risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 14/08/2020
Aceitação/Acceptance: 28/10/2020

Entornos Virtuales para Asistir el Aprendizaje de las Pruebas Software: Meta-Análisis para una Familia de Experimentos

Juan Ucán Pech[1], Omar S. Gómez[2], Julio Díaz Mendoza[1], Raúl Aguilar Vera[1 *]

juan.ucan@correo.uady.mx; ogomez@espoch.edu.ec; julio.diaz@correo.uady.mx; avera@correo.uady.mx

[1] Facultad de Matemáticas, Universidad Autónoma de Yucatán, C.P. 97000, Mérida, Yucatán, México.

[2] Facultad de Informática y Electrónica, Escuela Superior Politécnica de Chimborazo, Riobamba 060155, Ecuador.

*Autor para correspondencia.

Pages: 389-403

Resumen: El artículo reporta los hallazgos encontrados del Meta-Análisis realizado a una familia de experimentos, en dicho análisis se examina la efectividad en la detección de faltas en el código, con y sin apoyo de un Entorno Virtual Colaborativo (EVC). La Familia se integra por un experimento original (e1) y dos replicaciones (r1, r2) hechas en distintos tiempos. Como técnica de síntesis se estimaron los tamaños de efecto individual y global, empleando el enfoque de diferencia de medias ponderada, así como la determinación del grado de heterogeneidad en los experimentos de esta familia. Los resultados sugieren una equivalencia en la efectividad de la detección de defectos con respecto a usar o no usar un EVC, por lo que se sugiere que el uso de un EVC es una alternativa igualmente efectiva frente al modelo tradicional de verificación en colaboración (presencia física entre los miembros de un equipo).

Palabras-clave: Entornos Virtuales de Aprendizaje; Experimentación en Ingeniería de Software; Meta-Análisis; Pruebas de Software.

Intelligent Virtual Environments to Assist Software Testing Learning: Meta-Analysis for a Family of Experiments

Abstract: The article reports the findings of the Meta-Analysis carried out to a family of experiments, in this analysis the effectiveness in detecting errors in the code is examined, with and without the support of a Collaborative Virtual Environment (CVE). The Family is made up of an original experiment (e1) and two replications (r1, r2) made at different times. As a synthesis technique, the individual and global effect sizes were estimated, using the weighted mean difference approach, as well as the determination of the degree of heterogeneity in the experiments of this family.

The results suggest an equivalence in the effectiveness of defect detection with respect to using or not using an EVC, therefore it is suggested that the use of an EVC is an equally effective alternative to the traditional collaborative verification model (physical presence between members of a team).

Keywords: Meta-Analysis; Software Engineering Experimentation; Software Testing; Virtual Learning Environments.

1. Introducción

El estudio sobre el aprendizaje a través de Entornos Virtuales es un área de creciente interés de la Informática Educativa, particularmente para aquellos interesados en la línea del Aprendizaje Colaborativo Asistido por Computadora (Computer Supported Collaborative Learning: CSCL). Los entornos virtuales desarrollados bajo el paradigma del CSCL, se centran en el uso de las tecnologías de la información y de la comunicación, como herramientas de mediación en los métodos de colaboración de la instrucción, y tienen como finalidad, la construcción del conocimiento compartido a través de la interacción social de los miembros del grupo que lo integran (Stahl, Koschmann & Suthers, 2006).

En el ámbito de la Educación en Ingeniería de Software, particularmente en el área de la Programación, existen diversas propuestas de Entornos Virtuales Colaborativos diseñados exprofeso para facilitar su aprendizaje bajo un enfoque colaborativo (Bani-Salameg et al 2010; Esteves et al, 2011; Hupfer et al, 2004).

En el caso de la propuesta de Ucán (2015), de la cual se deriva la familia de experimentos que se analiza en el presente artículo, se estableció como objetivo de investigación, desarrollar un modelo para mejorar el proceso de aprendizaje de tareas vinculadas con la programación, particularmente, la detección de faltas comunes en la programación; así, utilizando el Entorno Virtual Colaborativo (EVC) descrito en (Ucán, 2015), los autores realizaron un conjunto de experimentos con la finalidad de explorar el impacto de utilizar un enfoque colaborativo con apoyo del EVC, en contraste con los esquemas tradicionales para revisión de código.

En el presente artículo se reportan los hallazgos derivados de una síntesis cuantitativa a la información obtenida de la familia de experimentos antes citada; dicho método de meta-análisis es utilizado para combinar de forma cuantitativa los resultados de estudios, replicaciones o experimentos similares, con la finalidad de generar nuevo conocimiento (Hedges & Olkin, 1985).

