Factores Personales que Influyen en la Realización de Tareas bajo *Test-Driven Development*: Meta-Análisis de una Familia de Experimentos

Geovanny Raura [1], **Efraín R. Fonseca C.** [1], **Oscar Dieste** [2]

[1]Departamento de Ciencias de la Computación – Universidad de las Fuerzas Armadas - ESPE
Quito, Ecuador

[2]Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software – Universidad Politécnica de Madrid
Madrid, España

{jgraura, erfonseca}@espe.edu.ec, odieste@fi.upm.es

Abstract. Background*: Experimentation in Software Engineering often requires experimental subjects to perform some task.* ***Objective****: To study the extent to which experimental tasks are carried out based on an experiment's instrumental aspects and the subjects' personal characteristics.* ***Method****: Meta-analysis of a family of experiments in Test-Driven Development.* ***Results****: Older subjects and those with more professional experience perform less work on the experimental tasks. The nature of the experimental task can influence the level of participation in the experiment.* ***Conclusions****: The motivations of experimental subjects should be studied, and control mechanisms should be established to ensure proper performance.*

Resumen. Antecedentes*: La experimentación en Ingeniería del Software requiere frecuentemente que los sujetos experimentales realicen algún tipo de tarea.* ***Objetivo****: Estudiar el grado en que se realizan las tareas experimentales, en función de los aspectos instrumentales de un experimento y de las características personales de los sujetos.* ***Método****: Meta-análisis de una familia de experimentos en Test-Driven Development.* ***Resultados****: Los sujetos de mayor edad y más experiencia profesional realizan menos trabajo en las tareas experimentales. La tarea experimental puede influir en el grado de participación en el experimento.* ***Conclusiones****: Deben estudiarse las motivaciones de los sujetos experimentales y establecer mecanismos de control que aseguren un correcto desempeño.*

1. Introducción

Las metodologías ágiles han ganado popularidad desde su aparición en la década de los 90, convirtiéndose en un pilar fundamental de las prácticas modernas de desarrollo de software. De entre estas metodologías, el Desarrollo Guiado por Pruebas (TDD, de las siglas del inglés *Test-Driven Development*) [Beck 2003] ha captado significativamente la atención de la academia, la industria y la comunidad científica, debido a su filosofía de mejora, tanto de la calidad del software, como de la productividad de los desarrolladores.

Existen varias revisiones de literatura que sintetizan los hallazgos acerca de TDD ([Rafique and Mišić 2012], [Munir et al. 2014], [Bissi et al. 2016a], entre otras). Un aspecto recurrente en estas investigaciones ha sido la comparación entre TDD y las metodologías tradicionales de programación, utilizando diversas aproximaciones como experimentos controlados [George and Williams 2004], estudios de caso [Nagappan et al. 2008] y encuestas [Bissi et al. 2016b]. Este tipo de estudios ha permitido identificar similitudes y diferencias significativas, enriqueciendo la comprensión sobre las fortalezas y limitaciones de TDD en una variedad de contextos y escenarios.

La literatura también destaca el papel de factores humanos, como la experiencia de los desarrolladores [Fucci et al. 2016], su familiaridad con las herramientas de desarrollo [Madeyski 2010], y el conocimiento de los lenguajes de programación [Pančur and Ciglarič 2011], como elementos influyentes en la aplicación de TDD y en los resultados obtenidos por los programadores. Sin embargo, la relación entre estos factores y los resultados de TDD sigue estando poco explorada, y las conclusiones de los estudios existentes no son concluyentes, lo que resalta la necesidad de investigaciones más profundas y focalizadas [Kollanus 2010].

En esta línea, hemos realizado una familia de experimentos centrados en estudiar factores humanos que influyen en la calidad del código y la productividad de los desarrolladores cuando utilizan TDD. Sin embargo, durante su ejecución, observamos discrepancias entre el comportamiento esperado y real de los sujetos participantes durante la realización de las tareas experimentales. Además de medir ciertas variables respuesta como la calidad y la productividad, revisamos manualmente el código entregado y encontramos que algunos participantes trabajaron considerablemente, mientras que otros apenas realizaron cambios. En casos extremos, los sujetos devolvieron el mismo código proporcionado al inicio.

En un experimento, se espera que los sujetos se comporten profesionalmente, de modo que los efectos observados en las variables respuesta se deriven de los tratamientos ensayados, y no de otros factores. En caso contrario, las conclusiones del experimento sufrirían de una severa amenaza a la validez. Por este motivo, hemos realizado un meta-análisis de nuestra familia de experimentos para averiguar si **el nivel de participación de los sujetos en la realización de las tareas experimentales es aleatorio o puede trazarse a alguna variable**, que en nuestro caso son los aspectos instrumentales de un experimento (tareas, objetos experimentales) y las características personales de los sujetos.

La principal contribución de este trabajo radica en mostrar en qué medida las características personales pueden influir en la ejecución de TDD, lo cual puede tener impacto en la aplicación práctica de esta estrategia de desarrollo, así como implicaciones metodológicas para la investigación empírica en Ingeniería de Software.

El artículo se estructura como sigue. La Sección 2 revisa el estado del arte. En la Sección 3 describimos el diseño experimental, mientras que en la Sección 4 describimos las características principales de las replicaciones experimentales. Los principales resultados del meta-análisis se reportan en la Sección 5. Las Secciones 6 y 7 finalizan el artículo describiendo las amenazas a la validez y las conclusiones, respectivamente.

