
- Modelos de efecto de Meta-Análisis: modelo de efecto fijo y modelo de efecto aleatorio.

Por tanto, se tienen 504 (21x4x3x2) combinaciones posibles. Como se dijo que se realizarían 1000 simulaciones por cada combinación de parámetros posibles, se tendrá un total de 504.000 simulaciones

Finalmente, tendremos un total de 504.000 + 600.000 = 1.104.000 simulaciones.

D. Resultados experimentales

Se desarrolla y prueba el simulador, y se procede a la realización de las simulaciones que arrojaron los resultados que se muestran a continuación. Para el simulador se utiliza VB.NET y su generador de números aleatorios. El desarrollo del simulador corresponde a un nivel de grado medio, por lo que su detalle se obvia en la presente Investigación.

D.1. Convención adoptada para lectura de las tablas que muestran los resultados de los experimentos de igual tamaño

Para analizar los resultados de las simulaciones y poder visualizarlos de manera rápida y clara, se adopta para la lectura de las tablas el marcado que se explica a continuación:

- Modelo de efecto fijo y tamaño 0,2: cuadro marcado en las tablas con una línea continua.
- Modelo de efecto fijo y tamaño 0,5: cuadro marcado en las tablas con una línea de puntos.
- Modelo de efecto fijo y tamaño 0,8: cuadro marcado en las tablas con una doble línea continua.
- Modelo de efecto fijo y tamaño 1,2: cuadro marcado en las tablas con una línea en zigzag.
- Modelo de efectos aleatorios y tamaño 0,2: cuadro marcado en las tablas con una línea de segmentos.
- Modelo de efectos aleatorios y tamaño 0,5: cuadro marcado en las tablas con una línea de segmentos y puntos.
- Modelo de efectos aleatorios y tamaño 0,8: cuadro marcado en las tablas con una triple línea continua.
- Modelo de efectos aleatorios y tamaño 1,2: cuadro marcado en las tablas con una línea doble en zigzag.

D.2. Cuadro comparativo de simulaciones para análisis de fiabilidad con desvío estándar de 10% e igual tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla VI, donde se muestran los resultados de la fiabilidad para experimentos de igual tamaño y desvío estándar del 10%

Se hace el análisis para cada modelo y para cada tamaño de efecto (la zona en gris indica los casos en que se alcanzó la fiabilidad del 95%):

- Modelo de efecto fijo y tamaño 0,2: para este caso se observa que sin importarla cantidad de sujetos ni la cantidad de experimentos, la fiabilidad propuesta se alcanza siempre, solo se observa una excepción para la agregación 28 sujetos distribuidos en dos experimentos (14 sujetos por experimentos).
- Modelo de efecto fijo y tamaño 0,5: en el cuadro señalado se alcanza la fiabilidad siempre y cuando la cantidad de sujetos y experimentos no sea alta. También no se alcanza la cota mínima para muchos sujetos y pocos experimentos (sí para un número medio de experimentos).
- Modelo de efecto fijo y tamaño 0,8: la tendencia observada en el caso anterior se acentúa, de tal forma que si antes no se alcanzaba la fiabilidad para muchos sujetos y pocos experimentos, ahora tampoco se alcanza para niveles medios de ambos. Ya podemos decir que se observa que a medida que se incrementa el tamaño de efecto, se pierde fiabilidad cuando el número de sujetos y experimentos crece.
- Modelo de efecto fijo y tamaño 1,2: siguiendo la tendencia observada hasta este caso, se ha perdido casi por completo la fiabilidad deseada, alcanzándose solamente cuando se agregan 2 experimentos con 4 u 8 sujetos cada uno (total de 8 y 16 sujetos experimentales respectivamente).
- Modelo de efectos aleatorios y tamaño 0,2: salvo para los casos de muchos sujetos experimentales, si la cantidad de experimentos a agregar son dos (2), entonces no se alcanza la fiabilidad, en todos los demás casos se está por encima del 95% esperado.
- Modelo de efectos aleatorios y tamaño 0,5: como en el caso del modelo de efecto fijo, ya se está en condiciones de asegurar que la tendencia para el modelo de efectos aleatorios es que a medida que el tamaño de efecto crece, la fiabilidad cae por debajo de la cota mínima a medida que la cantidad de experimentos decrece y la cantidad de sujetos crece. Así, para muchos sujetos y pocos experimentos no se llega al límite de fiabilidad del 95% (tenemos un caso de excepción para 2 sujetos y 20 experimentos).
- Modelo de efectos aleatorios y tamaño 0,8: se incrementa la tendencia, es decir, que mientras más grande es la cantidad de experimentos, debe ser más pequeña la cantidad de sujetos para alcanzar la fiabilidad de 95%. Tenemos el caso excepcional de 20 sujetos y 2 experimentos que sí alcanza la fiabilidad deseada.
- Modelo de efectos aleatorios y tamaño 1,2: la tendencia se incrementa al punto en que solamente se alcanza la fiabilidad para 10 experimentos o un nivel medio/alto de experimentos y pocos sujetos. También se presenta el caso excepcional de 20 sujetos y 2 experimentos que sí alcanza la fiabilidad deseada.
- Modelo de efecto fijo: para cada caso de tamaño de efecto, la fiabilidad crece al decrecer la cantidad de experimentos (de derecha a izquierda en cada cuadro) y al decrecer la cantidad de sujetos (de abajo hacia arriba en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la fiabilidad es mayor cuando el tamaño de efecto es menor, por lo que la fiabilidad crece al decrecer el tamaño de efecto.

TABLA VI. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS PARA COMPARACIÓN DE FIABILIDAD CON DESVÍO ESTÁNDAR DEL 10% (VALORES EXPRESADOS EN PORCENTAJES)

Fiabilidad Desvío 10%		Cantidad de experimentos (Modelo de Efecto Fijo)					Cantidad de experimentos (Modelo de Efectos Aleatorios)				
Efecto	Sujetos	2	4	6	8	10	2	4	6	8	10
0,2	4	99,0	99,4	100	99,0	99,0	92,3	96,6	99,2	99,8	100
	8	99,6	99,4	99,9	99,9	99,8	83,7	97,6	99,2	99,8	100
	10	97,6	99,9	99,6	99,8	99,3	85,8	97,0	98,7	99,9	100
	14	86,8	95,0	99,8	100	99,8	89,5	93,8	98,6	99,3	100
	20	95,8	100	100	99,8	100	98,1	98,7	100	100	100
0,5	4	98,3	99,3	99,9	98,9	96,5	87,6	95,2	98,0	99,8	100
	8	99,6	99,4	98,5	98,2	97,4	76,6	91,9	97,5	98,9	99,9
	10	96,4	100	96,9	97,9	89,6	78,7	88,6	95,3	98,6	99,7
	14	85,1	92,8	96,4	97,8	93,6	90,3	86,0	92,3	93,8	97,6
	20	92,5	99,6	97,4	93,0	90,1	98,7	80,0	90,6	96,2	99,5
0,8	4	97,2	97,9	97,1	94,0	89,2	80,9	91,8	96,9	98,8	99,8
	8	99,0	96,3	90,8	85,5	83,4	71,6	85,1	91,4	95,7	98,2
	10	94,5	96,2	87,3	83,6	74,1	76,1	78,4	86,4	95,9	98,1
	14	78,8	85,7	84,6	79,6	71,2	88,5	82,5	91,8	88,2	97,9
	20	85,5	90,8	81,1	69,4	62,0	97,9	80,6	87,1	96,1	97,5
1,2	4	95,9	91,8	86,4	78,4	70,5	74,7	90,2	95,0	96,9	99,0
	8	96,3	85,0	73,8	67,7	64,1	71,7	83,7	90,6	89,5	95,3
	10	91,1	84,4	71,3	61,9	53,5	78,2	76,4	82,1	94,4	95,2
	14	76,3	70,1	62,7	59,1	50,6	82,1	84,6	94,9	90,9	97,9
	20	73,6	67,1	58,3	50,4	44,5	99,6	84,9	94,4	93,0	97,5

Podría resumirse cada tendencia de la siguiente manera:

- Modelo de efectos aleatorios: para cada caso de tamaño de efecto, la fiabilidad crece al aumentar la cantidad de experimentos (de izquierda a derecha en cada cuadro) y al decrecer la cantidad de sujetos (de abajo hacia arriba en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la fiabilidad es mayor cuando el tamaño de efecto es menor, por lo que la fiabilidad crece al decrecer el tamaño de efecto.

D.3. Cuadro comparativo de simulaciones para análisis de fiabilidad con desvío estándar de 40% e igual tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla VII, donde se muestran los resultados de la fiabilidad para experimentos de igual tamaño y desvío estándar del 40%.

Analizando la Tabla VII, se observa que el comportamiento de ambos modelos es similar al caso en que se tiene un desvío estándar del 10%.

El comportamiento descrito para desvío del 10%, es también válido para desvío del 40%:

- Modelo de efecto fijo: para cada caso de tamaño de efecto (cada cuadro marcado), la fiabilidad crece al decrecer la cantidad de experimentos (de derecha a izquierda en cada cuadro) y al decrecer la cantidad de sujetos (de abajo hacia arriba en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la fiabilidad es mayor cuando el tamaño de efecto es menor, por lo que la fiabilidad crece al decrecer el tamaño de efecto.
- Modelo de efectos aleatorios: para cada caso de tamaño de efecto (cada cuadro marcado), la fiabilidad crece al aumentar la cantidad de experimentos (de izquierda a derecha en cada cuadro) y al decrecer la cantidad de sujetos (de abajo hacia arriba en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la fiabilidad es mayor cuando el tamaño de efecto es menor, por lo que la fiabilidad crece al decrecer el tamaño de efecto.

D.4. Cuadro comparativo de simulaciones para análisis de fiabilidad con desvío estándar de 70% e igual tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla VIII, donde se muestran los resultados de la fiabilidad para experimentos de igual tamaño y desvío estándar del 70%.

Analizando la Tabla VIII, se observa que el comportamiento de ambos modelos es similar al caso en que se tiene un desvío estándar del 10% y 40%.

Se puede concluir, que para análisis de fiabilidad de agregación de experimentos de igual tamaño, los modelos de efecto fijo y efectos aleatorios (al aplicar el método de Diferencia de Medias ponderadas, que es el método de Meta-Análisis que se está analizando) se comportan de igual manera para cualquier valor de la desviación estándar.

El análisis (que aquí se repite) es aplicable a cualquier tamaño de desvío estándar:

- Modelo de efecto fijo: para cada caso de tamaño de, la fiabilidad crece al decrecer la cantidad de experimentos (de derecha a izquierda en cada cuadro) y al decrecer la cantidad de sujetos (de abajo hacia arriba en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la fiabilidad es mayor cuando el tamaño de efecto es menor, por lo que la fiabilidad crece al decrecer el tamaño de efecto.