2. Características del Entorno Virtual Colaborativo

El EVC integra diversas prestaciones de los sistemas de trabajo en grupo (Wilson, 1991) y es complementado con un sistema experto como componente inteligente (Hernández, 2013); dicho entorno tiene como propósito asistir al aprendiz en la detección de faltas. La figura 1 presenta una vista arquitectónica —en tres capas— del EVC utilizado en la familia de experimentos.

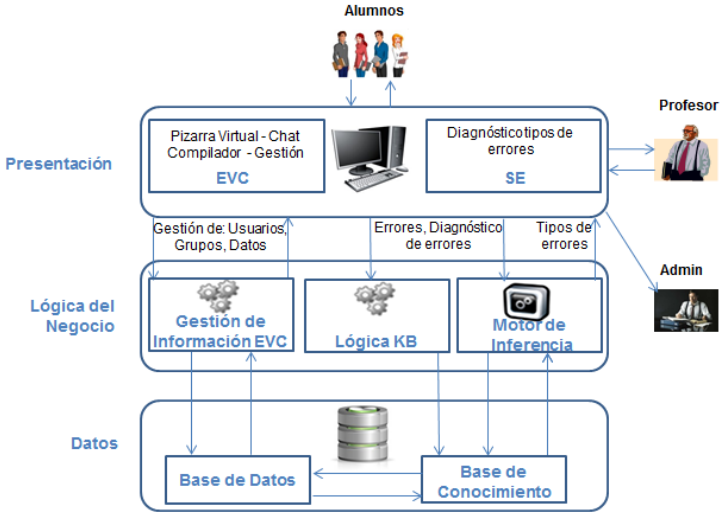

Figura 1 – Arquitectura del Entorno Virtual Colaborativo (Ucán, 2015)

De acuerdo con la capa de presentación, los estudiantes pueden visualizar —en una pizarra virtual— trozos de programas implementados bajo algún paradigma y deben encontrar las faltas incrustadas en dicho código, para dicha actividad de aprendizaje, podrán trabajar en grupos, de manera independiente, consultando al profesor, o incluso con el apoyo de un sistema experto integrado al EVC. Para el modelado del EVC, se utilizó la metodología UWE (*UML Web Engineering*) la cual contempla el modelado del análisis de requisitos, así como los modelos de contenido, navegación, representación y de proceso (Koch et al, 2008).

3. Familia de Experimentos

Los experimentos analizados en el presente reporte fueron realizados siguiendo un proceso de experimentación genérico (Gómez, O., Aguilar, R. y Ucan, J., 2019); dicha familia de experimentos consiste en un experimento original o primer experimento (E1) y dos replicaciones (R1) y (R2).

3.1. Contexto de los Experimentos

El experimento original o primer experimento (E1) se llevó a cabo en la Facultad de Matemáticas de la Universidad Autónoma de Yucatán (UADY) en noviembre de 2014 con estudiantes que cursaban la asignatura Fundamentos de Programación de las Licenciaturas en Ingeniería de Software y Ciencias de la Computación —curso mixto; dicho experimento fue replicado en dos ocasiones (R1 y R2) en la misma institución durante el siguiente año, aunque se tuvo una modificación en el parámetro vinculado con la experiencia de los estudiantes.

La Tabla 1 presenta —para la familia de experimentos— el número de estudiantes participantes, el semestre que cursaban los estudiantes, así como el mes de ejecución del experimento.

Primer Exp. (E1)	Replicación 1 (R1)	Replicación 2 (R2)
46 estudiantes Primer semestre Noviembre de 2014	28 estudiantes Octavo semestre Mayo de 2015	23 estudiantes Sexto semestre Mayo de 2015

Tabla 1 – Familia de Experimentos

Todos los participantes en nuestra familia de experimentos, de acuerdo con la clasificación sobre experiencia en programación creada por Dreyfus y Dreyfus (1986) se pueden clasificar entre estudiantes principiantes y avanzados. Por su parte, la tipología de replicaciones para estos experimentos se acoge a las recomendaciones propuestas por Gómez (2012). En esta familia de experimentos se observa el proceso de aprendizaje de la programación en atención a la detección de faltas en programas —código— de software, en este sentido, las variables fueron las siguientes:

Variable dependiente: detección de faltas comunes de la programación.

- Desde el punto de vista operacional, la variable dependiente utilizó como métrica el porcentaje de faltas observadas en un programa (código).

Variable independiente (Factor): Estrategia de revisión de código

- Tratamiento 1: Sin el uso de una herramienta.
- Tratamiento 2: Uso del Entorno Virtual Colaborativo.