2. Estado de la Cuestión

En la investigación experimental con humanos, se han realizado diversos estudios que no se centran únicamente en evaluar un determinado tratamiento o factor, sino que también persiguen comprender cómo los sujetos experimentales aplican dicho tratamiento o factor en contextos controlados.

El nivel de participación de los sujetos en las tareas experimentales puede estar influenciado por varios componentes de diseño, como la dificultad de la tarea, los incentivos, la duración del experimento, el diseño de la interfaz y las características del participante. La investigación realizada hasta el momento es limitada, pero algunos resultados son ilustrativos.

La complejidad de una tarea influye significativamente en la probabilidad de que los participantes participen activamente. Cuando una es demasiado compleja, los participantes experimentan sobrecarga cognitiva [Sweller 1988], lo que provoca frustración y abandono. Las tareas que requieren razonamiento abstracto o multitarea pueden abrumar a los participantes [Chandler et al. 2014]. Por el contrario, la baja dificultad puede provocar abandono por aburrimiento.

Los incentivos representan un aspecto muy importante que influye en la realización de las tareas experimentales. Cuando los incentivos son bajos, se produce un mayor abandono. Hauser, David y Schwarz [Hauser and Schwarz 2018] observaron que los participantes pueden abandonar si sienten que la tarea no compensa el tiempo empleado en ella.

La utilización de incentivos monetarios aumenta la retención de los sujetos experimentales. Sin embargo, algunos sujetos pueden realizar la tarea rápido y sin cuidado solo por obtener la recompensa [Oppenheimer et al. 2009]. En general, la motivación intrínseca es la que asegura un grado de compromiso mayor. Cuando los participantes encuentran a la tarea interesante o significativa, participan activamente y la completan con mayor frecuencia [Deci and Ryan 1985].

Los experimentos más largos tienden a presentar tasas de abandono más altas. Los participantes son más propensos a completar tareas de corta duración. En las tareas de duración media (10–30 minutos) la tasa de abandono es moderada. Zhou y Fishbach [Zhou and Fishbach 2016] observaron que, a medida que aumenta la duración de la tarea, la motivación tiende a disminuir, a menos que la tarea se divida en segmentos más cortos y manejables. Los estudios muestran una caída drástica en la retención después de los 30 minutos, especialmente en experimentos en línea [Kittur et al. 2008].

Las diferencias individuales influyen en quién tiene más probabilidades de abandonar la tarea. Las personas con altos niveles de escrupulosidad tienen más probabilidades de completar las tareas que les son asignadas [Diefendorff and Richard 2000]. Las personas con poca paciencia o excesivamente impulsivas abandonan con mayor frecuencia las tareas. En experimentos sobre gratificación diferida, como la famosa *Prueba del Malvavisco* [Mischel et al. 1989], los niños que tenían dificultades con el autocontrol eran más propensos a abandonar las tareas experimentales. De manera similar, los participantes que experimentan estrés o ansiedad tienen un rendimiento inferior, como lo demuestran los estudios sobre ansiedad ante los exámenes y rendimiento cognitivo [Eysenck et al. 2007]. Además, variables demográficas como el nivel educativo y la familiaridad con los proce-

dimientos experimentales afectan la persistencia en la realización de las tareas. Hauser, David y Schwarz [Hauser and Schwarz 2018] encontraron que los participantes con más experiencia en experimentos en línea eran menos propensos a abandonar prematuramente en comparación con los novatos.

El diseño deficiente de la interfaz genera frustración, mientras que las interfaces intuitivas mejoran la retención. Si los participantes tienen dificultades para entender las instrucciones, pueden abandonar prematuramente [Chandler et al. 2014]. Introducir elementos de gamificación (como recompensas o seguimiento del progreso) podrían el abandono [McGonigal 2011].

Los experimentos que generan malestar pueden llevar a un alto abandono. En el bien conocido experimento de obediencia de Milgram [Milgram 1963], algunos participantes se negaron a continuar debido a conflictos morales. En general, los estudios que incluyen dilemas morales reportan altos niveles de abandono [Kogut and Ritov 2005].

En el contexto de los experimentos en Ingeniería de Software, y particularmente en estudios sobre TDD, también se ha comenzado a prestar atención a la influencia que tiene la forma en que los participantes aplican la técnica durante el experimento. Karac et al.[Karac et al. 2025] investigan si esta aplicación práctica impacta los resultados obtenidos en experimentos sobre TDD, mostrando cómo las diferencias individuales pueden afectar los efectos medidos. Por su parte, Santos et al.[Santos et al. 2021] presentan una familia de experimentos centrada en TDD en la que se identifican retos relacionados con la ejecución consistente del tratamiento. Estos trabajos subrayan la importancia de considerar factores personales y contextuales al analizar los resultados de experimentos empíricos en prácticas de desarrollo de software como TDD.

3. Método de Investigación

La investigación en la que se basa este trabajo tenía como objetivo estudiar la influencia de los factores humanos en el desarrollo de software. Para ello, escogimos un experimento base en el área de TDD [Juristo 2016], del cual realizamos siete replicaciones literales nativas tanto internas como externas [Juristo and Gómez 2012] con ciertas diferencias en su contexto o metodología. Cinco replicaciones fueron realizadas en la academia y dos en la industria. A continuación, se presentan los elementos clave del diseño experimental.

