POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software

Proceso para la Agregación de Familias de Experimentos en Ingeniería del Software

Tesis Doctoral

Patricia de los Ángeles Riofrío Ojeda

Dirigida por

Dra. Sira Vegas Hernández y Dra. Natalia Juristo

2019

A mis padres y a Darwin.

Agradecimientos

La presente Tesis Doctoral, ha sido posible gracias al trabajo, esfuerzo y apoyo de todas aquellas personas que de alguna manera han formado parte y me han acompañado en este camino que decidí emprender hace cuatro años.

Quiero agradecer a mis directoras Dra. Natalia Juristo y Dra. Sira Vegas por brindarme esta oportunidad. Especialmente gracias a la Dra. Sira Vegas por todo su tiempo, su entrega y los conocimientos compartidos, porque siempre estuvo ahí para ayudarme. No existen palabras para agradecerle todo lo que hizo por mí.

A mis compañeros de laboratorio, por compartir conmigo su experiencia, conocimientos y hacer siempre amenos los momentos compartidos. Inicialmente compañeros, pero ahora amigos para toda la vida.

A mis padres, Patricia y Domingo, y a Darwin por confiar en mí y siempre brindarme palabras de ánimo sobretodo en los momentos más difíciles. Por intentar ayudarme en todo lo que estuvo a su alcance para hacer que este trabajo culmine exitosamente. A toda mi familia porque siempre estuvo orgullosa de mí y me dieron apoyo incondicional.

Finalmente, quiero agradecer profundamente a todas las personas que hicieron posible este trabajo que tanto anhelaba.

Un especial reconocimiento al proyecto TIN2014-60490-P, titulado Realizando experimentos en la industria del Software: Comprensión del paso de laboratorio a la realidad, del Ministerio de Ciencia e Innovación del Gobierno de España.

Resumen

La Ingeniería de Software Empírica (ISE) es el área de la Ingeniería de Software (IS), cuyo objetivo consiste en la obtención de evidencia sobre los métodos, técnicas, procesos y herramientas involucrados en el desarrollo del software. Existen distintos tipos de estudios empíricos entre los que se encuentran: las encuestas, los casos de estudio y los experimentos controlados. Este trabajo se centra en los experimentos controlados (también denominados simplemente experimentos).

En ISE, es bien sabido que la obtención de evidencia fiable requiere de la combinación de resultados de múltiples experimentos. Para combinar los resultados de distintos experimentos es necesario recopilar un conjunto de experimentos relacionados que investiguen la misma hipótesis para posteriormente analizar sus resultados de forma conjunta. Las familias de experimentos son un enfoque para recopilar un conjunto de experimentos relacionados. En las familias de experimentos, los investigadores tienen un conocimiento profundo sobre la configuración de cada experimento, tienen acceso a los datos recogidos en el mismo, y pueden cambiar de forma planificada y controlada la configuración experimental para dar una respuesta global a la hipótesis bajo investigación. Esta tesis se centra en cómo analizar familias de experimentos.

Un proceso para analizar familias de experimentos debe contemplar: 1. La identificación precisa de todas las similitudes y diferencias entre los experimentos. 2. El establecimiento de manera sistemática de un conjunto de variables (de diseño y moderadoras) que podrían estar afectando a los resultados. 3. El análisis de los datos de la familia, teniendo en cuenta de forma simultánea todas y cada una de las variables identificadas. En un estudio de mapeo sistemático sobre cómo se analizan las familias de experimentos en IS se ha descubierto que actualmente no existe un proceso que cumpla estas características.

El objetivo principal de esta tesis consiste en proponer un proceso para el análisis de familias de experimentos que resuelva las carencias encontradas. Para evaluar el proceso propuesto, se analizan dos familias de experimentos. La primera de ellas consta de 4 experimentos y estudia la efectividad relativa de distintas técnicas de evaluación de código. La segunda consta de 16 experimentos y estudia la calidad del código desarrollado utilizando desarrollo guiado por pruebas de software (TDD), frente al código desarrollado utilizando un enfoque de desarrollo tradicional.

Asimismo, el estudio de mapeo sistemático arroja que la variable moderadora más estudiada en las familias de experimentos en IS es el tipo de sujeto. Sin embargo, trabajos actuales sugieren que esta variable no es la más adecuada para explicar diferencias entre los participantes. En este trabajo, se estudia la evidencia empírica actual en relación con la idoneidad del uso de la etiqueta estudiante/profesional para reflejar las diferencias entre ambos colectivos. Para ello, se realiza un nuevo mapeo sistemático sobre experimentos que utilizan ambos tipos de sujetos.