TABLA VII. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE FIABILIDAD CON DESVÍO ESTÁNDAR DEL 40% (VALORES EXPRESADOS EN PORCENTAJES)

Fiabilidad Desvío 40%		Cantidad de experimentos (Modelo de Efecto Fijo)					Cantidad de experimentos (Modelo de Efectos Aleatorios)				
Efecto	Sujetos	2	4	6	8	10	2	4	6	8	10
0,2	4	98,4	99,5	99,6	99,4	99,5	90,4	96,6	99,3	100	100
	8	99,8	99,6	99,8	99,8	99,6	85,3	97,6	98,9	99,9	100
	10	98,2	100	99,6	99,8	99,3	87,3	97,0	99,0	99,9	100
	14	87,5	95,1	99,8	100	99,8	88,0	94,4	97,3	99,0	99,8
	20	95,2	100	100	100	100	97,9	98,7	99,8	100	100
0,5	4	97,6	99,0	99,9	98,6	96,2	84,8	94,7	99,1	99,7	100
	8	99,5	99,5	98,6	97,4	96,1	76,1	91,5	96,6	98,3	99,9
	10	96,9	100	97,5	97,7	89,7	78,2	88,5	95,6	98,2	100
	14	83,6	94,3	96,9	96,4	93,2	90,0	85,5	93,5	94,8	98,4
	20	93,4	99,6	98,6	91,2	91,5	97,5	79,7	89,1	95,3	99,6
0,8	4	96,8	97,6	95,8	93,6	87,4	81,8	92,5	96,2	98,8	99,7
	8	98,9	95,9	91,2	87,8	85,3	72,6	85,5	94,1	96,4	98,9
	10	95,4	97,1	86,1	83,7	71,7	74,7	75,3	86,6	95,0	98,4
	14	80,7	86,5	80,3	80,1	70,7	87,3	81,8	92,2	88,9	95,7
	20	84,2	88,0	83,3	70,1	65,2	97,7	80,4	87,1	96,0	96,1
1,2	4	96,7	92,5	85,6	77,4	69,3	78,6	91,5	94,7	96,8	98,9
	8	95,2	86,1	74,2	66,4	61,7	71,0	84,9	89,7	91,2	96,1
	10	90,3	83,1	69,5	62,0	55,0	80,0	78,9	81,2	95,5	95,6
	14	78,7	70,4	62,4	53,5	48,1	87,3	84,0	94,7	92,9	97,9
	20	77,9	67,5	58,4	49,5	43,8	99,1	87,9	95,3	92,9	97,9

TABLA VIII. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS PARA COMPARACIÓN DE FIABILIDAD CON DESVÍO ESTÁNDAR DEL 70% (VALORES EXPRESADOS EN PORCENTAJES)

Fiabilidad Desvío 70%		Cantidad de experimentos (Modelo de Efecto Fijo)					Cantidad de experimentos (Modelo de Efectos Aleatorios)				
Efecto	Sujetos	2	4	6	8	10	2	4	6	8	10
0,2	4	98,5	98,6	100	99,2	99,2	92,7	95,8	99,2	100	100
	8	99,5	99,7	100	100	99,6	83,7	98,2	99,3	99,7	100
	10	98,1	100	99,7	99,8	99,2	87,6	97,6	99,1	100	99,7
	14	85,9	95,4	99,7	100	99,9	87,1	94,4	97,9	98,6	100
	20	95,9	100	100	100	100	97,0	98,6	99,9	100	100
0,5	4	98,1	99,5	100	99,2	97,2	85,5	95,0	99,0	99,8	100
	8	99,5	99,3	98,8	96,8	96,9	77,6	91,7	96,8	98,9	99,8
	10	96,1	100	97,7	97,5	91,5	79,2	90,2	95,6	98,9	99,9
	14	84,5	93,3	96,9	97,9	92,3	89,4	85,7	94,7	93,8	98,2
	20	92,4	99,2	98,0	93,1	89,7	96,8	81,9	88,7	95,9	99,5
0,8	4	96,3	97,9	96,8	93,5	91,9	81,8	90,8	97,1	98,9	99,8
	8	99,2	95,8	92,4	88,0	81,2	73,2	84,9	92,1	94,2	98,9
	10	96,0	97,1	87,1	84,8	72,0	75,1	77,3	87,6	95,6	97,8
	14	84,6	85,7	81,1	79,2	72,5	88,5	81,1	91,9	89,2	96,3
	20	87,1	88,4	80,3	72,0	65,3	98,0	79,6	88,8	95,0	96,8
1,2	4	96,8	92,9	86,6	79,7	67,0	76,8	88,4	93,6	96,7	98,8
	8	97,1	86,2	72,8	68,5	62,8	73,6	83,2	90,4	90,0	95,4
	10	90,8	81,6	71,5	65,6	56,8	78,3	79,6	82,9	94,2	95,5
	14	76,9	71,6	61,4	58,4	48,2	84,8	86,1	94,6	91,1	98,4
	20	76,2	69,9	60,0	49,3	43,8	99,0	88,8	95,9	93,0	98,0

- Modelo de efectos aleatorios: para cada caso de tamaño de efecto, la fiabilidad crece al aumentar la cantidad de experimentos (de izquierda a derecha en cada cuadro) y al decrecer la cantidad de sujetos (de abajo hacia arriba en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la fiabilidad es mayor cuando el tamaño de efecto es menor, por lo que la fiabilidad crece al decrecer el tamaño de efecto.

D.5. Cuadro comparativo de simulaciones para análisis de potencia con desvío estándar de 10% e igual tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla IX, donde se muestran los resultados de la potencia para experimentos de igual tamaño y desvío estándar del 10%

Se hace el análisis para cada modelo y para cada tamaño de efecto (la zona en gris indica los casos en que se alcanzo la potencia del 80%):

- Modelo de efecto fijo y tamaño 0,2 (cuadro marcado en la tabla con una línea continua): en este caso, no se observa potencia en ninguna circunstancia.
- Modelo de efecto fijo y tamaño 0,5 (cuadro marcado en la tabla con una línea de puntos): la potencia se presenta para muchos sujetos y muchos experimentos. También hay casos de muchos sujetos y un nivel medio/alto de experimentos. Para un nivel medio de sujetos no se alcanza la potencia del 80%.
- Modelo de efecto fijo y tamaño 0,8 (cuadro marcado en la tabla con una doble línea continua): ya se puede observar la tendencia, a medida que la cantidad de experimentos aumenta (desplazamiento en la tabla de izquierda a derecha) y la cantidad de sujetos también (desplazamiento en la tabla de arriba hacia abajo), la potencia también lo hace.
- Modelo de efecto fijo y tamaño 1,2 (cuadro marcado en la tabla con una línea en zigzag): con la tendencia descripta (aumento de la potencia con el incremento de la cantidad de sujetos y experimentos) se llega a este caso en que solo no se alcanza la potencia del 80% para pocos experimentos y pocos sujetos.
- Para el modelo de efectos aleatorios no se observa potencia en ningún caso

En este caso, puede resumirse la tendencia como sigue:

- Modelo de efecto fijo: para cada caso de tamaño de efecto (cada cuadro marcado), la potencia crece al crecer la cantidad de experimentos (desplazamiento de izquierda a derecha en cada cuadro) y al crecer la cantidad de sujetos (de arriba hacia abajo en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la potencia es mayor cuando el tamaño de efecto es mayor, por lo que la potencia crece al crecer el tamaño de efecto. Al ver este análisis y compararlo con el análisis de fiabilidad, la tendencia es inversa.
- Modelo de efectos aleatorios: no alcanza la potencia en ningún caso.

D.6. Cuadro comparativo de simulaciones para análisis de potencia con desvío estándar de 40% e igual tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla X., donde se muestran los resultados de la potencia para experimentos de igual tamaño y desvío estándar del 40%.

TABLA IX. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE POTENCIA CON DESVÍO ESTÁNDAR DEL 10% (VALORES EXPRESADOS EN PORCENTAJES)

Potencia Desvío 10%		Cantidad de experimentos (Modelo de Efecto Fijo)					Cantidad de experimentos (Modelo de Efectos Aleatorios)				
Efecto	Sujetos	2	4	6	8	10	2	4	6	8	10
0,2	4	1,5	2,1	2,4	2,3	2,3	0	0,2	0	0	0
	8	2,6	4,4	4,3	4,9	6,7	0	0	0	0	0
	10	8,7	0,9	9,0	8,3	17,2	0	0	0	0	0
	14	20,3	12,8	11,1	8,4	16,2	0	0	0	0	0
	20	6,0	2,8	6,0	17,2	15,1	0	0	0	0	0
0,5	4	5,4	11,6	16,1	19,7	27,5	0	0,1	0	0	0
	8	14,5	30,3	43,9	59,5	74,7	0	0,9	0,1	0	0
	10	25,5	34,5	59,0	76,8	77,9	0	0,2	0	0,1	0
	14	39,3	53,4	73,5	95,5	96,3	0	0,5	0,3	0	0
	20	36,9	81,9	95,6	98,8	99,8	0	0,3	0,2	0,2	0
0,8	4	12,5	31,4	47,4	63,8	71,0	0,5	0,5	0,1	0	0
	8	39,2	68,4	87,9	95,9	98,7	0,1	1,6	0,2	0,1	0,1
	10	51,6	86,4	93,4	98,5	99,8	0,2	1,4	0,7	0	0
	14	62,6	91,1	98,4	99,8	99,9	0,8	1,5	0,1	0,3	0,1
	20	76,4	99,6	100	99,8	99,8	0	1,4	0,2	0,3	0,2
1,2	4	33,0	64,5	88,1	94,2	97,0	2,7	0,7	0,6	0,1	0
	8	78,1	98,0	99,1	99,8	99,8	1,1	0,4	0,5	0,4	0,2
	10	86,9	98,9	99,9	100	99,9	1,9	0,8	2,6	0,7	0,3
	14	90,3	99,9	99,7	100	100	6,2	0,9	0	0,6	0,2
	20	96,7	100	100	100	100	0,3	0,6	1,3	0,3	0

TABLA X. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE POTENCIA CON DESVÍO ESTÁNDAR DEL 40% (VALORES EXPRESADOS EN PORCENTAJES)

Potencia Desvío 40%		Cantidad de experimentos (Modelo de Efecto Fijo)					Cantidad de experimentos (Modelo de Efectos Aleatorios)				
Efecto	Sujetos	2	4	6	8	10	2	4	6	8	10
0,2	4	1,8	3,1	2,1	2,1	3,6	0	0	0	0	0
	8	1,9	4,7	3,7	4,6	7,4	0	0	0,1	0	0
	10	7,0	1,1	9,1	7,2	16,7	0	0	0	0	0
	14	16,4	13,0	12,2	6,6	18,9	0	0	0	0	0
	20	5,9	3,8	5,9	19,8	16,7	0	0	0	0	0
0,5	4	6,1	11,5	14,9	21,7	26,8	0,3	0,4	0	0	0
	8	12,6	27,5	42,6	59,6	74,6	0	0,6	0	0,1	0
	10	24,3	37,8	61,2	73,3	77,4	0	0,1	0,2	0	0
	14	38,1	55,3	75,1	93,2	96,2	0,1	0,1	0,4	0,1	0
	20	39,0	83,6	97,8	98,6	99,6	0	0,4	0,2	0	0
0,8	4	14,0	30,2	43,4	62,5	71,1	0,8	0,6	0,2	0	0
	8	39,9	69,6	88,1	96,2	99,1	0,2	1,8	0,2	0,4	0
	10	52,1	87,7	94,9	98,4	99,3	0,2	0,5	0,8	0	0
	14	64,0	91,5	98,9	100	99,9	0,7	1,0	0,2	0,6	0
	20	75,5	98,5	99,9	100	99,9	0	0,7	0,4	0,3	0,1
1,2	4	29,6	68,8	87,2	94,6	96,4	2,2	0,9	1,1	0,1	0
	8	79,4	96,9	99,6	99,7	99,9	1,6	0,6	0,2	0,7	0,5
	10	85,8	99,2	99,7	100	100	2,3	1,1	2,6	0,1	0,3
	14	90,5	99,4	99,9	100	100	3,8	0,9	0,4	0,4	0,3
	20	97,0	100	99,9	100	100	0,2	0,2	1,7	0	0,1