La *hipótesis de nulidad* planteada en esta familia se define como:

H_0. La efectividad medida como el porcentaje de faltas observadas es igual tanto para los grupos de participantes que emplearon el EVC como para los grupos de participantes que trabajaron Sin el EVC.

3.2. Diseño Experimental

La hipótesis de nulidad establecida en la sección previa se probó a través de las mediciones recabadas en la experimentación, dichas mediciones pertenecen a dos grupos de tratamientos, los grupos de sujetos que trabajaron con EVC y aquellos que trabajaron sin el EVC. En esta familia de experimentos se empleó un diseño cruzado con el fin de obtener un mayor número de observaciones (Kuehl, 2000), en un diseño cruzado las unidades experimentales, en este caso los grupos de sujetos conformados por tres personas reciben dos o más tratamientos y el orden de aplicación de los tratamientos está determinado por la estructura de este diseño. En particular, en cada experimento se empleó un diseño cruzado 2 x 2, dos tratamientos — sujetos Con EVC [C] y sujetos Sin EVCI [S]— en dos periodos distintos. Con dicha estructura, la mitad de los grupos de sujetos reciben el tratamiento C y en una sesión diferente reciben el tratamiento S,

mientras que la otra mitad reciben los tratamientos en orden inverso. Un inconveniente de este tipo de diseños es que los efectos de cierto tratamiento puedan extenderse más allá del periodo de aplicación —efecto remanente. Una estrategia para mitigar este inconveniente es no aplicar los tratamientos en periodos consecutivos; para esta familia de experimentos, en E1 se planificó un descanso de 4 días y para R1 y R2 un descanso de 3 días entre ambos periodos.

3.3. Ejecución del Experimento

La experimentación consistió en identificar faltas inyectadas intencionalmente en el código de dos programas de software escritos en un lenguaje de programación (para E1 en Lenguaje C y para R1 y R2 en Java), las faltas se tomaron como referencia de la clasificación propuesta en (Basili & Selby, 1987).

En relación con la asignación de las unidades experimentales a los tratamientos, se organizaron equipos de dos y tres estudiantes y se asignaron de forma aleatoria a alguna de las dos secuencias de tratamientos, tal como se ilustra en la Tabla 2.

	E1		R1		R2	
	C [EVC]	[S] EVC	C	S	C	S
Periodo 1 (Programa A)	8	8	5	4	4	4
Periodo 2 (Programa B)	8	8	4	5	4	4

Tabla 2 – Organización de Sujetos Experimentales en la Familia de Experimentos

En cada experimento de la familia, ambos grupos trabajaron con el mismo código; en el primer periodo unos equipos fueron provistos del empleo del EVC (Tratamiento C) para la identificación de las faltas, mientras que a los otros equipos se les proporcionó el programa de manera impresa, y por lo tanto, no utilizaron el EVC (Tratamiento S); en el segundo periodo los grupos de sujetos recibieron los tratamientos en orden inverso. Durante el experimento a todos los equipos se les proporcionó un documento impreso con las especificaciones del programa bajo análisis, y una plantilla impresa para el registro de la información referente a las faltas detectadas en el programa.

3.4. Análisis Experimental

En esta sección se presenta un análisis de los datos obtenidos en cada experimento, primero se presenta el estudio original con los análisis correspondientes, seguidamente se muestran los análisis de las dos réplicas de igual forma como se hizo en el experimento original.

3.4.1. Análisis del Experimento Base (Original, E1)

En la Tabla 3 se pueden apreciar las métricas obtenidas de la efectividad en la detección de defectos con respecto a los tratamientos, programas y secuencia de los tratamientos

del primer experimento; como se observa, la mayoría de los participantes detectó el mismo número de faltas en ambos tratamientos. En lo que respecta a los programas empleados, en promedio los sujetos lograron observar un mayor número de faltas en el programa B, mientras que los promedios en ambas secuencias parecen no tener diferencias significativas, por lo que pudiera intuirse la ausencia de efectos remanentes en los tratamientos.