Abstract

Empirical Software Engineering (ESE) is the area of Software Engineering (SE), that aims at obtaining evidence on the methods, techniques, processes and tools involved in software development. There are different types of empirical studies. The most common ones are: surveys, case studies and controlled experiments. This thesis focuses on controlled experiments (also called simply experiments).

It is well known that obtaining reliable evidence in ESE requires the combination of results from multiple experiments. In order to combine that results of several experiments it is necessary to compile a set of experiments that investigate the same hypothesis, to later analyze their results jointly. Families of experiments are an approach to collect a set of related experiments. In families of experiments, the researchers have in-depth knowledge of the configuration of each experiment, have access to the data gathered and can change the experimental configuration in a planned and controlled way to give a global answer to the hypothesis under investigation. This thesis focuses on how to analyse families of experiments.

A process to analyse families of experiments should encompass: 1. A precise identification of all similarities and differences between experiments. 2. Systematically determine a set of (design and moderator) variables that could be affecting the results. 3. Analyse data of the family, taking into account simultaneously all variables identified. In a systematic mapping study on how families of experiments in SE are analyzed, this thesis shows that there is currently no process with these characteristics.

The main objective of this thesis is to propose a process for the analysis of families of experiments that addresses shortcomings found in the state of the practice/art. To evaluate the proposed process, two families of experiments are analysed. The first one consists of 4 experiments and studies the relative effectiveness of different code evaluation techniques. The second consists of 16 experiments and studies the quality of the code developed using Test-Driven Development (TDD), compared to the code developed using a traditional development approach.

Likewise, the systematic mapping study shows that the most studied moderator variable in the families of experiments reported in SE is the subject type. However, current research suggests that this variable is not the most appropriate one to explain differences between subjects. This work studies current empirical evidence on the suitability of the use of the student / professional label to reflect the differences between both groups of subjects. For this, a new systematic mapping review is performed about experiments that use both subject types.

Índice General

Capítulo 1	Introducción	1
1.1	Área de Investigación	1
1.2	Problema	1
1.3	Aproximación a la Solución.....	3
1.4	Publicaciones Derivadas	4
1.5	Organización del Documento	4
Capítulo 2	Estado de la Cuestión	7
2.1	Estado de la Práctica	7
2.2	Trabajos Relacionados.....	27
Capítulo 3	Planteamiento del Problema.....	31
3.1	Descripción del Problema.....	31
3.2	Hipótesis y Objetivos de la Investigación	32
3.3	Metodología de Trabajo	35
Capítulo 4	Propuesta de Proceso para la Agregación de Familias de Experimentos	37
4.1	Visión General del Proceso Propuesto	37
4.2	Identificación de los Grupos de Agregación	40
4.3	Identificación de las Variables de Diseño	53
4.4	Identificación de las Variables Contextuales	63
4.5	Identificación de las Variables de Participantes	68
4.6	Realización del Análisis.....	70
4.7	Resumen del Proceso Propuesto	77
Capítulo 5	Estudio de la Viabilidad de la Propuesta: Agregación de una Familia de Experimentos de Técnicas de Prueba	79
5.1	Descripción de la Familia de Experimentos	79
5.2	Aplicación del Proceso Propuesto a la Familia de Experimentos	85
5.3	Discusión sobre el proceso	104
Capítulo 6	Estudio de la Viabilidad de la Propuesta: Agregación de una Familia de Experimentos sobre Desarrollo Guiado Por Pruebas (TDD).....	105
6.1	Descripción de la Familia de Experimentos	105
6.2	Analizando la Familia de Experimentos.....	110
Capítulo 7	Impacto del Uso de Estudiantes en Experimentos de IS	145
7.1	Introducción.....	145
7.2	Método de Investigación.....	147
7.3	RQ ₁ : Escala de Uso de Estudiantes y Profesionales	149
7.4	RQ ₂ : Caracterización de los Participantes	150
7.5	RQ ₃ : Incorporación del Tipo de Participante / Características en el Experimento	152
7.6	RQ ₄ : ¿Cuál es el impacto del uso de características / tipos de participantes en los resultados experimentales?	153
Capítulo 8	Conclusiones y Futuras Líneas de Investigación.....	167
8.1	Conclusiones.....	167
8.2	Contextualización del Proceso Propuesto	170
8.3	Líneas Futuras	171
Apéndices.....		173
Apéndice A	Diferencias entre experimentos (familias del estado de la cuestión)	173
Apéndice B	Aplicación web para familias de Experimentos.....	177
Apéndice C	Caracterización de los Experimentos.....	181