3.1. Factores y Niveles

3.1.1. Estrategia de Desarrollo

En este estudio se analizaron dos estrategias de desarrollo: **Desarrollo Guiado por Pruebas** (TDD) y **Desarrollo Iterativo con Pruebas Finales** (*Iterative Test-Last Development*, ITLD por sus siglas en inglés). En uno de los experimentos se introdujo una tercera estrategia denominada **A su manera** (*Your Way, YW por sus siglas en inglés*, pero su utilización no tuvo continuiada. En lo que respecta a TDD, los participantes siguieron el ciclo tradicional:

1. Dividir la tarea principal en sub-tareas pequeñas y manejables.
2. Escribir pruebas mínimas para cada sub-tarea antes de desarrollar el código de producción.

3. Ejecutar las pruebas y validar los resultados para determinar si la implementación es correcta.
4. Refactorizar el código de forma incremental para mejorar su diseño sin alterar su funcionalidad.

En ITLD, aunque se siguen los mismos principios generales, el código de producción se escribe antes que las pruebas, pero en incrementos progresivos. Este enfoque permite una comparación justa con TDD al incorporar pruebas de manera iterativa, evitando problemas asociados con enfoques como *Test Last Development*, donde el código se escribe en su totalidad antes de realizar las pruebas.

3.1.2. Nivel de Descomposición de las Tareas

En varias replicaciones se estudió el **nivel de especificación de las tareas**. Se consideraron dos variantes: **Sliced** (tareas descompuestas) y **No-sliced** (tareas integradas).

En la versión **Sliced**, cada tarea se presenta como una serie de pasos independientes, con especificaciones detalladas del comportamiento esperado y ejemplos de casos de prueba para cada paso. Por ejemplo, si la tarea consiste en implementar una calculadora, los participantes recibirían instrucciones para desarrollar primero la operación de suma con un caso de prueba asociado, luego la resta, y así sucesivamente.

En la versión **No-sliced**, la especificación de la tarea incluyó una descripción general de las funcionalidades requeridas, sin dividirla en pasos. Esto refleja un entorno más cercano a las condiciones reales en las que los requisitos son menos estructurados.

3.2. Tareas Experimentales

Utilizamos dos tareas experimentales:

- **MarsRover API (MR)**. Es un ejercicio de programación que tiene por objetivo el desarrollo de una serie de métodos públicos o API (Aplication Program Interface), que simula el movimiento de un vehículo hacia distintos puntos con diferentes orientaciones (Norte, Sur, Este, Oeste), dentro de un planeta representado por un plano de coordenadas. No se requiere la implementación de una interfaz de usuario.
- **Robert Martin's Bowling Score Keeper (BSK)**. Esta tarea tiene por objetivo calcular el marcador de un único juego de bolos. Igual que MR, esta tarea consiste en la implementación de un algoritmo de programación, sin que tampoco sea necesario el desarrollo de una interface de usuario.

Las tareas experimentales fueron realizadas utilizando el lenguaje de programación Java y JUnit como framework de pruebas. En todos los casos se entregó a los sujetos la estructura básica de código, disponible en [*cita anonimizada*]. Como IDE de desarrollo se utilizó Eclipse.

3.3. Variables de Respuesta y Métricas

El diseño experimental original contemplaba dos variables: la calidad del código y la productividad de los investigadores. El reporte de dichos resultados está en proceso. Sin embargo, al realizar el proceso de medición, pudimos apreciar que independientemente

de los valores obtenidos para las variables calidad y productividad, algunos sujetos entregaron código casi sin haber tocado el esqueleto entregado.

Este comportamiento motivó a que se incluyera dentro del reporte de mediciones, una nueva variable respuesta que se denominó: *nivel de participación en la tarea experimental*, con tres valores posibles (nótese que la variable es ordinal): *Nothing*, cuando el código entregado fue *prácticamente* el mismo código que el descargado de *github*; *Insignificant*, cuando los cambios realizados estaban restringidos a unas pocas líneas de código; y *Aceptable* en los casos restantes.

Es importante destacar que el nivel de participación no tiene el mismo significado que otras variables relacionadas, como el grado de adherencia a un procedimiento experimental (*process conformance* [Fucci et al. 2014]). La adherencia estudia en qué grado siguen los sujetos experimentales las instrucciones de los experimentadores. Sin embargo, el nivel de partición representa (a grano grueso) la cantidad de trabajo realizado por los sujetos, independientemente de si ese trabajo es conforme o no a las instrucciones proporcionadas.

Para hacerse una idea de la importancia del nivel de participación, la Tabla 1 muestra el nivel de participación desglosado por replicación experimental. Los niveles de participación *Nothing* e *Insignificant* superan, por mucho, las participaciones *Adequate*.

Tabla 1. Nivel de participación desglosado por replicación experimental

Nivel de participación	**Replicación**						
	1	2	3	4	5	6	7
Nothing	12	2	0	1	11	6	8
Insignificant	58	43	9	23	12	8	32
Adequate	5	23	8	6	18	2	13

3.4. Diseño Experimental

El diseño experimental sigue un enfoque factorial mixto, donde los factores principales (Estrategia de Desarrollo y Nivel de Descomposición de Tareas) son intra-sujetos. También es intra-sujetos el factor Tarea, aunque su interés es principalmente instrumental (son necesarias para la aplicación de los factores principales) y exploratorio. La variable independiente Grupo es entre-sujetos, aunque no tiene interés en la investigación más allá de aplicar las buenas prácticas indicadas en [Vegas et al. 2016]. Los participantes fueron asignados aleatoriamente a los siguientes grupos experimentales indicados en la Tabla 2.