TABLA XI. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE POTENCIA CON DESVÍO ESTÁNDAR DEL 70% (VALORES EXPRESADOS EN PORCENTAJES)

Potencia Desvío 70%		Cantidad de experimentos (Mod. Efecto Fijo)					Cantidad de experimentos (Mod. Efectos Aleatorios)				
Efecto	Sujetos	2	4	6	8	10	2	4	6	8	10
0,2	4	2,6	3,2	2,5	2,2	2,3	0	0,2	0	0	0
	8	2,7	4,6	3,7	4,1	6,2	0	0	0	0	0
	10	7,3	0,9	8,8	8,4	17,3	0	0,1	0	0	0
	14	19,4	14,5	13,1	8,0	16,2	0	0	0	0	0
	20	7,6	3,4	7,2	18,3	14,5	0	0	0	0	0
0,5	4	6,2	9,5	15,6	19,6	25,0	0,2	0,3	0	0	0
	8	14,1	26,8	44,0	60,9	73,3	0	0,3	0	0,2	0
	10	27,0	38,2	61,8	74,7	79,5	0	0,7	0,2	0	0
	14	38,0	54,9	73,0	95,1	96,1	0	0,3	0	0,1	0
	20	39,2	80,9	96,4	98,7	99,2	0	0,6	0,1	0,2	0
0,8	4	12,4	29,7	45,4	59,1	73,8	0,9	0,3	0,4	0	0
	8	37,8	69,5	90,2	96,8	98,2	0	1,9	0	0,3	0
	10	52,3	87,6	95,6	98,8	99,4	0,1	0,7	0,8	0	0
	14	64,4	92,4	98,7	99,5	100	1,0	1,0	0,2	0,2	0
	20	77,3	99,0	99,9	100	99,9	0	0,6	0,4	0,1	0
1,2	4	32,2	67,2	88,4	96,2	96,4	2,3	1,0	1,4	0,1	0,1
	8	80,1	97,7	99,2	99,9	99,9	0,6	0,7	0,2	0,5	0,4
	10	86,5	99,8	99,8	100	99,9	1,2	1,3	3,8	0,2	0,1
	14	91,1	99,4	99,9	100	100	4,6	1,0	0,2	0,5	0
	20	96,4	99,8	100	100	100	0,4	0,4	1,5	0	0

Analizando la Tabla X, se observa que el comportamiento de ambos modelos es similar al caso en que se tiene un desvío estándar del 10%. El comportamiento descrito para desvío del 10%, es también válido para desvío del 40%:

- Modelo de efecto fijo: para cada caso de tamaño de efecto (cada cuadro marcado), la potencia crece al crecer la cantidad de experimentos (desplazamiento de izquierda a derecha en cada cuadro) y al crecer la cantidad de sujetos (de arriba hacia abajo en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la potencia es mayor cuando el tamaño de efecto es mayor, por lo que la potencia crece al crecer el tamaño de efecto.
- Modelo de efectos aleatorios: no alcanza la potencia en ningún caso.

D.7. Cuadro comparativo de simulaciones para análisis de potencia con desvío estándar de 70% e igual tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla XI, donde se muestran los resultados de la potencia para experimentos de igual tamaño y desvío estándar del 70%

Al analizar la Tabla XI, se puede observar que los resultados son los mismos que los que se presentan para desvíos bajo y medio. Se observa la misma tendencia y resultados similares en todos los casos. Se puede concluir, que para análisis de potencia de agregación de experimentos de igual tamaño, los modelos de efecto fijo y efectos aleatorios (al aplicar el método de Diferencia de Medias ponderadas, que es el método de Meta-Análisis que se está analizando) se comportan de igual manera para cualquier valor de la desviación estándar. El análisis (que aquí se repite) es aplicable a cualquier tamaño de desvío estándar:

- Modelo de efecto fijo: para cada caso de tamaño de efecto (cada cuadro marcado), la potencia crece al crecer la cantidad de experimentos (desplazamiento de izquierda a derecha en cada cuadro) y al crecer la cantidad de sujetos (de arriba hacia abajo en cada cuadro). Además, para cada cuadro se ve que la cantidad de casos en que se alcanza la potencia es mayor cuando el tamaño de efecto es mayor, por lo que la potencia crece al crecer el tamaño de efecto.
- Modelo de efectos aleatorios: no alcanza la potencia en ningún caso.

D.8. Convención adoptada para lectura de las tablas que muestran los resultados de los experimentos de distinto tamaño

Para analizar los resultados de las simulaciones y poder visualizarlos de manera rápida y clara, se adopta para la lectura de las tablas el marcado que se explica a continuación:

- Modelo de efecto fijo y 3 experimentos: cuadro marcado en la tabla con una línea continua.
- Modelo de efecto fijo y 6 experimentos: cuadro marcado con una línea de puntos.
- Modelo de efecto fijo y 9 experimentos: cuadro marcado en la tabla con una doble línea continua.
- Modelo de efectos aleatorios y 3 experimentos: cuadro marcado en la tabla con una línea de segmentos.
- Modelo de efectos aleatorios y 6 experimentos: cuadro marcado en la tabla con una línea de segmentos y puntos.
- Modelo de efectos aleatorios y 9 experimentos: cuadro marcado en la tabla con una triple línea continua.

D.9. Cuadro comparativo de simulaciones para análisis de fiabilidad con desvío estándar de 10% y distinto tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla XII, donde se muestran los resultados de la fiabilidad para experimentos de distinto tamaño y desvío estándar del 10%.

TABLA XII. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE FIABILIDAD CON DESVÍO ESTÁNDAR DEL 10% (VALORES EXPRESADOS EN PORCENTAJES)

		Modelo de efecto fijo				Modelo de efectos aleatorios			
Sujetos	Experi-mentos	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2
4-4-14	3	94,3	93,3	91,2	83,8	96,2	87,7	78,0	72,6
4-4-20	3	98,4	98,6	97,9	89,4	99,6	93,4	78,3	68,5
4-14-14	3	91,3	88,2	81,5	70,4	96,4	90,7	85,2	82,0
4-14-20	3	92,1	90,2	80,6	70,3	99,8	98,0	86,7	83,2
4-20-20	3	96,6	93,3	84,5	73,9	97,3	93,0	81,9	76,3
14-14-20	3	91,4	87,1	80,2	71,5	98,9	86,2	83,3	84,0
14-20-20	3	94,1	90,2	82,3	67,8	96,6	88,5	84,8	89,6
4-4-4-4-14-14	6	97,5	93,0	85,3	70,3	100	98,6	96,0	93,1
4-4-4-4-20-20	6	97,1	92,8	82,1	65,5	100	99,9	96,3	92,7
4-4-14-14-14-14	6	96,2	92,5	82,5	65,3	99,1	92,4	86,8	86,2
4-4-14-14-20-20	6	94,1	90,2	76,6	59,6	100	98,0	87,7	86,9
4-4-20-20-20-20	6	100	98,3	81,8	60,7	100	99,0	94,6	92,4
14-14-14-14-20-20	6	97,1	91,2	74,1	57,1	100	98,0	94,4	96,2
14-14-20-20-20-20	6	99,1	91,7	78,0	57,5	100	93,9	89,1	92,8
4-4-4-4-4-4-14-14-14	9	96,6	90,9	79,4	65,1	100	99,5	98,2	96,0
4-4-4-4-4-4-20-20-20	9	100	97,5	83,2	61,6	100	100	100	97,4
4-4-4-14-14-14-14-14-14	9	99,9	91,8	72,7	53,6	100	99,2	97,1	97,0
4-4-4-14-14-14-20-20-20	9	98,5	91,5	73,5	55,7	100	99,7	96,1	91,8
4-4-4-20-20-20-20-20-20	9	100	95,3	74,2	51,3	100	99,3	97,1	95,6
14-14-14-14-14-14-20-20-20	9	99,8	88,6	72,6	47,9	100	99,9	96,8	97,9
14-14-14-20-20-20-20-20-20	9	99,8	91,4	68,4	53,9	100	98,2	93,7	94,1

Se hace el análisis para cada modelo y para cada cantidad de experimentos (la zona en gris indica los casos en que se alcanzó la fiabilidad del 95%):

- Modelo de efecto fijo y 3 experimentos (cuadro marcado en la tabla con una línea continua): en este cuadro se observa que se alcanza la fiabilidad para tamaño de efecto pequeño y combinaciones de sujetos por experimentos de 4-4-20 y 4-20-20. Para efecto medio y alto hay fiabilidad para 4-4-20. Eso significa que la fiabilidad se alcanza cuando la diferencia de tamaño entre experimentos a agregar es máxima.
- Modelo de efecto fijo y 6 experimentos (cuadro marcado en la tabla con una línea de puntos): en este caso, puede observarse que se alcanza la fiabilidad propuesta para tamaño de efecto pequeño (menos el caso en que se combinan 4-4-14-14-20-20 sujetos por experimento).
- Modelo de efecto fijo y 9 experimentos (cuadro marcado en la tabla con una doble línea continua): se alcanza fiabilidad en todos los casos de tamaño de efecto pequeño y para tamaño de efecto medio solo para los casos en que la diferencia de tamaño de experimentos a agregar es máxima (4-4-4-4-4-4-20-20-20 y 4-4-4-20-20-20-20).
- Modelo de efectos aleatorios y 3 experimentos (cuadro marcado en la tabla con una línea de segmentos): los casos en que se presenta fiabilidad es para tamaño de efecto pequeño.
- Modelo de efectos aleatorios y 6 experimentos (cuadro marcado en la tabla con una línea de segmentos y puntos): hay fiabilidad para tamaño de efecto pequeño y tamaño de efecto medio salvo en dos casos (4-4-14-14-14-14 y 14-14-20-20-20-20). También hay dos casos de fiabilidad para tamaño de efecto grande (4-4-4-4-14-14 y 4-4-4-4-20-20).
- Modelo de efectos aleatorios y 9 experimentos (cuadro marcado en la tabla con una triple línea continua): se alcanza la fiabilidad en casi todos los casos. Las excepciones son 3 (sobre 25): 14-14-14-20-20-20-20-20-20 para tamaño de efecto grande y muy grande y 4-4-4-14-14-14-20-20-20 para tamaño de efecto muy grande.

Se puede resumir la tendencia observada para desvío del 10% como sigue:

- Modelo de efecto fijo: a medida que se incrementa la cantidad de experimentos y el tamaño de efecto es pequeño, se incrementa la fiabilidad. También se observa que mejora la potencia cuando la diferencia de tamaño de experimentos a agregar es máxima (los casos de combinar experimentos de 4 y 20 sujetos).
- Modelo de efectos aleatorios: la fiabilidad se incrementa al subir la cantidad de experimentos y (el desplazamiento sobre la tabla es de cuadro a cuadro de arriba hacia abajo) y crecer el tamaño de efecto (el desplazamiento en la tabla en cada cuadro es de izquierda a derecha).