Apéndice D	Identificación de Variables de Diseño.....	203
Apéndice E	Resultados de la Agregación	225
Bibliografía		231

Índice Detallado

Capítulo 1	Introducción	1
1.1	Área de Investigación	1
1.2	Problema	1
1.3	Aproximación a la Solución	3
1.4	Publicaciones Derivadas	4
1.5	Organización del Documento	4
Capítulo 2	Estado de la Cuestión	7
2.1	Estado de la Práctica	7
2.1.1	Proceso de Búsqueda	7
2.1.2	Proceso de Selección	8
2.1.3	Extracción de Datos	10
2.1.4	P ₁ : ¿Existen Diferencias Entre los Experimentos que Componen una Familia?	12
2.1.5	P ₂ : ¿Cómo se identifican las diferencias entre los experimentos de una familia?	14
2.1.5.1	Caracterización de las replicaciones	14
2.1.5.2	Identificación de Diferencias	15
2.1.6	P ₃ : ¿Cómo se Identifican las Variables a Incluir en el Análisis en una Familia?	17
2.1.6.1	Cómo se Identifican las Variables a Incluir en el Análisis	17
2.1.6.2	Qué Variables se Identifican	18
2.1.7	P ₄ : ¿Cómo se Analizan Conjuntamente los Experimentos de una Familia?	19
2.1.7.1	Técnica de Agregación	19
2.1.7.2	Técnica de Análisis utilizadas	21
2.1.7.3	Estrategia de Análisis	22
2.1.8	Conclusiones generales	24
2.2	Trabajos Relacionados	27
2.2.1	Agregación Cuantitativa de Evidencia de Experimentos Controlados	28
2.2.2	Agregación de Familias de Experimentos	28
2.2.3	Resumen	29
Capítulo 3	Planteamiento del Problema	31
3.1	Descripción del Problema	31
3.2	Hipótesis y Objetivos de la Investigación	32
3.2.1	Creación de un Modelo de Caracterización de Experimentos	33
3.2.2	Identificación de un Conjunto de Variables de Análisis	33
3.2.3	Análisis Conjunto de los Experimentos	34
3.2.4	Estudio del impacto del tipo de sujeto	34
3.3	Metodología de Trabajo	35
Capítulo 4	Propuesta de Proceso para la Agregación de Familias de Experimentos	37
4.1	Visión General del Proceso Propuesto	37
4.2	Identificación de los Grupos de Agregación	40
4.2.1	Caracterización de los Experimentos de la Familia	40
4.2.1.1	Esquema de Caracterización Utilizado	40
4.2.1.2	Caracterización de una Familia de Experimentos	48
4.2.1.3	Herramienta para Recoger la Información de Familias de Experimentos	49
4.2.2	Identificación de los Grupos de Experimentos a Agregar	51
4.3	Identificación de las Variables de Diseño	53
4.3.1	Creación de un Diseño Agregado	53
4.3.2	Exploración de los Resultados de los Análisis Individuales	58