Existen varias consideraciones importantes respecto al diseño experimental base, que sólo podemos enunciar brevemente en este trabajo (los detalles completos están disponibles en [Raura Ruiz 2022]):

- El diseño mostrado en la Tabla 2 sólo fue utilizado en un experimento. En los restantes, se utilizó únicamente la versión **No-Sliced**, ya que la muestra de sujetos era limitada. De esta manera, se mejoró el poder estadístico de los experimentos individuales.
- Al utilizar únicamente la versión **No-Sliced**, el diseño de la Tabla 2 se simplifica a 2-grupos, 2-periodos, 2-tratamientos. En esta configuración es muy difícil identificar efectos de *carry-over*, por lo que conviene minimizarlos por diseño [Vegas et al. 2016].

Tabla 2. Diseño Experimental

Grupos	**Sesión Experimental 1 - ITLD**	**Sesión Experimental 2 - TDD**
G1	BSK /Sliced	MR/No Sliced
G2	BSK/ No Sliced	MR/Sliced
G3	MR/Sliced	BSK/No Sliced
G4	MR/No Sliced	BSK/Slicing

- El *carry-over* es el efecto que el primer factor intra-sujetos ejerce sobre el segundo factor [Senn 2003, 10 ss.]. En medicina, el *carry-over* tiene su fundamento en la permanencia de la medicación en el organismo durante un cierto tiempo después de finalizar su administración. En IS, el fundamento es menos evidente y podría obedecer a múltiples causas. En esta investigación, el aspecto que más claramente podría producir *carry-over* es la formación, ya que podría ocurrir que los sujetos utilizasen la estrategia TDD en ITLD. Al contrario es más improbable, ya que ITLD sólo añade la prueba automatizada a los conocimientos en programación que ya poseen los sujetos, y dichas pruebas son un prerequisito para TDD. Por este motivo, se decidió aplicar ITLD siempre antes que TDD, aunque ello implique confundir la Estrategia de Desarrollo y la sesión experimental (esto es, el diseño es entre-sujetos pero no *cross-over*).

3.5. Covariables

Estudiamos los efectos de diferentes covariables, esto es, otras variables independientes que podrían influir en el estudio, tales como la edad de los sujetos, la experiencia en desarrollo y el conocimiento previo de TDD, entre otros aspectos indicados en la Tabla 3.

Habitualmente, las covariables no son de interés para la investigación y se consideran únicamente para reducir la variabilidad en la medición de la variable respuesta. Este no es nuestro caso; de hecho, las covariables son un objeto principal de nuestra investigación. La principal característica de estas covariables (y por esta razón no las consideramos como factores) es que no son controlables por los investigadores: en todos los casos, son características propias de los sujetos experimentales. Los valores de estas covariables fueron obtenidos mediante la aplicación de un cuestionario al inicio de cada experimento. Sin embargo, dependiendo de las condiciones de cada experimento, algunas de ellas fueron descartadas posteriormente.

4. Familia de Experimentos

Tal y como se ha mencionado con anterioridad, este trabajo se basa en siete experimentos replicados en diferentes contextos, los cuales se resumen en la Tabla 4. Todos los experimentos están afectados por el fenómeno de abandono de las tareas.

Cada uno de estos experimentos fue analizado independientemente con el propósito de encontrar relaciones entre el nivel de participación y:

Tabla 3. Covariables

Covariable	Descripción	Métrica
Edad	Indica la edad del sujeto participante	Medida mediante un valor numérico: Tipo entero positivo
Nivel de Educación	Determina el nivel de educación del sujeto, el mismo que puede ser estudiante de pre-grado, estudiante de post-grado u otro.	Medida mediante escala nominal: 1 Bachiller en Ciencias de la Computación 2 Master 3 Otro
Experiencia en programación	Indica la experiencia del sujeto en programación	Medida mediante escala ordinal de likert: 1 Sin experiencia (<2 años) 2 Novato (2 - 5 años) 3 Intermedio (6 - 10 años) 4 Experto (>10 años)
Uso de herramientas de pruebas	Indica si el sujeto ha utilizado o no herramientas automáticas de pruebas	Medida mediante escala nominal: 1 Si 2 No
Experiencia en lenguaje Java	Determina si el sujeto tiene experiencia con el uso del lenguaje Java	Medida mediante escala escala ordinal de likert: 1 Sin experiencia (<2 años) 2 Novato (2-5 años) 3 Intermedio (6-10 años) 4 Experto (>10 años)
Experiencia en JUnit	Indica si el sujeto tiene experiencia previa utilizando el framework JUnit	Medida mediante escala escala ordinal de likert: 1 Sin experiencia (<2 años) 2 Novato (2-5 años) 3 Intermedio (6-10 años) 4 Experto (>10 años)
Uso de la técnica TDD	Permite conocer si el sujeto ha tenido conocimiento previo en la técnica TDD	Medida mediante escala nominal: 1 Si 2 No
Experiencia en TDD	Si el sujeto tiene experiencia en la técnica de TDD, este indicador permite determinar los años de experiencia en su uso	Medida mediante escala escala ordinal de likert: 1 Sin experiencia (<2 años) 2 Novato (2-5 años) 3 Intermedio (6-10 años) 4 Experto (>10 años)
Entrenamiento previo en desarrollo de pruebas unitarias	Determina si el sujeto ha recibido entrenamiento previo en el desarrollo de pruebas unitarias	Medida mediante escala nominal: 1 Si 2 No
Conocimiento del entorno Eclipse	Determina si el sujeto tiene conocimiento previo del IDE de desarrollo Eclipse	Medida mediante escala nominal: 1 Si 2 No
Función actual en la organización	Indica cuál es la función que se encontraba desempeñando el sujeto dentro de la Organización	Medida mediante escala nominal: 1 Manager 2 Developer 3 Analyst 4 Other

- Los componentes instrumentales del experimento: niveles y tareas experimentales.
- Los factores personales indicados en la Tabla 3.