		Efectividad
Tratamiento	[C]	43.45%
	[S]	45.33%
Periodo	Prog. A	37.88%
	Prog. B	50.90%
Secuencia	C→S	41.07%
	S→C	47.71%

Tabla 3 – Efectividad respecto a Tratamientos, Programas y Secuencia en E1

Para probar las hipótesis del estudio original, se analizó un modelo que se describe en la ecuación (1).

$$y_{ijk} = \mu + \alpha_i + b_{ij} + \gamma_k + \tau_d + \lambda_c + \varepsilon_{ijk} \quad (1)$$

Donde μ es el promedio general, αi es el efecto de la secuencia, b_{ij} es el efecto aleatorio para cada sujeto con promedio 0 y varianza σb2, γ_k es el efecto del periodo, τ_d es el efecto directo del tratamiento, λ_c es el efecto remanente, y ε_{ijk} es el error aleatorio independiente con promedio 0 y varianza σ^2. En la Tabla 4 se muestran los resultados de la prueba paramétrica del Análisis de Varianza (ANOVA) con respecto a la efectividad medida como el porcentaje de faltas observadas por los grupos de sujetos. Dados los componentes analizados con el ANOVA, se observa una diferencia significativa a un nivel α del 0.05 en el programa. Con respecto al tratamiento y al efecto remanente no se observan diferencias significativas.

Componente	**SC Parcial**	**gl**	**SM**	**F**	**p-value**
Tratamiento	290.22	1	290.22	0.9	0.3517
Periodo (programa)	1546.17	1	1546.17	4.78	0.0373
Efecto remanente	353.02	1	353.02	1.09	0.3051
Residuos	9058.20	28	323.51		

Tabla 4 – Resultados del ANOVA (Efectividad)

La medida de separación entre el tratamiento y el efecto remanente arroja un valor cercano al 30% (29.2893%), esta medida indica el grado de ortogonalidad entre ambos efectos. Un diseño cruzado 2 x 2 como el aquí usado está expuesto a un porcentaje bajo en la medida de separación. Se recomienda emplear este diseño cuando se intuya la ausencia de efectos remanentes. Dado que no se ha observado un efecto remanente significativo, los resultados del ANOVA pueden considerarse confiables. Para considerar válidos los resultados en E1, se deben verificar los supuestos de normalidad, homocedasticidad e independencia de los datos; los primeros dos supuestos fueron validados mediante las pruebas de Shapiro-Wilk y Levene, respectivamente; en el caso del tercer supuesto, se generaron gráficos de residuos vs. secuencia para evaluar visualmente; por restricciones de espacio no serán incluidos.

3.4.2. Análisis de la primera Replicación (R1)

La primera réplica continúa con el análisis de efectividad, eficiencia y costo promedio en la búsqueda de faltas con respecto a los tratamientos, programas y secuencia, sin embargo, esta prueba se realiza con nuevos sujetos. Como se indica en el contexto del experimento, en este caso se emplearon estudiantes avanzados de pregrado. La Tabla 5 presenta las observaciones acerca de la efectividad.

		Efectividad
Tratamiento	[C]	53.70%
	[S]	50.00%
Periodo	Prog. A	61.10%
	Prog. B	42.59%
Secuencia	C→S	54.16%
	S→C	50.00%

Tabla 5 – Efectividad respecto a Tratamientos, Programas y Secuencia en R1

Como se observa en la Tabla 5, los participantes detectaron un porcentaje ligeramente mayor de faltas con el uso del EVC [C] que sin uso del EVC [S]. Por lo que respecta a los programas empleados, en promedio los sujetos lograron observar un mayor número de faltas en el programa A, mientras que los promedios en ambas secuencias parecen no tener diferencias sustanciales por lo que pudiera intuirse la ausencia de efectos remanentes en los tratamientos.

Para probar las hipótesis de R1, se utilizó y analizó el mismo modelo que se describió la ecuación (1). En la Tabla 6 se muestran los resultados del análisis de varianza con respecto a la efectividad medida como el porcentaje de faltas observadas por los grupos de sujetos. Dados los componentes analizados en la ANOVA, se observa una diferencia significativa a un nivel α del 0.05 en el programa. Con respecto al tratamiento y al efecto remanente no se observan diferencias significativas.

Componente	SC Parcial	gl	SM	F	p-value
Tratamiento	222.20	1	222.20	0.98	0.3384
Periodo (programa)	1209.89	1	1209.89	5.35	0.0365*
Efecto remanente	77.15	1	77.15	0.34	0.5685
Residuos	3166.57	14	226.18		

Tabla 6 – Resultados del ANOVA en R1

3.4.3. Análisis de la segunda Replicación (R2)

En forma similar al experimento original y al primer experimento, en la segunda réplica se analizó la variable efectividad en la búsqueda de faltas con respecto a los tratamientos, programas y secuencia.