4.3.3	Identificación del Conjunto de Variables de Diseño	62
4.4	Identificación de las Variables Contextuales	63
4.4.1	Identificación de las Diferencias Contextuales	63
4.4.2	Creación de un Conjunto Inicial de Variables Contextuales	66
4.4.3	Refinamiento del Conjunto Inicial de Variables Contextuales	67
4.5	Identificación de las Variables de Participantes	68
4.5.1	Identificación de las Diferencias entre los Participantes	68
4.5.2	Creación de un Conjunto Inicial de Variables de Participantes	69
4.5.3	Refinamiento del Conjunto Inicial de Variables de Participantes	69
4.6	Realización del Análisis	70
4.6.1	Identificación del Modelo de Análisis	72
4.6.2	Análisis del Modelo	74
4.6.3	Interpretación de los Resultados	76
4.6.4	Interpretación Conjunta de los Resultados de los Distintos Grupos	77
4.7	Resumen del Proceso Propuesto	77
Capítulo 5 Estudio de la Viabilidad de la Propuesta: Agregación de una Familia de Experimentos de Técnicas de Prueba		79
5.1	Descripción de la Familia de Experimentos	79
5.1.1	Técnicas Utilizadas	80
5.1.2	Programas	80
5.1.3	Faltas	81
5.1.4	Variable respuesta	82
5.1.5	Participantes	82
5.1.6	Procedimiento experimental	82
5.1.7	Cambios entre los experimentos	84
5.1.7.1	UPM-2	84
5.1.7.2	UPV	84
5.1.7.3	ORT	84
5.2	Aplicación del Proceso Propuesto a la Familia de Experimentos	85
5.2.1	Identificación de los Grupos de Agregación	85
5.2.1.1	Caracterización de los Experimentos de la Familia	85
5.2.1.2	Identificación de los Grupos de Experimentos a Agregar	86
5.2.2	Identificación de las Variables de Diseño	88
5.2.2.1	Creación de un Diseño Agregado	88
5.2.2.2	Exploración de los Resultados de los Análisis Individuales	90
5.2.2.3	Identificación del Conjunto de Variables de Diseño	91
5.2.3	Identificación de las Variables Contextuales	92
5.2.3.1	Identificación de Diferencias Contextuales	92
5.2.3.2	Creación de un Conjunto Inicial de Variables Contextuales	95
5.2.3.3	Refinamiento del Conjunto Inicial de Variables Contextuales	95
5.2.4	Identificación de las Variables de Participantes	96
5.2.4.1	Identificación de las Diferencias entre los Participantes	96
5.2.4.2	Creación de un Conjunto Inicial de las Variables de Participantes	97
5.2.4.3	Refinamiento del Conjunto Inicial de las Variables de Participantes	97
5.2.5	Realización del Análisis	98
5.2.5.1	Identificación del Modelo de Análisis	99
5.2.5.2	Análisis del Modelo	100
5.2.5.3	Interpretación de los Resultados	102
5.3	Discusión sobre el proceso	104
Capítulo 6 Estudio de la Viabilidad de la Propuesta: Agregación de una Familia de Experimentos sobre Desarrollo Guiado Por Pruebas (TDD)		105
6.1	Descripción de la Familia de Experimentos	105

6.1.1 Factores y Tratamientos.....	106
6.1.1.1 Método de Desarrollo.....	106
6.1.1.2 Granularidad.....	107
6.1.1.3 Efecto de Aprendizaje.....	107
6.1.1.4 Realización del Trabajo.....	107
6.1.1.5 Objetos experimentales.....	107
6.1.2 Variable respuesta y Medición.....	109
6.2 Analizando la Familia de Experimentos.....	110
6.2.1 Identificación de los Grupos de Agregación.....	110
6.2.1.1 Caracterización de los Experimentos de la Familia.....	110
6.2.1.2 Identificación de los Grupos de Experimentos a Agregar.....	112
6.2.2 Identificación de las Variables de Diseño.....	113
6.2.2.1 Creación de un Diseño Agregado.....	113
6.2.2.2 Exploración de los Resultados de los Experimentos Individuales.....	119
6.2.2.3 Identificación del Conjunto de Variables de Diseño.....	121
6.2.3 Identificación de las Variables Contextuales.....	122
6.2.3.1 Identificación de Diferencias Contextuales.....	122
6.2.3.2 Identificación de Variables Contextuales.....	125
6.2.3.3 Refinamiento del Conjunto Inicial de Variables Contextuales.....	126
6.2.4 Identificación de las Variables de Participantes.....	126
6.2.4.1 Identificación de las Diferencias entre los Participantes.....	126
6.2.4.2 Creación de un Conjunto Inicial de las Variables de los Participantes.....	128
6.2.4.3 Refinar el Conjunto Inicial de las Variables de los Participantes.....	129
6.2.5 Conducir Análisis.....	131
6.2.5.1 Identificación del Modelo de Análisis.....	131
6.2.5.2 Análisis del Modelo.....	135
6.2.5.3 Interpretación de los Resultados.....	139
Capítulo 7 Impacto del Uso de Estudiantes en Experimentos de IS.....	145
7.1 Introducción.....	145
7.2 Método de Investigación.....	147
7.2.1 Proceso de Búsqueda.....	147
7.2.2 Proceso de Selección.....	148
7.2.3 Extracción de Datos.....	149
7.2.4 Análisis.....	149
7.3 RQ ₁ : Escala de Uso de Estudiantes y Profesionales.....	149
7.4 RQ ₂ : Caracterización de los Participantes.....	150
7.5 RQ ₃ : Incorporación del Tipo de Participante / Características en el Experimento.....	152
7.6 RQ ₄ : ¿Cuál es el impacto del uso de características / tipos de participantes en los resultados experimentales?.....	153
7.6.1 Impacto del Tipo de Participantes al Analizarlos por Separado.....	153
7.6.2 Impacto del Tipo de Participantes al Analizarlos Conjuntamente.....	157
7.6.3 Impacto de las Características de los Participantes en los Resultados.....	161
7.6.4 Discusión.....	162
Capítulo 8 Conclusiones y Futuras Líneas de Investigación.....	167
8.1 Conclusiones.....	167
8.2 Contextualización del Proceso Propuesto.....	170
8.3 Líneas Futuras.....	171
Apéndices.....	173
Apéndice A Diferencias entre experimentos (familias del estado de la cuestión).....	173
A.1 Operacionalización.....	173
A.2 Ejecución experimental.....	173
A.3 Instrumentación.....	174