D.10. Cuadro comparativo de simulaciones para análisis de fiabilidad con desvío estándar de 40% y distinto tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla XIII, donde se muestran los resultados de la fiabilidad para experimentos de distinto tamaño y desvío estándar del 40%. Analizando la Tabla XIII, se observa que el comportamiento de ambos modelos es similar al caso en que se tiene un desvío estándar del 10%. El comportamiento descrito para desvío del 10%, es también válido para desvío del 40%:

- Modelo de efecto fijo: a medida que se incrementa la cantidad de experimentos y el tamaño de efecto es pequeño, se incrementa la fiabilidad. También se observa que mejora la potencia cuando la diferencia de tamaño de experimentos a agregar es máxima (los casos de combinar experimentos de 4 y 20 sujetos).
- Modelo de efectos aleatorios: la fiabilidad se incrementa al subir la cantidad de experimentos y (el desplazamiento sobre la tabla es de cuadro a cuadro de arriba hacia abajo) y crecer el tamaño de efecto (el desplazamiento en la tabla en cada cuadro es de izquierda a derecha).

TABLA XIII. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE FIABILIDAD CON DESVÍO ESTÁNDAR DEL 40% (VALORES EXPRESADOS EN PORCENTAJES)

		Modelo de efecto fijo				Modelo de efectos aleatorios			
Sujetos	Experi-mentos	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2
4-4-14	3	93,4	92,5	91,8	82,9	96,3	89,6	80,2	73,6
4-4-20	3	98,4	98,0	96,6	90,1	99,2	91,8	76,0	70,3
4-14-14	3	92,9	87,4	80,3	70,0	96,6	91,3	84,6	84,1
4-14-20	3	92,6	90,9	81,8	72,0	99,3	97,5	87,0	82,9
4-20-20	3	96,5	92,0	87,4	73,7	98,0	92,0	82,9	79,6
14-14-20	3	89,6	87,5	81,9	69,6	98,9	86,5	83,1	84,3
14-20-20	3	94,1	90,1	83,1	67,3	97,1	88,6	82,9	90,1
4-4-4-4-14-14	6	97,8	92,8	86,5	69,2	99,9	99,8	96,2	92,6
4-4-4-4-20-20	6	97,7	92,5	80,5	66,6	100	99,8	96,1	91,0
4-4-14-14-14-14	6	96,1	91,5	81,3	66,8	99,3	93,6	85,2	84,7
4-4-14-14-20-20	6	95,1	91,2	75,5	63,2	100	98,6	88,9	86,0
4-4-20-20-20-20	6	100	98,6	83,7	59,8	100	98,2	94,1	92,8
14-14-14-14-20-20	6	95,7	90,1	74,1	57,9	100	97,6	94,9	96,8
14-14-20-20-20-20	6	99,5	92,9	76,6	57,3	100	94,7	88,3	92,2
4-4-4-4-4-4-14-14-14	9	96,6	92,2	79,5	61,9	100	99,1	98,5	96,5
4-4-4-4-4-4-20-20-20	9	100	97,5	82,1	62,6	100	100	100	97,1
4-4-4-14-14-14-14-14-14	9	99,9	91,1	75,2	54,8	100	99,3	97,1	97,7
4-4-4-14-14-14-20-20-20	9	98,8	91,6	70,6	54,0	100	100	96,9	93,5
4-4-4-20-20-20-20-20-20	9	100	94,4	73,6	53,7	100	99,8	97,1	94,5
14-14-14-14-14-14-20-20-20	9	100	89,2	66,5	46,5	100	99,9	97,7	97,2
14-14-14-20-20-20-20-20-20	9	100	90,6	69,7	49,7	100	98,1	93,5	94,3

D.11. Cuadro comparativo de simulaciones para análisis de fiabilidad con desvío estándar de 70% y distinto tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla XIV, donde se muestran los resultados de la fiabilidad para experimentos de distinto tamaño y desvío estándar del 70%. Analizando la Tabla XIV, se observa que el comportamiento de ambos modelos es similar al caso en que se tiene un desvío estándar del 10%. Se puede concluir, que para análisis de fiabilidad de agregación de experimentos de igual tamaño, los modelos de efecto fijo y efectos aleatorios (al aplicar el método de Diferencia de Medias ponderadas, que es el método de Meta-Análisis que se está analizando) se comportan de igual manera para cualquier valor de la desviación estándar. El análisis (que aquí se repite) es aplicable a cualquier tamaño de desvío estándar:

- Modelo de efecto fijo: a medida que se incrementa la cantidad de experimentos y el tamaño de efecto es pequeño, se incrementa la fiabilidad. También se observa que mejora la potencia cuando la diferencia de tamaño de experimentos a agregar es máxima (los casos de combinar experimentos, donde los mismos tienen 4 o 20 sujetos, como ser 4-4-20 o 4-20-20).
- Modelo de efectos aleatorios: la fiabilidad se incrementa al subir la cantidad de experimentos y (el desplazamiento sobre la tabla es de cuadro a cuadro de arriba hacia abajo) y crecer el tamaño de efecto (el desplazamiento en la tabla en cada cuadro es de izquierda a derecha).

D.12. Cuadro comparativo de simulaciones para análisis de potencia con desvío estándar de 10% y distinto tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla XV, donde se muestran los resultados de la potencia para experimentos de distinto tamaño y desvío estándar del 10%. Se hace el análisis para cada modelo y para cada cantidad de experimentos (la zona en gris indica los casos en que se alcanzo la potencia del 80%):

- Modelo de efecto fijo y 3 experimentos (cuadro marcado en la tabla con una línea continua): en este caso, se alcanza la potencia propuesta del 80% cuando el tamaño de efecto es muy grande, o para tamaño de efecto grande si la cantidad total de sujetos es elevada (que es para los casos en que se combinan 14 y 20 sujetos por experimento)
- Modelo de efecto fijo y 6 experimentos (cuadro marcado en la tabla con una línea de puntos): hay potencia cuando el tamaño de efecto es grande y muy grande (para el caso de tamaño de efecto de 0,8, no se alcanza la potencia cuando la cantidad de sujetos totales es baja, es decir, 4-4-14). También hay potencia para tamaño de efecto medio si la cantidad de sujetos totales es alta.
- Modelo de efecto fijo y 9 experimentos (cuadro marcado en la tabla con una doble línea continua): en este caso se alcanza la potencia del 80% para tamaño de efecto grande y muy grande. Para tamaño de efecto medio, se alcanza la potencia mínima deseada para una cantidad de sujetos totales media/alta.
- Para el modelo de efectos aleatorios no se observa potencia en ningún caso.

Analizando la tendencia observada a lo largo de los cuadros de la Tabla XV, puede resumirse el comportamiento de cada modelo para desvío estándar del 10% como sigue:

- Modelo de efecto fijo: la potencia se incrementa al subir la cantidad de experimentos (desplazamiento en la tabla de arriba hacia debajo de cuadro a cuadro) y al crecer el tamaño de efecto (desplazamiento dentro de cada cuadro de izquierda a derecha).
- Para el modelo de efectos aleatorios no se observa potencia en ningún caso

TABLA XIV. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE FIABILIDAD CON DESVÍO ESTÁNDAR DEL 70% (VALORES EXPRESADOS EN PORCENTAJES)

		Modelo de efecto fijo				Modelo de efectos aleatorios			
Sujetos	Experi-mentos	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2
4-4-14	3	94,0	92,4	90,6	85,0	96,1	89,5	77,3	73,1
4-4-20	3	99,2	98,6	96,8	89,0	99,3	92,8	78,5	69,8
4-14-14	3	90,5	84,5	82,7	72,5	95,7	91,9	81,0	83,3
4-14-20	3	92,7	88,9	80,6	70,6	99,7	96,8	87,8	84,1
4-20-20	3	96,8	91,5	87,7	74,7	99,3	92,6	82,2	75,4
14-14-20	3	90,3	86,5	81,1	71,2	98,6	89,2	82,5	82,7
14-20-20	3	93,7	90,9	84,3	69,9	97,4	87,2	84,7	90,0
4-4-4-4-14-14	6	97,7	93,3	84,2	68,5	100	99,4	96,8	94,4
4-4-4-4-20-20	6	97,1	93,5	81,0	63,3	100	99,9	96,2	91,5
4-4-14-14-14-14	6	96,8	91,1	79,3	63,1	99,0	92,1	84,3	82,9
4-4-14-14-20-20	6	95,9	90,7	78,6	61,3	100	98,4	86,2	86,9
4-4-20-20-20-20	6	100	98,6	82,0	60,6	100	98,8	94,5	92,3
14-14-14-14-20-20	6	96,2	88,3	76,5	55,1	100	97,7	93,2	96,5
14-14-20-20-20-20	6	99,1	92,0	78,2	55,3	99,9	94,3	88,8	91,7
4-4-4-4-4-4-14-14-14	9	95,9	90,6	79,0	62,2	99,9	99,4	98,2	96,0
4-4-4-4-4-4-20-20-20	9	100	97,8	83,2	61,9	100	100	100	97,0
4-4-4-14-14-14-14-14-14	9	100	92,8	73,6	52,8	100	99,3	97,2	97,9
4-4-4-14-14-14-20-20-20	9	98,0	93,2	75,7	54,4	100	99,6	97,4	92,4
4-4-4-20-20-20-20-20-20	9	100	94,7	74,9	50,7	100	99,6	96,7	94,7
14-14-14-14-14-14-20-20-20	9	99,9	88,5	69,8	43,7	100	99,9	98,0	98,1
14-14-14-20-20-20-20-20-20	9	99,8	91,7	71,0	47,2	100	99,1	94,5	94,7

TABLA XV. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE POTENCIA CON DESVÍO ESTÁNDAR DEL 10% (VALORES EXPRESADOS EN PORCENTAJES)

		Modelo de efecto fijo				Modelo de efectos aleatorios			
Sujetos	Experi-mentos	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2
4-4-14	3	7,3	25,1	49,1	84,4	0	0,2	0,9	2,7
4-4-20	3	4,3	26,4	67,6	96,0	0	0,1	0,9	2,5
4-14-14	3	10,9	33,4	58,6	90,9	0	0,1	0	0,9
4-14-20	3	8,5	29,3	70,5	94,0	0	0	0	0,4
4-20-20	3	9,0	38,4	79,1	98,6	0	0,5	0,8	0,4
14-14-20	3	17,2	45,4	81,0	98,9	0	0,1	0,7	0,8
14-20-20	3	13,8	54,4	87,4	98,8	0	1,5	1,4	0,4
4-4-4-4-14-14	6	7,8	35,3	78,6	98,7	0	0	0	0,2
4-4-4-4-20-20	6	6,7	46,1	88,5	99,1	0	0	0,1	0,7
4-4-14-14-14-14	6	10,6	61,4	92,2	99,7	0	0,3	0,6	0,5
4-4-14-14-20-20	6	17,1	64,8	93,8	99,9	0	0	0,9	0,5
4-4-20-20-20-20	6	3,4	82,8	99,3	100	0	0,1	0,3	0,5
14-14-14-14-20-20	6	16,7	78,1	99,3	100	0	0	0,1	0,1
14-14-20-20-20-20	6	10,9	87,5	99,6	100	0	0	0,4	0,3
4-4-4-4-4-4-14-14-14	9	7,3	55,7	91,5	99,7	0	0	0	0,8
4-4-4-4-4-4-20-20-20	9	7,4	78,0	99,4	100	0	0	0	0,2
4-4-4-14-14-14-14-14-14	9	10,4	73,5	98,4	99,9	0	0	0,1	0,2
4-4-4-14-14-14-20-20-20	9	18,4	88,3	99,9	100	0	0	0,1	0,8
4-4-4-20-20-20-20-20-20	9	7,4	96,3	100	100	0	0	0,1	0,2
14-14-14-14-14-14-20-20-20	9	16,5	93,6	100	100	0	0	0,1	0,1
14-14-14-20-20-20-20-20-20	9	27,5	98,6	100	100	0	0	0,2	0,2