Los análisis estadísticos no mostraron, con pocas excepciones, diferencias estadísticamente significativas. La única excepción fue la **experiencia en programación**, la cual mostró una correlación positiva con el nivel de participación en las tareas en algunos experimentos, indicando que los participantes con mayor experiencia tendieron a finalizar más tareas de manera aceptable. No obstante, este resultado no se confirmó en el meta-análisis posterior, como se indicará en la siguiente sección.

La ausencia de resultados significativos no debería extrañarnos. De existir relación entre los componentes instrumentales o factores personales, y el nivel de participación, dicha relación relación debería estar asociada a un tamaño de efecto muy alto para poder ser detectable con el número de sujetos por replicación disponibles. Por este motivo, decidimos aplicar meta-análisis para analizar todas las replicaciones en conjunto.

Tabla 4. Descripción de los experimentos realizados

Rep.	Contexto	# sujetos	Niveles				
			TDD	ITLD	YOUR WAY	SLICED	NO-SLICED
1	Académico (pregrado)	43	X	X		X	X
2	Académico (pregrado)	35	X	X		X	X
3	Académico (postgrado)	9	X	X		X	X
4	Académico (pregrado)	15	X	X		X	X
5	Industrial	23	X	X	X	X	
6	Industrial	10	X	X		X	X
7	Académico (postgrado)	27	X	X		X	X
Sujetos		162					

5. Síntesis de la Familia de Experimentos

El procedimiento que vamos a utilizar para la síntesis es el meta-análisis de datos agregados[1]. Dado que el nivel de participación es ordinal, lo dicotomizaremos en dos categorías: *NothingOrInsignificant* y *Acceptable*. La primera categoría indica que el trabajo realizado por el sujeto experimental no ha sido satisfactorio, y la segunda lo contrario. Existen otras opciones de meta-análisis que no exigen la dicotomización, como los *proportional odds models* y las tablas de contingencia. Evitamos los primeros por su difícil interpretación, y los segundos porque no nos permiten considerar los experimentos por separado ni determinar la homogeneidad de los mismos.

El procedimiento de meta-análisis aplicado el *odds ratio* (OR) o diferencia media (d) utilizando el método genérico de ponderación por inversa de varianza. Para el cálculo de la heterogeneidad hemos usado el procedimiento de DerSimonian y Laird. Dado que este estudio no es confirmatorio y el tamaño muestral es limitado, usamos el nivel de significación $\alpha = 0{,}1$ para rechazar la hipótesis nula, aunque entendemos que los resultados son mucho más fiables a un nivel de significación $\alpha = 0{,}05$.

A este respecto, el p-valor del "*test for overall effect*" no ha sido incluido en los *forest plots* con el fin de evitar sobrecargar las figuras. Sin embargo, su carácter significativo al nivel $\alpha = 0{,}05$ puede inferirse del valor del intervalo de confianza al 95 %. Sí incluye el p-valor correspondiente al análisis de heterogeneidad, ya que este permite determinar el modelo de efectos (fijos o aleatorios) que debe ser interpretado en cada caso.

El meta análisis ha arrojado tres resultados positivos. En primer lugar, el tipo de tarea experimental influye significativamente en nivel de participación. La influencia de la Tarea puede observarse en la Figura 1. En particular, la tarea **Bowling Score Keeper** propocia que los sujetos realicen una mayor cantidad de trabajo en comparación con **Mars Rover**. El efecto observado es estadísticamente significativo ($p - valor = 0{,}06$) al nivel $\alpha = 0{,}1$, con un tamaño de efecto $OR = 3{,}15$, indicando una influencia moderada en el nivel de participación en la tarea de generar el código.

En segundo lugar, la edad de los participantes también juega un papel en el nivel de participación en las tareas. En este análisis, se consideró la edad como variable dependiente, revelando que los **sujetos de mayor edad tienden a entregar tareas que denotan menor participación**. La influencia de la Edad puede observarse en la Figura 2.

[1]Los datos experimentales están disponibles en `https://doi.org/10.5281/zenodo.15075335`.