		Efectividad
Tratamiento	[C]	45.83%
	[S]	56.25%
Periodo	Prog. A	60.41%
	Prog. B	35.41%
Secuencia	C→S	41.66%
	S→C	54.16%

Tabla 7 – Efectividad respecto a Tratamientos, Programas y Secuencia en R2

Como podemos observar en la Tabla 7, los participantes detectaron un porcentaje ligeramente menor de faltas con el uso del EVC [C] que sin uso del EVCI [S]. Similar a R1, en lo que respecta a los programas empleados, en promedio los sujetos lograron observar un mayor número de faltas en el programa A, mientras que los promedios en ambas secuencias parecen no tener diferencias sustanciales por lo que pudiera intuirse la ausencia de efectos remanentes en los tratamientos.

En la Tabla 8 se muestran los resultados del análisis de varianza con respecto a la efectividad. De acuerdo con los componentes analizados, se mantiene la observación de una diferencia significativa a un nivel α del 0.05 en el tratamiento, con respecto al programa y al efecto remanente se continúa sin observar diferencias significativas.

Componente	SC Parcial	gl	SM	F	p-value
Tratamiento	1701.19	1	1701.19	4.82	0.0485*
Periodo (programa)	0.00	1	0.00	0.00	1.0000
Efecto remanente	1406.25	1	1406.25	3.98	0.0691
Residuos	4235.89	12	352.99		

Tabla 8 – Resultados del ANOVA en R2

4. Síntesis Cuantitativa de la Familia de Experimentos

La síntesis cuantitativa es un método de meta-análisis utilizado para combinar de forma cuantitativa los resultados de estudios, replicaciones o experimentos similares para generar nuevas piezas de conocimiento, las cuales serán más generales y fiables en virtud de que se encuentran sustentadas por una mayor cantidad de evidencia empírica (Hedges & Olkin, 1985).

De manera general, el meta-análisis consiste en:

1. estimar el tamaño del efecto de un par de tratamientos por cada uno de los experimentos que se desean examinar,
2. estimar un tamaño de efecto global a partir de los tamaños de efecto estimados individualmente, y
3. estimar el nivel de heterogeneidad del conjunto de tamaños de efecto observados.

Para poder aplicar el método de síntesis cuantitativa es necesario calcular los promedios, tamaños de muestra y desviaciones estándar de los pares de tratamientos de los diferentes estudios que conformarán el meta análisis. La Tabla 9 presenta los promedios, tamaños de muestra y desviaciones estándar calculadas con base en los datos de la familia de experimentos descrita en secciones previas; como se puede apreciar, los tamaños de muestra empleados en los experimentos de esta familia pareciera ser una limitación, no obstante, en el ámbito de la experimentación en ingeniería de software, suele ser una limitación contar con suficientes sujetos en la realización de experimentos, es por ello que de manera paulatina comienzan a desarrollarse familias de experimentos en los que se aplican técnicas de síntesis cuantitativa para minimizar esta limitación (Santos, Gómez & Juristo, 2018).

Estudio	n.evc	avg.evc	sd.evc	n.sevc	avg.sevc	sd.sevc
E1	16	43.45%	20.03	16	45.33%	17.70
R1	9	53.70%	20.03	9	50.00%	14.43
R2	8	45.83%	21.36	8	56.25%	23.47

Tabla 9 – Estadísticos descriptivos de la familia de experimentos

4.1. Estimación de tamaños de efecto individual

El tamaño del efecto es un índice que se utiliza para medir las diferencias entre dos variables, en este caso un par de tratamientos.; el tamaño del efecto permite conocer qué tanto es mejor un tratamiento frente a otro. Para calcular el tamaño del efecto de cada replicación de la partición seleccionada se utiliza la técnica de Glass (1976), conocida como diferencia de medias ponderada WMD (*Weighted Mean Difference*). En la Tabla 10 se muestran los tamaños de efectos calculados en cada estudio, así como el intervalo de confianza en el que se encuentran estos estimados; en nuestro caso, los tamaños de efecto se han calculado entre usar un EVC con respecto a no utilizarlo. Tamaños de efecto negativos indican un tamaño de efecto en favor de no usar un EVC, mientras que tamaños de efecto positivo indican un tamaño de efecto a favor del uso de un EVC.