TABLA XVI. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE POTENCIA CON DESVÍO ESTÁNDAR DEL 40% (VALORES EXPRESADOS EN PORCENTAJES)

		Modelo de efecto fijo				Modelo de efectos aleatorios			
Sujetos	Experi-mentos	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2
4-4-14	3	8,8	25,1	51,7	83,2	0	0	0,9	2,8
4-4-20	3	5,1	27,3	67,4	96,2	0	0,1	0,7	1,7
4-14-14	3	9,8	33,3	58,8	91,2	0	0	0,1	0,5
4-14-20	3	7,4	28,9	71,0	95,4	0	0	0	0,4
4-20-20	3	6,9	44,0	82,4	97,2	0	0,7	0,8	0,6
14-14-20	3	17,1	44,9	82,7	98,5	0	0	1,2	1,0
14-20-20	3	10,9	52,4	87,4	99,5	0,1	0,6	1,8	0,1
4-4-4-4-14-14	6	7,9	39,0	78,8	98,5	0	0	0,2	0,6
4-4-4-4-20-20	6	6,3	46,7	87,4	99,7	0	0	0	0,6
4-4-14-14-14-14	6	11,1	57,6	92,3	99,7	0,1	0	0,6	0,5
4-4-14-14-20-20	6	17,0	66,8	93,6	99,8	0	0	0,1	1,4
4-4-20-20-20-20	6	3,5	83,4	99,4	100	0	0	0,2	1,1
14-14-14-14-20-20	6	16,7	76,6	99,2	99,9	0	0,1	0	0
14-14-20-20-20-20	6	13,7	87,8	99,4	100	0	0	0,4	0,4
4-4-4-4-4-4-14-14-14	9	9,6	57,3	92,0	99,6	0	0	0	0,4
4-4-4-4-4-4-20-20-20	9	4,2	77,6	98,8	99,9	0	0	0	0,2
4-4-4-14-14-14-14-14-14	9	9,3	73,8	99,5	100	0	0	0,1	0,1
4-4-4-14-14-14-20-20-20	9	19,6	87,8	99,8	100	0	0	0,1	1,0
4-4-4-20-20-20-20-20-20	9	6,8	95,6	99,8	100	0	0	0	0,3
14-14-14-14-14-14-20-20-20	9	16,0	94,3	99,9	100	0	0	0,1	0,3
14-14-14-20-20-20-20-20-20	9	26,2	98,4	99,9	100	0	0	0,4	0,4

D.13. Cuadro comparativo de simulaciones para análisis de potencia con desvío estándar de 40% y distinto tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla XVI, donde se muestran los resultados de la potencia para experimentos de distinto tamaño y desvío estándar del 40%

Analizando la Tabla XVI, se observa que el comportamiento de ambos modelos es similar al caso en que se tiene un desvío estándar del 10%. El comportamiento descrito para desvío del 10%, es también válido para desvío del 40%:

- Modelo de efecto fijo: la potencia se incrementa al subir la cantidad de experimentos (desplazamiento en la tabla de arriba hacia debajo de cuadro a cuadro) y al crecer el tamaño de efecto (desplazamiento dentro de cada cuadro de izquierda a derecha).
- Para el modelo de efectos aleatorios no se observa potencia en ningún caso

D.14. Cuadro comparativo de simulaciones para análisis de potencia con desvío estándar de 70% y distinto tamaño de experimentos

Los resultados de las simulaciones hechas para los modelos de efecto fijo y efecto aleatorio se sintetizan en la Tabla XVII, donde se muestran los resultados de la potencia para experimentos de distinto tamaño y desvío estándar del 70%

Al analizar la Tabla XVII, se puede observar que los resultados son los mismos que los que se presentan para desvíos bajo y medio. Se observa la misma tendencia y resultados similares en todos los casos.

Se puede concluir, que para análisis de potencia de agregación de experimentos de igual tamaño, los modelos de efecto fijo y efectos aleatorios (al aplicar el método de Diferencia de Medias ponderadas, que es el método de Meta-Análisis que se está analizando) se comportan de igual manera para cualquier valor de la desviación estándar.

El análisis (que aquí se repite) es aplicable a cualquier tamaño de desvío estándar:

- Modelo de efecto fijo: la potencia se incrementa al subir la cantidad de experimentos (desplazamiento en la tabla de arriba hacia debajo de cuadro a cuadro) y al crecer el tamaño de efecto (desplazamiento dentro de cada cuadro de izquierda a derecha).
- Para el modelo de efectos aleatorios no se observa potencia en ningún caso

E. Resumen de los resultados obtenidos y recomendaciones

A continuación se mostrará un resumen del análisis de los resultados obtenidos en la experimentación para luego elaborar un conjunto de recomendaciones sobre el modelo de Meta-Análisis a utilizar para agregar experimentos en Ingeniería de Software Empírica, según sean las características dadas.

E.1. Resumen de los resultados obtenidos en la experimentación

Los resultados obtenidos en la sección V.D, deben agruparse bajo un criterio que haga simple su análisis visual (donde se utiliza las convenciones adoptadas y explicadas en las secciones D.1 y D.8), para lo cual se diseñaron las tablas que siguen a continuación, donde se resumen los resultados de la comparación de ambos métodos para agrupar experimentos de igual tamaño (Tabla XVIII) y experimentos de distinto tamaños (Tabla XIX).

La alta fiabilidad se indica con (+) y la baja con espacio en blanco. En tanto que la alta potencia se indica con (//) y la baja potencia con espacio en blanco.

TABLA XVII. CUADRO DE SIMULACIONES PARA AMBOS MODELOS DE EFECTO DE META-ANÁLISIS, PARA COMPARACIÓN DE POTENCIA CON DESVÍO ESTÁNDAR DEL 70% (VALORES EXPRESADOS EN PORCENTAJES)

		Modelo de efecto fijo				Modelo de efectos aleatorios			
Sujetos	Experi-mentos	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2
4-4-14	3	8,7	22,6	51,0	85,2	0,1	0	0,8	2,4
4-4-20	3	5,0	28,6	67,5	96,6	0	0	0,2	1,0
4-14-14	3	12,7	35,3	61,4	92,4	0	0	0	0,6
4-14-20	3	7,9	30,4	70,0	94,1	0	0	0	0,1
4-20-20	3	5,4	41,8	83,4	97,5	0	0,4	0,9	0,9
14-14-20	3	15,3	44,9	82,3	99,4	0	0	0,8	1,5
14-20-20	3	11,9	52,0	88,9	99,1	0	0,9	1,7	0,2
4-4-4-4-14-14	6	8,1	34,7	78,2	98,3	0	0	0	0,1
4-4-4-4-20-20	6	7,0	46,6	87,8	99,6	0	0	0	0,6
4-4-14-14-14-14	6	11,4	54,9	89,8	99,5	0	0,2	0,9	0,6
4-4-14-14-20-20	6	18,5	64,0	94,0	100	0	0	0,3	1,0
4-4-20-20-20-20	6	3,5	82,2	99,2	99,6	0	0	0,2	0,6
14-14-14-14-20-20	6	18,4	76,2	99,2	100	0	0	0	0,1
14-14-20-20-20-20	6	13,7	86,9	99,8	100	0	0	0,3	0,2
4-4-4-4-4-4-14-14-14	9	9,1	55,6	91,4	99,5	0	0	0,1	0,6
4-4-4-4-4-4-20-20-20	9	3,8	78,4	99,1	99,9	0	0	0	0,3
4-4-4-14-14-14-14-14-14	9	9,6	74,9	99,5	99,9	0	0	0,1	0,2
4-4-4-14-14-14-20-20-20	9	19,2	90,4	99,6	100	0	0	0,1	0,8
4-4-4-20-20-20-20-20-20	9	8,0	96,3	99,7	100	0	0	0,5	0,4
14-14-14-14-14-14-20-20-20	9	15,2	93,7	100	100	0	0	0	0
14-14-14-20-20-20-20-20-20	9	25,8	98,4	100	100	0	0	0,5	0,2

TABLA XVIII. CUADRO COMPARATIVO DE AMBOS MODELOS PARA EXPERIMENTOS DE IGUAL TAMAÑO. COMPARACIÓN DE FIABILIDAD (+) Y POTENCIA ESTADÍSTICA (//)

Fiabilidad y Potencia		Cantidad de experimentos (Modelo de Efecto Fijo)					Cantidad de experimentos (Modelo de Efectos Aleatorios)				
Efecto	Sujetos	2	4	6	8	10	2	4	6	8	10
0,2	4	+	+	+	+	+		+	+	+	+
	8	+	+	+	+	+		+	+	+	+
	10	+	+	+	+	+		+	+	+	+
	14		+	+	+	+			+	+	+
	20	+	+	+	+	+	+	+	+	+	+
0,5	4	+	+	+	+	+			+	+	+
	8	+	+	+	+	+			+	+	+
	10	+	+	+	+				+	+	+
	14			+	+ //	+ //					+
	20		+ //	+ //	//	//	+			+	+
0,8	4	+	+	+					+	+	+
	8	+	+	//	//	//				+	+
	10	+	+ //	//	//	//				+	+
	14		//	//	//	//					+
	20		//	//	//	//	+			+	+
1,2	4	+		//	//	//				+	+
	8	+	//	//	//	//					+
	10		//	//	//	//				+	+
	14		//	//	//	//					+
	20		//	//	//	//	+		+		+

TABLA XIX. CUADRO COMPARATIVO DE AMBOS MODELOS PARA EXPERIMENTOS DE DISTINTO TAMAÑO. COMPARACIÓN DE FIABILIDAD (+) Y POTENCIA ESTADÍSTICA (//)

		Modelo de efecto fijo				Modelo de efectos aleatorios			
Sujetos	Experi-mentos	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2	Efecto 0,2	Efecto 0,5	Efecto 0,8	Efecto 1,2
4-4-14	3				//	+			
4-4-20	3	+	+	+	//	+			
4-14-14	3				//	+			
4-14-20	3				//	+	+		
4-20-20	3	+			//	+			
14-14-20	3				//	+			
14-20-20	3				//	+			
4-4-4-4-14-14	6	+			//	+	+	+	
4-4-4-4-20-20	6	+		//	//	+	+	+	
4-4-14-14-14-14	6	+		//	//	+			
4-4-14-14-20-20	6			//	//	+	+		
4-4-20-20-20-20	6	+	//	//	//	+	+		
14-14-14-14-20-20	6	+		//	//	+	+		
14-14-20-20-20-20	6	+	//	//	//	+			
4-4-4-4-4-4-14-14-14	9	+		//	//	+	+	+	+
4-4-4-4-4-4-20-20-20	9	+	+	//	//	+	+	+	+
4-4-4-14-14-14-14-14-14	9	+		//	//	+	+	+	+
4-4-4-14-14-14-20-20-20	9	+	//	//	//	+	+	+	
4-4-4-20-20-20-20-20-20	9	+	+ //	//	//	+	+	+	+
14-14-14-14-14-14-20-20-20	9	+	//	//	//	+	+	+	+
14-14-14-20-20-20-20-20-20	9	+	//	//	//	+	+		

En todos los casos, cuando se hace referencia a baja o alta potencia y a baja o alta fiabilidad, es en referencia a los valores establecidos como límites aceptables en la sección C.2. Esto es un valor del 95% o superior para la fiabilidad y del 80% o mayor para la potencia estadística.