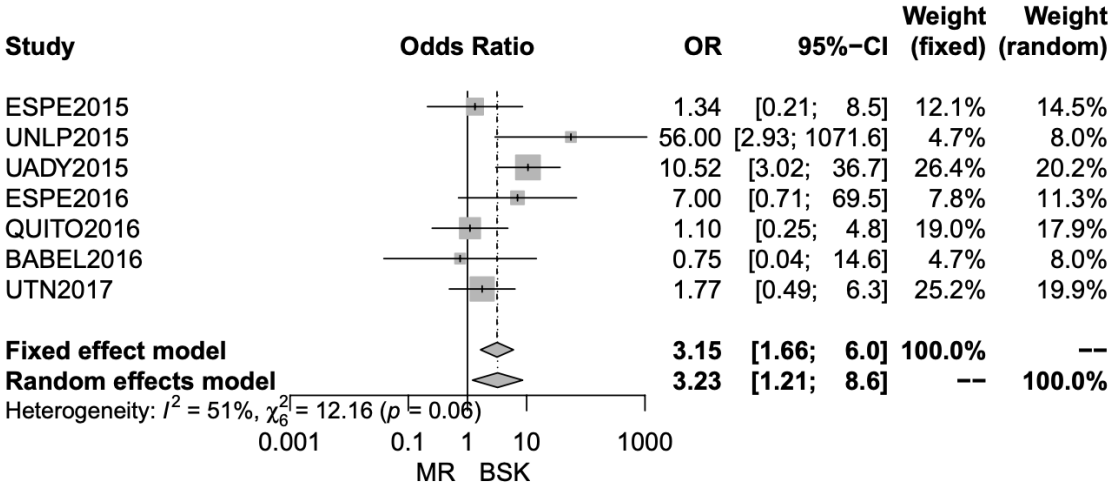

Figura 1. Meta-análisis para el componente instrumental Tarea

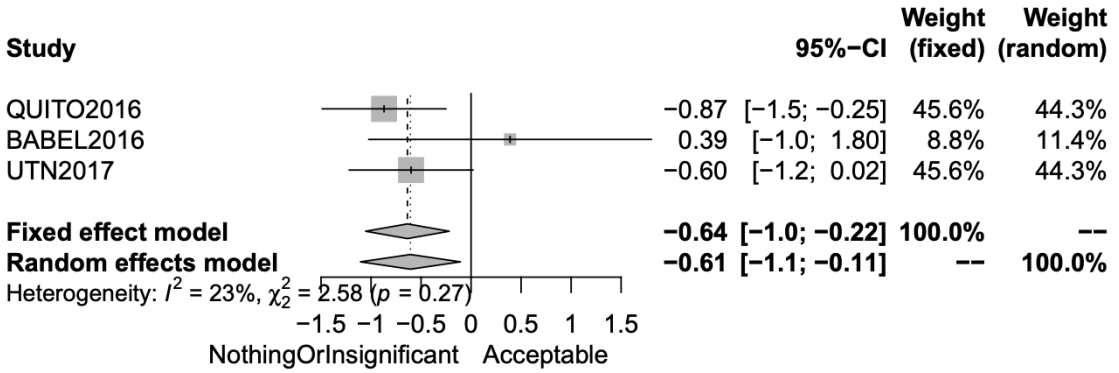

Figura 2. Meta-análisis para el aspecto personal Edad

El tamaño del efecto fue $d = -0{,}64$, lo que representa un efecto medio estadísticamente significativo ($p - value \leq 0{,}0029$). Este hallazgo sugiere que la edad podría estar relacionada con la capacidad o la motivación para completar las tareas asignadas, lo que podría deberse a diferencias en experiencia, hábitos de trabajo o familiaridad con el entorno experimental.

Finalmente, el nivel de educación de los participantes, es una variable ordinal (between-subjects), por lo que hemos realizado un análisis mediante una tabla de contingencia, con un análisis *post-hoc* posterior, en lugar de aplicar un metá-análisis de OR o d. Se obtiene un $p - valor = 0{,}0655$, por debajo del nivel $\alpha = 0{,}1$, lo que lleva a descartar la hipótesis nula, aunque los resultados no son completamente fiables. El análisis *post-hoc* muestra que las diferencias entre grupos no son estadísticamente significativas, pero se observa una tendencia: los participantes con educación de *Master* y *Bachelor* completaron más tareas aceptables (32.5 %), en comparación con el 19.9 % del grupo *Undergraduate*. El grupo *Other* (sin educación universitaria) no completó ninguna tarea de manera aceptable.

El resto de aspectos instrumentales y aspectos personales no han arrojado resultados estadísticamente significativos:

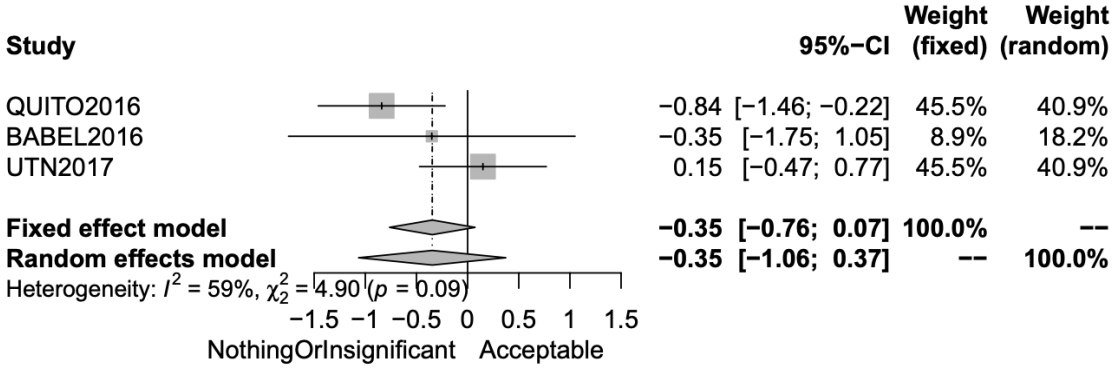

Figura 3. Meta-análisis para la Experiencia Profesional

- Las estrategias TDD e ITLD, esto es, el tipo de estrategia de desarrollo no influencia el nivel de participación.
- El entrenamiento en *Unit Testing*, esto es, el conocimiento previo en pruebas unitarias no afecta al nivel de participación.
- El conocimiento del entorno de desarrollo *Eclipse*.
- La experiencia en programación.
- La experiencia profesional, aunque en este caso el tamaño del efecto $d = -0{,}35$ es pequeño, con un $p - valor \leq 0{,}1055$ no significativo, lo que sugiere que podemos estar ante un error de tipo II. El *forest plot* correspondiente se muestra en la Figura 3.
- El uso de herramientas de prueba.
- La experiencia en *JUnit*.
- La experiencia en el lenguaje de programación Java.