Estudio	MD	95% - CI
E1	-1.8800	[-14.9774; 11.2174]
R1	3.7000	[-12.4283; 19.8283]
R2	-10.4200	[-32.4106; 11.5706]

Tabla 10 – Tamaños de efecto estimados en los experimentos de la familia

4.2. Estimación del tamaño del efecto general

Una vez estimados los tamaños de los efectos y sus intervalos de confianza, a continuación se combinan estos efectos para obtener un efecto global de los estudios en cuestión. El efecto global es un promedio ponderado del conjunto de tamaños de efecto estimados; a cada tamaño de efecto estimado se le asigna un determinado nivel de ponderación o peso de acuerdo con el nivel de precisión que éste tiene. Los tamaños de efecto que tienen mayor nivel de precisión se les asigna mayor peso, cabe mencionar que la precisión en el tamaño de efecto está en función, principalmente, del tamaño de muestra utilizado en el experimento. A mayor tamaño de muestra los intervalos de confianza del tamaño de efecto estimado se reducen, por lo que el tamaño de efecto estimado se aproxima mucho más al tamaño de efecto verdadero. Por otra parte, a menor tamaño de muestra los intervalos de confianza del tamaño de efecto estimado se amplían, por lo que es más probable que el tamaño de efecto estimado se aproxime menos al tamaño de efecto verdadero. Como se observa en la Tabla 11, E1 ha recibido un grado mayor de ponderación (49.6%); este experimento es el que cuenta con mayor tamaño de muestra, por el contrario, R2 ha recibido la menor ponderación (17.6) y contiene un tamaño de muestra menor, aunque los intervalos de confianza son más anchos.

Estudio	n	MD	95% - CI	%W
E1	16	-1.8800	[-14.9774; 11.2174]	49.6
R1	9	3.7000	[-12.4283; 19.8283]	32.7
R2	8	-10.4200	[-32.4106; 11.5706]	17.6

Tabla 11 – Ponderación y Tamaño de muestra en los experimentos de la familia

Dos de los modelos estadísticos más utilizados para calcular el efecto global de un conjunto de estudios son, el modelo de efectos fijos y el modelo de efectos aleatorios, la elección de un modelo u otro depende del supuesto que se tenga acerca de la distribución de los tamaños de los efectos que provienen del conjunto de replicaciones, en este caso de la partición (Hedges & Olkin, 1985). En el modelo de efectos fijos, se supone que todas las réplicas —de la partición— comparten el mismo tamaño de efecto verdadero. En otras palabras, supone un tamaño de efecto verdadero que es fijo y que subyace a todas las repeticiones, de modo que las diferencias observadas en los tamaños de efecto estimados son producto del error de muestreo y no el producto de la influencia de otros

factores o de las condiciones del experimento. En este modelo, el efecto global es un estimador del tamaño de efecto verdadero que es común en las replicaciones de la partición. Por el contrario, en el modelo de efectos aleatorios se asume que el tamaño de efecto verdadero puede variar de un experimento a otro, es decir, en las replicaciones de la partición pueden existir condiciones que influyen en el tamaño del efecto verdadero. El modelo se asume que los tamaños de efecto verdaderos se distribuyen de acuerdo con la distribución normal, y el efecto global es entonces un estimador del promedio de la distribución de tamaños de efecto verdaderos.

En la práctica, es muy poco probable tener replicaciones exactas de algún experimento porque siempre existirán condiciones o factores que varían entre la(s) replicación(es) y el experimento base (Gómez, 2012). Al haber en las replicaciones cambios tanto deliberados como no deliberados, los tamaños de efecto verdaderos varían entre las replicaciones. Ante a esta situación, el modelo estadístico que utilizamos para calcular el efecto global en nuestra familia de experimentos es el modelo de efectos aleatorios.

Como se observa en la Tabla 12, el tamaño de efecto global estimado se encuentra entre los tamaños de efecto -10.7856 y 7.6715; de acuerdo con el intervalo de confianza estimado, el tamaño de efecto global verdadero puede ser 0, por lo que estos resultados sugieren que la efectividad en la detección de defectos es equivalente con el uso de un ECV o sin el uso de éste.

Estudio	MD (Global)	95% - CI	Z	p-value
Modelo de Efectos Aleatorios	-1.5571	[-10.7856; 7.6715]	-0.33	0.7409

Tabla 12 – Tamaño de efecto global en esta familia de experimentos

Los resultados antes descritos pueden también ser representados, a través del uso de un diagrama de bosque (*forest plot*), como se muestra en la figura 2.