De esta manera, pueden agruparse los resultados similares en solamente dos tablas, donde no se indica la potencia y fiabilidad alcanzada, sino simplemente si se llegó (o superó) la cota superior propuesta como objetivo (sección C.2).

E.2. Recomendaciones para elegir un modelo de Meta-Análisis en la agregación de experimentos en Ingeniería de Software Empírica

Habiendo agrupado los resultados de las simulaciones obtenidas durante la experimentación, se formulan las recomendaciones a los investigadores en el campo de la Ingeniería de Software Empírica, sobre la elección del modelo de Meta-Análisis que es conveniente escoger para agregar experimentos con el método de Diferencia de Medias Ponderadas, de acuerdo a las características del estudio a desarrollar.

E.2.1. Recomendaciones para elegir un modelo de Meta-Análisis en la agregación de experimentos de igual tamaño

La primera recomendación para utilizar Meta-Análisis como método de investigación en ingeniería de Software Empírica, es para casos en que se trabaja con un tamaño de efecto medio. En este caso, se aconseja utilizar el método de Diferencia de Medias ponderadas bajo el modelo de efecto fijo utilizando entre 14 y 20 sujetos y realizando de 4 a 10 experimentos. Bajo estas circunstancias, puede esperarse que los resultados obtenidos tengan fiabilidad y potencia estadística aceptables.

Para cualquier otro caso, la potencia y fiabilidad no pueden asegurarse simultáneamente. Sin embargo, para casos de tamaño de efecto pequeño a mediano, se recomienda utilizar el modelo de efecto fijo, que presenta fiabilidad y si bien no presenta la potencia deseada, si tiene mayor potencia que el modelo de efectos aleatorios.

Solamente se recomienda (con reservas) utilizar el modelo de efectos aleatorios para tamaños de efecto grande y muy grande, y para el caso en que se agreguen de 8 a 10 experimentos o 6 experimentos con 20 sujetos. Sin embargo, debe aclararse que en estos casos, el modelo no tiene potencia estadística (no solo no alcanza la cota deseada, sino que el valor es realmente bajo).

En todo caso, para las dos últimas recomendaciones, debe tenerse en cuenta dos cosas, primero que la segunda recomendación es para usar un modelo que no alcanza la potencia deseada, pero que de todas formas presenta ciertos valores de potencia que dependiendo de las circunstancias de la experimentación pueden llegar a ser aceptables, en tanto que el tercer método no tiene potencia alguna. En segundo lugar, debe recordarse que una aplicación estadística que no presenta potencia indica una alta probabilidad de cometer un error de Tipo II sin poder detectarse, es decir, el error que proviene de aceptar la hipótesis nula cuando en realidad se verifica la hipótesis alternativa. Esto quiere decir que no se puede asegurar que el tratamiento experimental supera al de control cuando en realidad sí lo hace. En todo caso, nunca se hará una recomendación que incremente el error de Tipo I, ya que éste implica que se acepta erróneamente el tratamiento experimental como superior al de control.

E.2.2. Recomendaciones para elegir un modelo de Meta-Análisis en la agregación de experimentos de distinto tamaño

En este caso, puede observarse que no coincide para ningún modelo fiabilidad y potencia, por lo que la primera recomendación es la de no agregar experimentos de distinto tamaño.

Si de todas formas, deben agregarse experimentos de distinto tamaño, se recomienda utilizar el modelo de efecto fijo para tamaño de efecto pequeño y más de 6 experimentos.

En este caso, también puede utilizarse el modelo de efectos aleatorios si la cantidad de experimentos es superior a 6 y cualquier tamaño de efecto o cualquier cantidad de experimentos para tamaño de efecto pequeño. De todas maneras, se recuerda que siempre que se escoja el modelo de efectos aleatorios para agregar experimentos, éste no presenta nunca potencia estadística.

Finalmente, se recomienda utilizar siempre el modelo de efecto fijo para los casos que presenta fiabilidad, ya que siempre tiene mayor potencia que el modelo de efectos aleatorios, aunque esta elección se deja en manos del investigador.

E.2.3. Resumen de recomendaciones para elegir un modelo de Meta-Análisis en la agregación de experimentos

En la Tabla XX se resumen las recomendaciones antes detalladas, que busca ser una herramienta de apoyo al investigador que se dedica al campo de la Ingeniería de Software Empírica.

VI. Conclusiones

A. Análisis de los resultados obtenidos e interpretación

A.1. Análisis comparativo de simulaciones para análisis de fiabilidad e igual tamaño de experimentos

Según se vio en el capítulo V (Experimentación), el modelo de efecto aleatorio presenta fiabilidad para tamaño de efecto pequeño y mediano, en tanto, para tamaños de efecto alto y muy alto, la fiabilidad disminuye al incrementarse la cantidad de sujetos y experimentos.

Por otro lado, el modelo de efectos aleatorios presenta fiabilidad para un número grande de experimentos. La fiabilidad con este modelo crece al disminuir la cantidad de sujetos por experimentos y el tamaño de efecto.

Ambos modelos presentan mayor fiabilidad para los casos de tamaño de efecto pequeño (con la cantidad de sujetos y experimentos con que se realizó el proceso de simulación), pero en general, para tamaños de efecto pequeño y medio, es mayor la fiabilidad del modelo de efecto fijo.

A.2. Análisis comparativo de simulaciones para análisis de potencia e igual tamaño de experimentos

La potencia para el modelo de efecto fijo se presenta cuando la cantidad de sujetos por experimentos y la cantidad de experimentos a agregar es alta. En general, el modelo de de efecto fijo presenta potencia para 80 sujetos totales (en la combinación de todos los experimentos a agregar) y tamaño de efecto medio, 40 sujetos totales para tamaño de efecto grande y 20 sujetos para tamaño de efecto muy grande. No hay potencia para tamaño de efecto pequeño.

El modelo de efectos aleatorios no alcanza la potencia deseada (80 %) en ningún caso.

A.3. Análisis comparativo de simulaciones para análisis de fiabilidad y distinto tamaño de experimentos

El modelo de efecto fijo presenta fiabilidad para 6 y 9 experimentos (independientemente de la cantidad de sujetos que se combinen en cada experimento) y tamaño de efecto pequeño.

El modelo de efectos aleatorios tiene fiabilidad para tamaño de efecto pequeño y 3 experimentos, tamaño de efecto pequeño y medio y 6 experimentos y siempre que se combinen 9 experimentos.

En este caso, el modelo de efectos aleatorios presenta más fiabilidad que el modelo de efecto fijo.

A.4. Análisis comparativo de simulaciones para análisis de potencia y distinto tamaño de experimentos

En este caso, el modelo de efecto fijo tiene potencia para tamaño de efecto muy alto y 3 experimentos, tamaño de efecto alto y muy alto y 6 experimentos y para tamaño de efecto medio a muy alto y 9 experimentos.

Para el modelo de efectos aleatorios no se observa potencia en ningún caso (por debajo del 80 % en todos los casos).

TABLA XX. Cuadro De Simulaciones Para Ambos Modelos De Efecto De Meta-Análisis, Para Comparación De Potencia Con Desvío Estándar Del 40% (Valores Expresados En Porcentajes)

Características de la experimentación					Recomendación	
Tamaños de los experimentos a agregar	Tamaño de efecto	Desvío estándar	Sujetos experimentales	Cantidad de experimentos	Modelo recomendado	Nota a la recomendación
Experimentos igual tamaño	Medio	Indiferente	>14	>4	FIJO	Potencia y fiabilidad deseada
	Bajo y medio	Indiferente	Indiferente	Indiferente	FIJO	Fiabilidad deseada y potencia baja
	Grande y muy grande	Indiferente	Indiferente	>8	NINGUNO (ALEATORIO)	Fiabilidad deseada y potencia NULA
	Grande y muy grande	Indiferente	20	6	NINGUNO (ALEATORIO)	Fiabilidad deseada y potencia NULA
Experimentos de distinto tamaño	Pequeño	Indiferente	Indiferente	>6	FIJO	Fiabilidad deseada y potencia baja
	Indiferente	Indiferente	Indiferente	>6	NINGUNO (ALEATORIO)	Fiabilidad deseada y potencia NULA
	Pequeño	Indiferente	Indiferente	Indiferente	NINGUNO (ALEATORIO)	Fiabilidad deseada y potencia NULA

B. Conclusiones y recomendaciones

B.1. Conclusiones

La presente investigación ha mostrado la baja potencia del modelo de efectos aleatorios dentro del contexto de simulación desarrollado, esto provoca en la práctica que el resultado final del meta-análisis hecho con este tipo de modelo tienda a dar diferencias no significativas en todo momento, no permitiendo de esta forma poder detectar la comisión de un error de tipo II (afirmar que un tratamiento es mejor que otro cuando en realidad los es).

Por otro lado, la fiabilidad del modelo de efectos aleatorios, es superior a la del modelo de efecto fijo cuando el efecto es alto o muy alto, cuando se combinan un número elevado de sujetos por experimentos y la cantidad de experimentos agregados es igualmente elevada. Esto surge en principio como consecuencia directa de los amplios tamaños de los intervalos de confianza que el método DMP arroja bajo el supuesto que indica utilizar el modelo de efectos aleatorios.

Se puede observar que el modelo de efecto fijo se comporta mejor que el modelo de efectos aleatorios, presentando potencia con más de 80 sujetos/experimentos (para tamaño de efecto medio que es cuando además tiene fiabilidad) cuando el modelo de efecto aleatorio no posee potencia en ninguno de los casos analizados, y presenta fiabilidad (el modelo de efecto fijo) para todos los casos en que la varianza es baja o media. Cuando los efectos poblacionales son altos o muy altos, el modelo de efecto fijo tiende a perder fiabilidad sobre todo cuando se incrementa la cantidad de experimentos y la cantidad de sujetos experimentales. Este hecho se produce por una reducción en el tamaño del intervalo de confianza y por una subestimación del tamaño de efecto por diferencias en los valores del desvío estándar, pero se compensa, en parte, con el aumento de la potencia estadística, lo cual permite a los investigadores asegurar que uno de los tratamientos es mejor que el otro a pesar de que el tamaño de efecto indicado no sea exacto, algo que no sucede con el modelo de efectos aleatorios (debido a la falta de potencia que presenta).

B.2. Recomendaciones

De los análisis anteriores surge en primera instancia, que nunca se aconseja utilizar el modelo de efectos aleatorios, ya que carece de potencia en todos los casos (posibilidad de cometer un error de tipo II sin ser detectado).

Sin embargo, el modelo de efecto fijo no presenta fiabilidad y potencia en todos los casos, y se recomienda utilizar éste modelo bajo las siguientes circunstancias:

i) Combinar siempre experimentos de igual tamaño (ya que para los casos en que se combinan experimentos de distinto tamaño, el modelo presenta fiabilidad cuando pierde la potencia y viceversa)
ii) Para que los resultados de la agregación tenga potencia (>= 80%) y fiabilidad (>= 95%), deben combinarse al menos un total de 80 sujetos para casos de tamaño de efecto medio. En estas circunstancias, el desvío estándar es indiferente.