6. Amenazas a la Validez

Existen diversas amenazas a la validez [Shadish et al. 2002] que podrían influir en los resultados obtenidos. Estas amenazas están principalmente relacionadas con el diseño experimental, las condiciones en las que se llevaron a cabo los experimentos y la muestra de participantes utilizada.

La **validez de conclusión** se ve amenaza por dos aspectos principales. En primer lugar, el **tamaño muestral es limitado**, lo que influye en el poder estadístico del análisis. Para paliar esta amenaza, se han realizado varias replicaciones totalizando 162 sujetos, y los resultados de las replicaciones individuales se han combinado mediante meta-análisis. Sin embargo, dado el elevado número de resultados no significativos obtenidos, no podemos descartar la existencia de errores de tipo II, independientemente de que muchos de los efectos implicados serán pequeños, y por lo tanto de difícil detección y no muy interesantes en la práctica.

Otra potencial amenaza a la conclusión es que **sólo se ha considerado el efecto sumativo**, y no interactivo, de las covariables. Podría ocurrir que la edad interaccionase negativamente con TDD pero no con ITLD, y que la experiencia profesional interaccionase inversamente. Sin embargo, estos efectos de segundo orden normalmente son pequeños y difíciles de detectar. Por ello, su estudio debe llevarse a cabo mediante análisis *post-hoc*, que no tienen cabida en esta investigación de tipo experimental.

Una de las principales **amenazas a la validez interna** es el efecto de **práctica**, ya que los participantes pueden mejorar en la tarea a medida que adquieren experiencia con el tiempo. Sin embargo, dado que TDD es una técnica nueva para la mayoría de los sujetos, se espera que la práctica les ayude a comprender mejor sus beneficios en lugar de sesgar los resultados. La **fatiga** es otra amenaza importante, pues las sesiones de entrenamiento y los experimentos se realizaron de manera intensiva, lo que pudo afectar el rendimiento de los participantes, especialmente en tareas repetitivas y de alta concentración.

Otra posible **amenaza a la validez interna** es el **efecto de arrastre** (*carry-over*), ya discutido en la sección 3.4, que ocurre cuando la experiencia adquirida en una condición experimental afecta la siguiente. No podemos descartar que se produzca *carry-over*. Sin embargo, dado que ITLD y TDD comparten estrategias similares, este efecto no se considera problemático e incluso podría haber facilitado la transferencia de conocimientos entre técnicas. Dado que el *carry-over* se confunde con el periodo y grupo experimental [Vegas et al. 2016], además de con la Estrategia de Desarrollo en este diseño concreto, no es posible realizar correcciones durante el análisis que eliminen dicho efecto.

Asimismo, el **orden o periodo de las pruebas** podría ser una fuente de sesgo, pero dado que las sesiones experimentales y de entrenamiento se realizaron secuencialmente, es poco probable que eventos externos hayan influido significativamente en la motivación o habilidades de los participantes.

La medición de la variable respuesta Nivel de Participación se ha realizado de forma subjetiva. Aunque en la medición han participado dos investigadores, es necesario una definición más precisa que evite problemas de **validez de constructo**.

En cuanto a las **amenazas a la validez externa**, la diversidad de los participantes representa una limitación importante. La mayoría de los sujetos experimentales han sido estudiantes, aunque también incorporamos un numero significativo de desarrolladores que trabajan en la industria. Aun así, sería conveniente una muestra más diversa que permitiera generalizar de los resultados.

7. Conclusiones y Trabajos Futuros

La principal contribución de este trabajo es poner de manifiesto que los sujetos experimentales, por diversas razones, pueden realizar de manera inadecuada (lo que denominamos nivel de participación) las tareas experimentales.

Observamos que, conforme los participantes tienen mayor edad, el nivel de participación en las tareas experimentales disminuye. Quizás este efecto podría explicarse por la falta de motivación. En nuestra investigación, los sujetos de mayor edad fueron profesionales que participaban en cursos de entrenamiento y, en ciertos casos, no podían abandonar dichos cursos (ni tampoco los experimentos asociados) dado que debían realizar el entrenamiento dentro de su horario de trabajo.

Otro factor humano que está relacionado con un bajo nivel de participación es la experiencia profesional. Aunque se podría buscar una explicación a dicha relación, lo más probable es que se deba a que está fuertemente correlacionada con la edad (r = 0.66, p-valor <0,001), por lo que exhibe un comportamiento semejante, aunque el meta-análisis, en este caso, no sea estadísticamente significativo al nivel $\alpha = 0{,}05$.

Un factor que se acercó al nivel de significación estadística del 5 % fue el nivel de

educación para el caso de los experimentos realizados con estudiantes. El grupo de participantes de *Bachelor* y *Master* fueron quienes entregaron las tareas experimentales más completas. Esto puede parecer lógico en el sentido de que los estudiantes más formados realizan las tareas *mejor* que los menos formados.