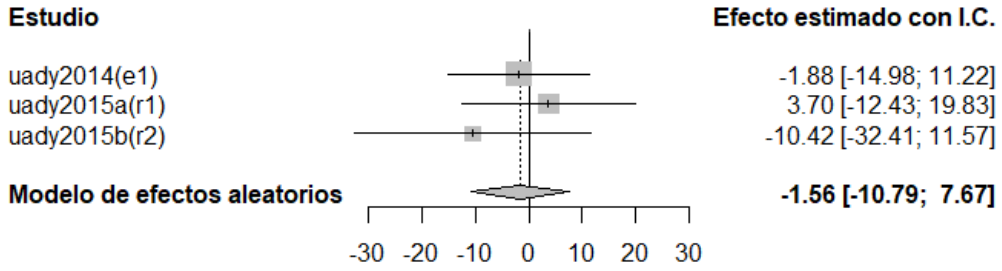

Figura 2 – Diagrama de bosque con los efectos estimados en esta familia de experimentos

4.3. Grado de Heterogeneidad

La prueba de heterogeneidad utiliza el índice Q para evaluar si existe heterogeneidad estadísticamente significativa en un conjunto de tamaños de efecto estimados, esta

prueba parte de la hipótesis nula de la existencia de un tamaño de efecto que es común en los tamaños de efecto de un conjunto de experimentos. Bajo la hipótesis nula Q sigue una distribución χ2 (chi-cuadrada) con k – 1 grados de libertad donde k es el número de experimentos; el no rechazar la hipótesis nula asume que los efectos estimados son homogéneos, es decir, que no hay variabilidad en los resultados.

No obstante, esta prueba estadística tiene una potencia baja para detectar una verdadera heterogeneidad cuando se tiene un conjunto reducido de experimentos (Borenstein et al, 2009). Debido a esta situación se utiliza también el estadístico T2 descrito anteriormente, que consiste en estimar la varianza inter-estudios de un conjunto de experimentos. El otro estadístico que también se utiliza es el índice I2 (Higgins & Thompson, 2002) que se interpreta como el porcentaje de variabilidad total en un conjunto de tamaños de efecto estimados, que se debe a una verdadera heterogeneidad. Los autores proponen tres valores de referencia para este índice, sugieren que valores alrededor de 25%, 50% y 75% representan respectivamente un nivel de heterogeneidad bajo, medio y alto.

Como se observa en la Tabla 13, el valor p resultante no es significativo, sugiriendo una homogeneidad en los resultados de la familia de experimentos. Complementariamente también se han estimado los valores de Tau^2 e I^2 obteniendo 0, y 0.0%, respectivamente. Sugiriendo también una homogeneidad entre los tamaños de efecto de los experimentos de esta familia.

Q	df	p-value
1.03	2	0.5962

Tabla 13 – Prueba de Heterogeneidad

5. Discusión

De acuerdo con los hallazgos aquí reportados, los resultados en los tres experimentos sugieren una equivalencia con respecto a la detección de defectos tanto para los grupos de participantes que trabajaron de forma colaborativa virtual (con el EVC) como para aquellos que trabajaron de manera tradicional. Para este tipo de tarea estudiada —detección de faltas en el código— los resultados arrojados sugieren que el uso de esta herramienta resulta igualmente de efectiva que trabajar de forma colaborativa bajo un enfoque tradicional —colaboración e interacción presencial entre miembros de los equipos.

Al llevar a cabo la síntesis cuantitativa, el efecto global estimado sugiere una efectividad equivalente entre ambos enfoques de colaboración —tamaño de efecto global no significativo— a través de un EVC y de modo tradicional; estos resultados son soportados por la prueba de heterogeneidad la cual sugiere que los tamaños de efecto estimados de manera individual —en los tres experimentos— son consistente entre sí.

En cuanto a las limitaciones del estudio, con respecto a las amenazas de validez interna, en los tres experimentos las sesiones se llevaron a cabo acorde a lo planificado, sin presentar algún incidente —efecto de historia— y dado que se planificó un periodo de descanso

entre la aplicación de tratamientos, se puede asumir un efecto de maduración mínimo. El efecto de sesgo por selección se ha reducido debido al mecanismo de aleatorización empleado en el diseño experimental. Durante las sesiones del experimento se observó interés por parte de los estudiantes en realizar las actividades por lo que también puede asumirse un efecto de desmoralización mínimo. Por otro lado, en lo referente a las amenazas de validez externa los resultados aquí presentados son generalizables a estudiantes con características similares a los estudiantes empleados en el experimento aquí reportado, gracias a las replicaciones efectuadas, el tamaño de la muestra se ha incrementado, así también, la confiabilidad observada en los resultados consistentes entre los experimentos de la familia.

6. Conclusiones

Con base en los resultados del meta-análisis para la familia de experimentos analizada, se puede concluir que existe equivalencia en la efectividad para la detección de faltas en el código, entre los grupos de estudiantes que trabajaron de forma colaborativa virtual (con el EVC), y aquellos que trabajaron de manera tradicional (sin el EVC).