Para cualquier otro caso, el método DMP para ambos modelos no presenta fiabilidad y potencia simultáneamente, por lo que no existe recomendación alguna para estos casos.

B.3. Futuros trabajos

Según lo estableciera [36], el DMP es la técnica de agregación de experimentos que mejor fiabilidad y potencia presenta en general. En el presente trabajo, además se estableció qué modelo de Meta-Análisis utilizar.

Sin embargo, en [36] se compararon cuatro técnicas de agregación de experimentos, y para los casos en que el DMP pierde fiabilidad, el Response Ratio (RR) la ganaba. De hecho, para casos de muchos sujetos y muchos experimentos, se recomienda utilizar el RR.

En la investigación presente se estudiaron ambos modelos de Meta-Análisis con la técnica DMP, pero no se realizaron simulaciones para analizar ésta otra técnica (RR).

Debido a que ya fue analizado el RR para el modelo de efecto fijo e igual tamaño de experimentos, se plantea como futuras líneas de trabajo:

i) Analizar el RR para el modelo de efecto fijo y distintos tamaños de experimentos.
ii) Analizar el RR para el modelo de efectos aleatorios para distintos tamaños de experimentos.
iii) Analizar el RR para el modelo de efectos aleatorios para igual tamaño de experimentos.

REFERENCIAS

[1] Agarwal, R., Tanniru, M., 1990, Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation, Journal of Management Information System, M.E. Sharpe, Vol. 7 N. 1.

[2] Bailey, J. y Basili, V. 1981: A Meta-Model for Software Development Resource Expenditures, IEEE Press, pp. 107-116

[3] Banker y Keremer, 1989, Scale economies in new software development. IEEE Transactions on Software Engineering. (15): 10, pp. 1199-1205.

[4] Basili, V., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sörumgård, S. y Zelkowitz, M. 1996: The empirical investigation of perspective-based reading, International Journal on Empirical Software Engineering, Vol. 1, No. 2, pp. 133–164.

[5] Basili, V. y Weiss, D. 1981: Evaluation of a Software Requirements Document by Analysis of Change Data, IEEE Press, pp. 314-323

[6] Borenstein, M., Hedges, L. y Rothstein, H. 2007: Meta-Analysis Fixed Effect vs. random effect, www.Meta-Analysis.com

[7] Brassard, G. y Bratley, P. 1988: Algorithmics: Theory and Practice, Prentice-Hall

[8] 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, pp. 136-145.

[9] 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, pp. 167-178.

[10] Cabrero García, L. y Richart Martínez, M. 1996: El debate investigación cualitativa frente a investigación cuantitativa Enfermería clínica, pp. 212-217.

[11] Crandall Klein, B. y Asociados, 1989: A Comparative Study Of Think-Aloud And Critical Decision Knowledge Elicitation Method. SIGAR Newsletter, April 1989, Number 108, Knowledge Acquisition Special Issue, pp. 144-146.

[12] Chalmers, I., Hedges, L. y Cooper, H. 2002: A brief history of research synthesis, Eval Health Prof March, pp. 2–37.

[13] Ciolkowski, M. 2009: What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering, 3rd International Symposium on Empirical Software Engineering and Measurement, pp. 133-144.

[14] Cochran, W. 1954: The combination of estimates from different experiments, Biometrics, pp. 101–129.

[15] Cochrane, 2008: Curso Avanzado de Revisiones Sistemáticas, www.cochrane.es/?q=es/node/198

[16] Cochrane collaboration, 2011: Open learning material, http://www.cochranenet. org/openlearning/html/mod0.htm, disponible al 26 de agosto de 2012.

[17] Cohen, J. 1988: Statistical Power Analysis for the Behavioral Sciences (2nd ed.), ISBN 0-8058-0283-5.

[18] Cooper, H. y Hedges, L. 1994: The Handbook of Research Synthesis, Russell Sage Foundation: New Cork, NY.
[19] 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.
[20] Cruzes, D. y Dybå, T. 2010: Synthesizing evidence in software engineering research, Proceedings of ACM-IEEE International Symposium on Empirical Software Engineering and Measurement.
[21] Daren S. Starnes, Dan Yates, David Moore, 2010. The Practice of Statistics. Editorial Freeman, ISBN 978-142924-559-3
[22] Davies, P. 1999: What is evidence-based education?, British Journal of Educational Studies, pp. 108-121.
[23] Davis, D. y Holt, C. 1992: Experimental Economics, Princeton University Press
[24] Davis, A., Dieste, O., Hickey, A. Juristo, N. y Moreno, A. 2006: Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review, 14th IEEE International Requirements Engineering Conference (RE'06), pp. 179-188
[25] DerSimonian, R. y Laird, N. 1986: Meta-Analysis in clinical trials, Control Clin Trials, pp. 177-88.
[26] Dieste, O. y Griman, A. 2007: Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews, IEEE Press
[27] Dieste, O., y Juristo, N. 2009: Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques, IEEE Transactions on Software Engineering, TSE-2009-03-0052, http://main.grise.upm.es/reme/publicaciones_download.aspx?type=REV&id=64
[28] Dieste, O., Fernández, E., García, R., y Juristo, N, 2010: "Hidden Evidence Behind Useless Replications", 1st RESER
[29] Dieste, O., Fernández, E., García, R. y Juristo, N. 2011. "Comparative analysis of meta-analysis methods: when to use which?" 6th EASE Durham (UK)
[30] Dixon-Woods, M., Agarwal, S., Jones, D., Young, B. y Sutton, A. 2005: Synthesising qualitative and quantitative evidence: a review of possible methods, Journal of Health Services Research and Policy, 10, 1, 45-53B (9)
[31] Dyba, T., Aricholm, E., Sjoberg, D., Hannay, J. y Shull, F. 2007: Are two heads better than one? On the effectiveness of pair programming. IEEE Software, pp. 12-15.
[32] Dyba, T., Kampenes, V. y Sjoberg, D. 2006: A systematic review of statistical power in software engineering experiments, Information and Software Technology, vol. 48, pp. 745-755
[33] El Emam, K. y Laitenberger, O. 2001: Evaluating Capture-Recapture Models with Two Inspectors, IEEE Transaction on Software Engineering, pp. 851-864.
[34] Everitt, B. 2003: The Cambridge Dictionary of Statistics, CUP, ISBN: 0-521-81099-x
[35] Fenton, N. y Pfleeger, S. 1997: Software metrics. A rigurous and practical approachFuente, PWS Publishing Company
[36] Fernández, E. 2013: Proceso de agregación para estudios experimentales en Ingeniería de Software, Tesis Doctoral, Facultad de Informática, UNLP.
[37] Fernández, E., Pollo, F., Amatriain, H., Dieste, O., Pesado, P. y García-Martínez, R. 2009: Aplicabilidad de los Métodos de Síntesis Cuantitativa de Experimentos en Ingeniería de Software. Proceedings XV Congreso Argentino de Ciencias de la Computación. Workshop de Ingeniería de Software, pp. 752-761. ISBN 978-897-24068-4-1.
[38] Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, pp. 507-521.
[39] Friedrich, J., Adhikari, N. y 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
[40] Gambara, H., Botella, J. y Gempp, R. 2002: Empty time and full time. A meta-analysis of age-related changes perceiving time, © 2002 by Fundación Infancia y Aprendizaje, ISSN: 0210-9395
[41] García, R. 2004: Inferencia Estadística y Diseño de Experimentos, eudeba, Buenos Aires Argentina
[42] Gavaghan, D., Moore, A. y McQay, H. 2000: An evaluation of homogeneity tests in meta-analysis in pain using simulations of patient data, Pain, vol. 85, pp. 415-424.
[43] Glass, G. 1976: Primary, secondary, and meta-analysis of research, Educational Researcher, pp. 3-8
[44] Glass, G. 2000: Meta-analysis at 25, http://glass.ed.asu.edu/gene/papers/meta25.html
[45] Good, P. y Hardin, J. 2006: Common Errors in Statistics (and How to Avoid Them), second edition, wiley & Sons, ISBN-13: 978-0-471-79431-8.
[46] Goodman, C. 1996: Literature Searching and Evidence Interpretation for Assessing Health Care Practices, SBU, Stockholm.
[47] Graham, J. y Schafer, J. 1999: On the Performance of Multiple Imputation for Multivariate Data With Small Sample Size, v-29, Sage PublicationsCarpetas
[48] Guerra Romero, L., 1996: La medicina basada en evidencia: un intento de acercar la ciencia al arte de la práctica médica, Plan Nacional sobre el Sida. Ministerio de Sanidad y consumo, Madrid, España, Med Clin (Barc) 1996, 107, pp. 377-382
[49] Gurevitch, J. y Hedges, L. 2001: Meta-analysis: Combining results of independent experiments, Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch), Oxford University Press, Oxford, pp. 347–369.
[50] Hardy, R.J. y Thompson, S.G, 1998: Detecting and describing heterogeneity in meta-analysis.
[51] Hedges, L 1982: Fitting categorical model to effect size from a series of experiments, journal educational statistics, pp. 119-137.
[52] Hedges, L. 1993: Statistical Considerations, Russell Sage Foundation, First edition
[53] Hedges, L., Gurevitch, J. y Curtis, P. 1999: The Meta-Analysis of Response Ratio in Experimental Ecology, The Ecological Society of America
[54] Hedges, L., Gurevitch, J. y Curtis, P. 1999: Meta Analysis http://www.bio.mq.edu.au/pgrad/SIBS/Meta_analysis.PPT
[55] Hedges, L. y Olkin, I. 1985: Statistical methods for meta-analysis. Academic Press
[56] Higgins, J. y Green, S. 2011: Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0, The Cochrane Collaboration
[57] Higgins, J y Thompson, S. 2002: Quantifying heterogeneity in a meta-analysis, Statistics in Medicine, vol. 21, pp. 1539-1558
[58] Hoyle, R. 1999: Preface, v-vii, Sage Publications
[59] Hu, Q., 1997, Evaluating Alternative Software Production Function. IEEE Transactions on Software Engineering. (23): 6, pp. 379-387.
[60] Hunt, M. 1997: How Science takes stock: the story of meta-analysis, Russell Sage Foundation: New York
[61] Hunter, J. y Schmidt, F. 2004: Methods of meta-analysis: correcting error and bias in research findings, Sage Publications
[62] Ioannidis, J., Patsopoulos, N. y Evangelou, E. 2007: Uncertainty in heterogeneity estimates in metaanalyses, BMJ, 335: 914 doi: 10.1136/bmj.39343.408449.80
[63] Johnson, D. W. y Curtis, P. S. 2001: Effects of forest management on soil C and N storage: meta-analysis, Forest Ecology and Management, pp. 227-238
[64] Jones, M., O'Gorman, T., Lemke, J. y Woolson, R. 1989: A Monte Carlo Investigation of Homogeneity Tests of the Odds Ratio under Various Sample Size Configurations, Biometrics, Vol. 45, No. 1
[65] Jørgensen, M. 2004: A Review of Studies on Expert Estimation of Software Development Effort, Journal of Systems and Software, (70): 1-2, pp. 37-60.
[66] Judd, C., Smith, E. y Kidder, L. 1991: Research Methods in Social Relations, Hartcourt Brace Jovanovich College Publishers, Orlando, Florida