Sin embargo, el nivel de participación no mide dicho efecto, sino el grado en el que los estudiantes de *Bachelor* y *Master* han trabajado durante el experimento, independientemente de si han conseguido realizar con éxito las tareas experimentales. Nuestra interpretación es que los sujetos formados en la universidad están más acostumbrados y probablemente aprecian más los cursos de formación del estilo impartido durante la realización del experimento, y por lo tanto se involucran en mayor medida.

Desde nuestra perspectiva, hemos obtenido ciertos hallazgos interesantes que apuntan en dos direcciones: 1) deben estudiarse las motivaciones de los sujetos experimentales y 2) es conveniente establecer mecanismos de control que aseguren un correcto desempeño de los sujetos durante la realización de experimentos.

Finalmente, queremos indicar que el presente estudio es de carácter exploratorio, dado que no hemos pretendido establecer una relación causa-efecto, sino más bien despertar el interés en la realización de un mayor número de estudios empíricos sobre los factores humanos que podrían incidir en los resultados. En particular, el limitado número de sujetos experimentales que participaron en los siete experimentos hace que nuestros resultados deban tomarse con cautela.

Referencias

Beck, K. (2003). *Test Driven Development: By Example*. Addison-Wesley.

Bissi, W., Neto, A. G. S. S., and Emer, M. C. F. P. (2016a). The effects of test driven development on internal quality, external quality and productivity: A systematic review. *Information and Software Technology*, 74:45–54.

Bissi, W., Scanniello, G., Romano, S., and Tortora, G. (2016b). On the perception of test driven development: A replicated study. *Journal of Visual Languages & Computing*, 34:11–24.

Chandler, J., Mueller, P., and Paolacci, G. (2014). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46(1):112–130.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum, New York.

Diefendorff, J. M. and Richard, E. M. (2000). Antecedents and consequences of emotional display rule perceptions. *Journal of Applied Psychology*, 88(2):284–294.

Eysenck, M. W., Derakshan, N., Santos, R., and Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, 7(2):336–353.

Fucci, D., Erdogmus, H., Turhan, B., Oivo, M., and Juristo, N. (2016). A dissection of the test-driven development process: does it really matter to test-first or to test-last? *IEEE Transactions on Software Engineering*, 43(7):597–614.

Fucci, D., Turhan, B., and Oivo, M. (2014). Impact of process conformance on the effects of test-driven development. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 1–10.

George, B. and Williams, L. (2004). A structured experiment of test-driven development. *Information and Software Technology*, 46(5):337–342.

Hauser, D. J. and Schwarz, N. (2018). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1):400–407.

Juristo, N. (2016). Experiences conducting experiments in industry: The eseil fidipro project. In *2016 IEEE/ACM 4th International Workshop on Conducting Empirical Studies in Industry (CESI)*, pages 1–3.

Juristo, N. and Gómez, O. S. (2012). *Replication of Software Engineering Experiments*, pages 60–88. Springer Berlin Heidelberg, Berlin, Heidelberg.

Karac, I., Panach, J. I., Turhan, B., and Juristo, N. (2025). Does treatment adherence impact experiment results in tdd? *IEEE Transactions on Software Engineering*, 51(1):135–152.

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456.

Kogut, T. and Ritov, I. (2005). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, 18(3):157–167.

Kollanus, S. (2010). Test-driven development: still a promising approach? *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 1–4.

Madeyski, L. (2010). The impact of test-first programming on branch coverage and mutation score indicator of unit tests: An experiment. *Information and Software Technology*, 52(2):169–184.

McGonigal, J. (2011). *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*. Penguin Press, New York.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67(4):371–378.

Mischel, W., Shoda, Y., and Rodriguez, M. L. (1989). Delay of gratification in children. *Science*, 244(4907):933–938.

Munir, H., Moayyed, M., and Petersen, K. (2014). Considering rigor and relevance when evaluating test driven development: A systematic review. *Information and Software Technology*, 56(4):375–394.

Nagappan, N., Maximilien, E. M., Bhat, T., and Williams, L. (2008). Realizing quality improvement through test driven development: Results and experiences of four industrial teams. *Empirical Software Engineering*, 13(3):289–302.

Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872.

Pančur, M. and Ciglarič, M. (2011). Impact of test-driven development on productivity, code and tests: A controlled experiment. *Information and Software Technology*, 53(6):557–573.

Rafique, Y. and Mišić, V. B. (2012). Effects of test driven development on external quality and productivity: A systematic review. *Empirical Software Engineering*, 18(5):1094–1117.

Raura Ruiz, J. G. (2022). *Impacto de las características personales de los programadores en la efectividad de Test-Driven Development (TDD)*. PhD thesis, Universidad Nacional de La Plata, La Plata, Argentina. Tesis de doctorado.

Santos, A., Vegas, S., Dieste, O., and Juristo, N. (2021). A family of experiments on test-driven development. *Empirical Software Engineering*, 26(42).

Senn, S. (2003). *Cross-over Trials in Clinical Research*. Statistics in Practice. Wiley.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Vegas, S., Apa, C., and Juristo, N. (2016). Crossover designs in software engineering experiments: Benefits and perils. *IEEE Transactions on Software Engineering*, 42(2):120–135.

Zhou, X. and Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4):493–504.