A.4	Población	175
A.5	Experimentador	175
Apéndice B	Aplicación web para familias de Experimentos.....	177
B.1	Listado de usuarios	177
B.2	Creación de nuevo usuario.....	177
B.3	Listado de familias de experimentos	178
B.4	Listado de experimentos de una familia	178
B.5	Detalle de la operacionalización de un experimento	179
B.6	Detalle de la Población de un experimento	179
B.7	Edición de un experimento (Operacionalización).....	180
Apéndice C	Caracterización de los Experimentos.....	181
C.1	Experimentos de Técnicas de Prueba	181
C.2	Experimentos de Test Driven Development (TDD)	186
Apéndice D	Identificación de Variables de Diseño.....	203
D.1	Experimentos de Técnicas de Prueba	203
D.2	Experimentos de Test Driven Development (TDD)	208
Apéndice E	Resultados de la Agregación	225
E.1	Experimentos de Técnicas de Prueba	225
E.2	Experimentos de Test Driven Development (TDD)	228
Bibliografía.....		231
	Lista de Estudios Primarios (Familias de Experimentos)	233
	Lista de Estudios Primarios (Impacto de los Estudiantes)	235

Índice de Tablas

Tabla 2.1 Número de publicaciones obtenidas durante el proceso de búsqueda	8
Tabla 2.2 Matriz de distancia para la familia [EP14]	12
Tabla 2.3 Grupos de variación: todos los elementos de la configuración experimental	13
Tabla 2.4 Cambios para cada uno de los elementos de la configuración experimental	13
Tabla 2.5 Instrumentos utilizados para la caracterización de familias de experimentos	14
Tabla 2.6 Siguiendo guías y recomendaciones de otros autores	14
Tabla 2.7 Identificación de diferencias entre experimentos	15
Tabla 2.8 Método para la identificación de variables moderadoras	17
Tabla 2.9 Identificación de variables moderadoras	18
Tabla 2.10 Técnica de agregación utilizada en cada familia	20
Tabla 2.11 Técnica de análisis estadístico (IPD e IPD-E)	23
Tabla 2.12 Identificación de variables moderadoras	24
Tabla 2.13 Resumen de los valores para las preguntas P_2 , P_3 y P_4 (1/2)	25
Tabla 2.14 Resumen de los valores para las preguntas P_2 , P_3 y P_4 (2/2)	26
Tabla 2.15 Resultados para el estado del arte	29
Tabla 4.1 Propuesta de esquema de caracterización para replicaciones (Gómez, 2012)	41
Tabla 4.2 Diferencias con la estructura inicial	42
Tabla 4.3 Plantilla para la caracterización	44
Tabla 4.4 Caracterización familia [EP14] (1/2)	48
Tabla 4.5 Caracterización familia [EP14] (2/2)	49
Tabla 4.6 Elementos para la identificación de grupos de experimentos	51
Tabla 4.7 Identificación de grupos de experimentos, familia [EP14]	52
Tabla 4.8 Elementos para la identificación de variables de experimento	53
Tabla 4.9 Ejemplo de un diseño Entre-sujetos	54
Tabla 4.10 Ejemplo de un diseño Intra-sujetos sin cruzar tareas	54
Tabla 4.11 Ejemplo de un diseño Intra-sujetos cruzando tareas	54
Tabla 4.12 Ejemplo de un diseño Crossover	54
Tabla 4.13 Ejemplo de un diseño de Un factor	55
Tabla 4.14 Ejemplo de un diseño de Un factor con variables de bloque	55
Tabla 4.15 Ejemplo de un diseño Factorial	55
Tabla 4.16 Diseño de los experimentos de la familia [EP14]	58
Tabla 4.17 Ejemplo de los resultados de los análisis individuales	60
Tabla 4.18 Ejemplo de obtención de significancia práctica y/o estadística de todos los experimentos	60
Tabla 4.19 Resultados experimento UPV1 [EP14]	60
Tabla 4.20 Resultados experimento UPV2 [EP14]	61
Tabla 4.21 Resultados experimento UPV3 [EP14]	61
Tabla 4.22 Resultados experimento UNIBAS [EP14]	61
Tabla 4.23 Resumen de la significancia para todos los experimentos de la familia [EP14]	62
Tabla 4.24 Elementos para la identificación de variables contextuales	63
Tabla 4.25 Identificación de diferencias contextuales familia [EP14]	64
Tabla 4.26 Ejemplo de distribución de variables	66
Tabla 4.27 Ejemplo de conjunto inicial de variables familia [EP14]	67
Tabla 4.28 Ejemplo de distribución de variables	67
Tabla 4.29 Ejemplo de conjunto final de variables familia [EP14]	68
Tabla 4.30 Elementos para la identificación de variables de participantes	68
Tabla 4.31 Ejemplo de diferencias entre participantes	68