[67] Juristo, N. y Moreno, A. 2001: Basics of Software Engineering Experimentation. Boston: Kluwer Academic Publisher.
[68] Juristo, N. y Moreno, A. 2002: Reliable Knowledge for Software Development, IEEE Software, pp. 98-99.
[69] Juristo, N, Moreno, A. y Vegas, S. 2004: Towards building a solid empirical body of knowledge in testing techniques, Acm Sigsoft Software Engineering Notes (Sigsoft), pp. 1-4
[70] Juristo N. y Vegas, S. 2011: The Role of Non-Exact Replications in Software Engineering Experiments, Journal: Empirical Software Engineering
[71] Kampenes, V., Dyba, T., Hannay, J. y Sjøberg, D. 2007: A systematic review of effect size in software engineering experiments, Information and Software Technology 49, pp. 1073–1086
[72] Kim, W., 2000: A Meta-Analysis of Fear Appeals: Implications for Effective Public Health Campaigns
[73] Kitchenham, B. 2004: Procedures for performing systematic reviews. Keele University, TR/SE-0401. Keele University Technical Report.
[74] Knuth, D. E. 1997: The Art of Computer Programming, Addison-Wesley, vol 2, 1997
[75] Laitenberger, O., Atkinson, C., Schlich, M. y El Emam, K. 2000: An experimental comparison of reading techniques for defect detection in UML design documents, J.Syst.Software, 53, 2, pp. 183-204
[76] Laitenberger, O. y Rombach, D. 2003: (Quasi-)Experimental Studies in Industrial Settings, World Scientific
[77] Lajeunesse, M. y Forbes, M. 2003: Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. Ecology Letters, 6, pp. 448-454.
[78] Liang, K. y Self, S. 1985: Tests for Homogeneity of Odds Ratio When the Data are Sparse, Biometrika, Vol. 72, No. 2
[79] Lipsitz, S., Dear, K., Laird, N. y Molenberghs, G. 1998: Tests for homogeneity of the risk difference when data are sparse, Biometrics, vol. 54, pp. 148-160.
[80] Meta-Analysis, 2011: disponible en http://www.meta-analysis.com/
[81] Metropolis, N. y Ulam, S. 1949: The Monte Carlo Method, Journal of the American Statistical Association, 44(247), pp. 335-341.
[82] Miguez, E. y Bollero, G. 2005: Review of Corn Yield Response under winter cover cropping systems using Meta-Analytic Methods, Crop Science Society of America
[83] Miller, J. 1999, Can Results from Software Engineering Experiments be Safely Combined?, IEEE METRICS, pp. 152-158
[84] Miller, J. 2000: Meta-analytical Procedures to Software Engineering Experiments, Journal of Systems and Software, 54, 1, pp. 29-39
[85] Mix, 2011: disponible en http://www.mix-for-meta-analysis.info/
[86] Mohagheghi, P. y 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).
[87] Morales Vallejo, P. 2011: Tamaño necesario de la muestra ¿Cuántos sujetos necesitamos?, www.upcomillas.es/personal/peter/investigacion/TamañoMuestra.pdf
[88] Myers, D. y Lamm, H. 1975: The polarizing effect of group discussion, American Scientist, 63, pp. 297-303Navarro, Giribet y Aguinaga, 1999: Psiquiatría basada en la evidencia: Ventajas y limitaciones, Psiquiatría Biológica, 6, pp. 77-85.
[89] Noortgate, W. y Onghena, P. 2003: Estimating the mean effect size un meta-analysis: Bias, precision, and mean squared error of different weighting methods. Behavioral research methods, instruments and computers, 35, pp. 504-511
[90] Pearson, K. 1904: Report on certain enteric fever inoculation statistics, BMJ 3, pp. 1243-1246.
[91] Petrosino, A., Boruch, R., Soydan, H., Duggan, L. y Sánchez-Meca, J. 2001: Meeting the challenges of Evidence-Based Policy: The Campbell Collaboration, Annals of the American Academy of Political & Social Science, 578, pp. 14-34.
[92] Pfleeger, S. 1999: Albert Einstein and Empirical Software Engineering, Computer, pp. 32-37.
[93] Reichart, C. y Cook, T. 1986: Hacia una superación del enfrentamiento entre los métodos cualitativos y cuantitativos, En: Cook TD, Reichart ChR (ed). Métodos cualitativos y cuantitativos en investigación evaluativa. Madrid: Morata.
[94] Richy, F., Ethgen, O., Bruyere, O., Deceulaer, F. y Reginster, J. 2004: From Sample Size to Effect-Size: Small Study Effect Investigation (SSEi), The Internet Journal of Epidemiology, 1, 2
[95] Rogers, D. 2006: Fifty years of Monte Carlo simulations for medical physics, Physics in Medicine and Biology, 51, pp. R287-R301
[96] Sabaliauskaite, G., Kusumoto, S. & Inoue, K. 2004: Assessing defect detection performance of interacting teams in object-oriented design inspection, Information and Software Technology 46 (2004), pp. 875–886, Available online at: www.sciencedirect.com
[97] Sabaliauskaite, G., Matsukawa, F., Kusumoto, S. & Inoue, K. 2002: An experimental comparison of checklist-based reading and perspective-based reading for UML design document inspection, Empirical Software Engineering, pp. 148-157
[98] Sackett, D. y Wennberg, J. 1997: Choosing the best research design for each question, BMJ, 315:1636
[99] Sanchez-Meca, J. y Botella, J. 2010: Revisiones Sistemáticas y Meta-Análisis: Herramientas Para La práctica Profesional, Papeles del Psicólogo, Vol. 31, Núm. 1, pp. 7-17
[100] Sánchez-Meca, J. y Marín-Martínez, F. 1998: Testing continuous moderators in meta-analysis: A comparison of procedures, British Journal of Mathematical and Statistical Psychology, 51:311–26.
[101] Sawilowsky, S. y Fahoome, G. 2002: Statistics Through Monte Carlo Simulation with Fortran, ed: JMASM.
[102] 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), pp. 245-253.
[103] Schmidt, F. y Hunter, J. 2003: Handbook of Psychology, Research Methods in Psychology, Chapter 21, "Meta-Análisis", Schinka, J., Velicer, W., Weiner, I. Editors, Volume 2.
[104] Shekelle, P., Maglione, M., Morton, S., 2003, Judging What to Do About Ephedra, http://rand.org/publications/randreview/issues/spring2003/evidence.htm
[105] Shull, F., Carver, J., Travassos, G. H., Maldonado, J. C., Conradi, R., and Basili, V. R., 2003: Replicated Studies: Building a Body of Knowledge about Software Reading Techniques. Lecture Notes on Empirical Software Engineering. World Scientific. Chapter 2, pp. 39-84.
[106] Sidhu, D. y Leung, T. 1989: Formal Methods for Protocol Testing: A Detail Study, IEEE Transaction on Software Engineering, 15(4) pp. 413-426.
[107] Sjoberg, D. 2005: A survey of controlled Experiments in Software Engineering, IEEE Transactions on Software Engineering, Vol 31 Nro. 9.
[108] Song, F., Sheldon, T., Sutton, A., Abrams, K. y Jones, D. 2001: Methods for Exploring Heterogeneity in Meta-Analysis, Evaluation and The Health professions, vol. 24 no. 2, pp. 126-151.
[109] Strain, D. y Lee, J. 1984: Variance Component Testing in the Longitudinal Mixed Effects Model, Biometrics, vol. 50, pp. 1171-1177.
[110] Takkouche, B., Cadarso-Suarez, C. y Spiegelman, D. 1999: Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis, Am. J. Epidemiol, PubMed ChemPort,150, pp. 206–215,
[111] Thalheimer, W. y Cook, S. 2002: How to calculate effect sizes from published research: A simplified methodology, A Work-Learning Research Publication.
[112] Tichy, W. 1971: Should computer scientists experiment more?, IEEE Computer, vol. 31, pp. 32-40

[113] Tichy, W. 1998: Should Computer Scientists Experiment More?, IEEE Computer, vol. 31, pp. 32-40
[114] Vander Wiel y Votta, 1993: Assessing Software Design Using Capture-Recapture Methods, IEEE Transaction on Software Engineering, 19(11), pp. 1045-1054.
[115] Weinberg, G. 1971: The Psychology of Computer Programming, Van Nostrand Reinhold, New York
[116] Wohlin, C., Runeson, P., H^st, M., Ohlsson, M. C., Regnell, B. y WesslÈn, A. 2000: Experimentation in Software engineering: An Introduction, International Series in Software Engineering, id: 29, Record: 5370, Volume: 6
[117] Woody, J., Will, R., Blanton, J., 1996, Enhancing Knowledge Elicitation using the Cognitive Interview, Expert system with application, Vol. 10 N. 1
[118] Worn, B, Barbier, E., Beaumont, N., Duffy, J., Folke, C., Halpern, B., Jackson, J., Lotze, H., Micheli, F., Palumbi, S., Sala, E., Selkoe, K., Stachowics, J. y Watson, R. 2007: Supporting Online Material: Impacts of biodiversity loss on ocean ecosystem services.

Hernán Amatriain. Es Profesor Adjunto del Área de Ingeniería de Software en la Licenciatura en Sistemas de la Universidad Nacional de Lanús (UNLa). Es Investigador del Grupo de Investigación en Sistemas de Información (GISI) y dirige el Laboratorio de Investigación y Desarrollo en Ingeniería del Software del GISI del Departamento de Desarrollo Productivo y Tecnológico de la UNLa. Es Ingeniero en Sistemas de Información por la Universidad Tecnológica Nacional y Magister en Ingeniería de Software por la Facultad de Informática de la Universidad Nacional de La Plata.

Eduardo Diez. Es Profesor Asociado del Área de Ingeniería de Software en la Licenciatura en Sistemas de la Universidad Nacional de Lanús (UNLa). Es Investigador del Grupo de Investigación en Sistemas de Información (GISI) y dirige el Laboratorio de Investigación y Desarrollo en Aseguramiento de Calidad de Software del GISI del Departamento de Desarrollo Productivo y Tecnológico de la UNLa. Es Analista de Sistemas y Licenciado en Sistemas por la Universidad de Belgrano. Es Especialista en Ingeniería de Sistemas Expertos y Magister en Ingeniería de Software por el Instituto Tecnológico de Buenos Aires y por la Facultad de Informática de la Universidad Politécnica de Madrid.

Rodolfo Bertone. Es Profesor Titular del área de Bases de Datos. Es Investigador del Instituto de Investigación en Informática (III-LIDI) de la Facultad de Informática de la Universidad Nacional de La Plata (UNLP). Es Licenciado en Informática y Magister en Ingenieria de Software por la Universidad Nacional de La Plata. Es Docente Investigador Categoría II en el Programa de Incentivos del Ministerio de Educación.

Enrique Fernández. Es Investigador Adscripto al Grupo de Investigación en Sistemas de Información de la Licenciatura en Sistemas de la Universidad Nacional de Lanús. Es Docente de la Cátedra de Sistemas de Programación no Convencional de Robots de la Facultad de Ingeniería de la Universidad de Buenos Aires. Es Doctor en Ciencias Informáticas por la Universidad Nacional de La Plata.