Adicionalmente, algunas de las ventajas observadas tras utilizar el EVC son las siguientes:

1. No se requiere que los estudiantes estén situados en un mismo espacio físico, lo cual, para la educación mixta o para la virtual, la disponibilidad de un EVC resulta apropiada; ni que decir para escenarios como el de la pandemia del COVID-19, en la cual las sesiones presenciales han tenido que migrar a sesiones virtuales.
2. La prestación de una pizarra virtual en el EVC, dicho componente tiene integrado un IDE para escribir y compilar trozos de código de un lenguaje de programación en forma colaborativa, bajo el enfoque tradicional de aprendizaje, dicho escenario es imposible de generar.
3. La prestación del componente inteligente; dicho módulo —sistema experto— ha sido desarrollada para asistir la interacción entre los estudiantes durante actividades de aprendizaje de la programación, así como asistir a los estudiantes durante la identificación de defectos en una serie de programas instrumentados.

Como indican los hallazgos, tener una familia de experimentos permite corroborar los hallazgos anteriores, lo que aumenta la confiabilidad de los resultados. En futuros estudios, tenemos la intención de desarrollar más repeticiones en otros contextos para comprender mejor los límites externos para mantener resultados consistentes con nuevos tipos de réplicas del experimento original.

Agradecimientos

Agradecemos el apoyo brindado por la Secretaría de Educación Pública (México) a través del proyecto P/PROFEXCE-2020-31MSU0098J-13.

Referencias

Bani-Salameh, H., Jeffery, C. & Al-Gharaibeh, J. (2010) A Social Collaborative virtual environment for software development. *Proceedings of the International Symposium on Collaborative Technologies and Systems*. Chicago, IL. DOI: 10.1109/CTS.2010.5478525

Basili, V. & Selby, R. (1987) Comparing the effectiveness of software testing strategies. *IEEE Transactions on Software Engineering*. Vol. SE-13. No. 12. 1278-1296. DOI: 10.1109/TSE.1987.232881

Borenstein, M., Hedges, L. V., Higgins, J. & Rothstein, H. (2009) *Introduction to MetaAnalysis*. John Wiley & Sons, Ltd, United Kingdom.

Dreyfus, H. & Dreyfus, S. (1986) *Mind over Machine. The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Basil Blackwell.

Esteves, M. , Fonseca, B. , Morgado, L. & Martins, P. (2011) Improving teaching and learning of computer programming through the use of the Second Life virtual world. *British Journal of Educational Technology*. https://doi.org/10.1111/j.1467-8535.2010.01056.x

Glass, G. (1976) Primary, secondary, and meta-analysis of research. *Educational Researcher*. Vol. 5(10), 3–8.https://doi.org/10.3102%2F0013189X005010003

Gómez, O. (2012) *Tipología de Replicaciones para la Síntesis de Experimentos en Ingeniería del Software*. (Tesis Doctoral). Universidad Politécnica de Madrid, Madrid: España.

Gómez, O., Aguilar, R. y Ucán, J. (2019) Experimentación en Ingeniería de Software. En Aguilar, R. (Editor). *Ingeniería de Software en México: Educación, Industria e Investigación* (pp. 205-230). CDMX México: Academia Mexicana de Computación.

Hedges, L. & Olkin, I. (1985) *Statistical Methods for Meta-analysis*. Academic Press, Orlando:USA.

Hernández G. (2013) *Análisis del uso de la inteligencia colaborativa como herramienta para la construcción de bases de conocimiento consensuadas en procesos de diagnóstico médico*. (Tesis Doctoral). Universidad Carlos III de Madrid. Madrid: España.

Higgins, J. & Thompson, S. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*. Vol. 21(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hupfer, S., Cheng, L., Ross, S. & Patterson, J. (2004) Introducing Collaboration into an Application Development Environment. *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. https://doi.org/10.1145/1031607.1031611

Koch, N, Knapp A., Zhang, G., and Baumeister, H. (2008) UML-based web engineering. In *Web Engineering: Modelling and Implementing Web Applications*. Springer, London.

Kuehl, R. (2000) *Design of experiments: statistical principles of research design and analysis*. Duxbury/Thomson Learning.

Santos, A., Gómez, O. & Juristo, N. (2018) *Analyzing Families of Experiments in SE: A Systematic Mapping Study*. https://arxiv.org/abs/1805.09009. Online; accessed 22 May 2020.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), Cambridge handbook of the learning sciences (pp. 409-426). Cambridge, UK.

Ucán, J. (2015) *Aprendizaje de la Programación Asistido con Entornos Virtuales Colaborativos Inteligentes*. (Tesis Doctoral). Universidad del Sur, Mérida: México.

Wilson, P. (1991) *Computer supported cooperative work: An introduction*. Springer.