Tabla 4.32 Ejemplo de conjunto inicial de variables de participantes.....	69
Tabla 4.33 Ejemplo de asociación de variables a sus respectivos niveles	72
Tabla 4.34 Asociación de variables a sus respectivos niveles para la familia [EP14].....	73
Tabla 4.35 Resumen de proceso propuesto.....	78
Tabla 5.1 Procedimiento experimental.....	83
Tabla 5.2 Procedimiento experimental.....	84
Tabla 5.3 Caracterización del experimento UPM-1 (1/2)	85
Tabla 5.4 Caracterización del experimento UPM-1 (2/2)	86
Tabla 5.5 Elementos de la Operacionalización	87
Tabla 5.6 Diseño del experimento UPM-1	88
Tabla 5.7 Diseño del experimento UPM-2	88
Tabla 5.8 Diseño del experimento UPV	88
Tabla 5.9 Diseño del experimento ORT.....	88
Tabla 5.10 Superposición de los diseños UPM04-UPM05	89
Tabla 5.11 Superposición de los diseños UPM04-UPM05-UPV05.....	89
Tabla 5.12 Superposición UPM04-UPM05-UPV05-ORT05.....	90
Tabla 5.13 Variables del análisis	90
Tabla 5.14 Resultados de los análisis individuales.....	91
Tabla 5.15 Elementos del diseño	92
Tabla 5.16 Elementos de la instrumentación	93
Tabla 5.17 Elementos de la población	93
Tabla 5.18 Elementos de los experimentadores	94
Tabla 5.19 Listado de variables	95
Tabla 5.20 Distribución del conjunto final de variables.....	96
Tabla 5.21 Distribución de las variables de los participantes.....	97
Tabla 5.22 Conjunto final de las variables de participantes.....	97
Tabla 5.23 Definición de las variables de análisis.....	99
Tabla 5.24 AICs para el modelo básico familia de técnicas de prueba	100
Tabla 5.25 Resultados del análisis para el modelo refinado	101
Tabla 5.26 Resultados del análisis para el modelo básico	102
Tabla 5.27 Resultados del análisis	103
Tabla 6.1 Detalles de la familia FiDiPro.....	106
Tabla 6.2. Número de casos de prueba y aserciones.....	109
Tabla 6.3 Caracterización Universidad de Oulu 1	111
Tabla 6.4 Elementos de la Operacionalización	112
Tabla 6.5 Diseño FSH, FSK, FSO y EB	113
Tabla 6.6 Diseño OY1	113
Tabla 6.7 Diseño PT.....	113
Tabla 6.8 Diseño UPV	114
Tabla 6.9 Diseño UPM.....	114
Tabla 6.10 Diseño ER.....	114
Tabla 6.11 Diseño UB	114
Tabla 6.12 Diseño USD2	114
Tabla 6.13 Diseño PAF y ETAPA	115
Tabla 6.14 Diseño USD	115
Tabla 6.15 Diseño OY2	115
Tabla 6.16 Diseño NoviSad	115
Tabla 6.17 Diseño agregado (Paso 1).....	116
Tabla 6.18 Diseño agregado (Paso 2, añadir PT)	116
Tabla 6.19 Diseño agregado (Paso 3, añadir OY2)	116
Tabla 6.20 Diseño agregado (Paso 4, añadir UPM).....	116
Tabla 6.21 Diseño agregado (Paso 7, añadir NoviSad).....	117

Tabla 6.22 Diseño agregado (Paso 6, añadir PAF y ETAPA).....	117
Tabla 6.23 Diseño agregado (Paso 5, añadir ER)	117
Tabla 6.24 Diseño agregado (Paso 8, añadir OY1)	118
Tabla 6.25 Diseño agregado (Paso 9, añadir UB).....	118
Tabla 6.26 Diseño agregado (Paso 10, añadir USD).....	119
Tabla 6.27 Diseño agregado (Paso 11, añadir USD2).....	119
Tabla 6.28 Pruebas estadísticas y factores analizados individualmente	120
Tabla 6.29 Significación estadística o práctica en cada experimento (2/2)	121
Tabla 6.30 Elementos de la ejecución experimental	122
Tabla 6.31 Elementos de la Población	123
Tabla 6.32 Elementos de la Instrumentación	124
Tabla 6.33 Elementos de los Experimentadores	125
Tabla 6.34 Variables para el análisis	126
Tabla 6.35 Número de observaciones para cada variable (2/2).....	127
Tabla 6.36 Distribución de las variables de los participantes.....	128
Tabla 6.37 Conjunto final de las variables de participantes.....	129
Tabla 6.38 Definición de las variables de análisis.....	134
Tabla 6.39 AICs para el modelo básico familia TDD.....	135
Tabla 6.40 Resultados del análisis para el modelo refinado	136
Tabla 6.41 Resultados del análisis para el modelo básico	140
Tabla 6.42 Resultados del análisis	141
Tabla 7.1 Artículos que han reportado experimentos con participantes	148
Tabla 7.2 Número de artículos y experimentos por categoría de participantes.....	150
Tabla 7.3 Número de participantes usados en experimentos.....	150
Tabla 7.4 Incorporación de tipo/características de participantes al análisis en el experimento ...	152
Tabla 7.5 Resumen: el tipo de participante se analiza por separado (1/2)	155
Tabla 7.6 Resumen: el tipo de participante se analiza por separado (2/2)	156
Tabla 7.7 Resumen: el tipo de participante se analiza conjuntamente (1/2)	159
Tabla 7.8 Resumen: el tipo de participante se analiza conjuntamente (2/2)	160
Tabla 7.9 Resumen: para las características de los participantes (1/3)	164
Tabla 7.10 Resumen: para las características de los participantes (2/3)	165
Tabla 7.11 Resumen: para las características de los participantes (3/3)	166
Tabla 8.1 Circunstancias que afectan al proceso asociadas a cada paso de este.....	171
Tabla A.1 Grupos de variación para los elementos de la operacionalización	173
Tabla A.2 Grupos de variación para los elementos de la operacionalización	173
Tabla A.3 Grupos de variación para los elementos de la instrumentación.....	174
Tabla A.4 Grupos de variación para los elementos de la Población	175
Tabla A.5 Grupos de variación para los elementos de Experimentador.....	176
Tabla C.1.6 Caracterización UPM-2 (1/2).....	181
Tabla C.1.7 Caracterización UPM-2 (2/2).....	182
Tabla C.1.8 Caracterización UPV (1/2)	182
Tabla C.1.9 Caracterización UPV (2/2)	183
Tabla C.1.10 Caracterización ORT (1/2)	184
Tabla C.1.11 Caracterización ORT (2/2)	185
Tabla C.2.12 Caracterización FSecure (1/2).....	186
Tabla C.2.13 Caracterización FSecure (2/2).....	187
Tabla C.2.14 Caracterización Elektrobit (1/2).....	187
Tabla C.2.15 Caracterización Elektrobit (2/2).....	188
Tabla C.2.16 Caracterización UPM (1/2)	188
Tabla C.2.17 Caracterización UPM (2/2)	189
Tabla C.1.18 Caracterización PlayTech (1/2)	189
Tabla C.2.19 Caracterización PlayTech (2/2)	190

Tabla C.2.20 Caracterización UPV	191
Tabla C.2.21 Caracterización Ericsson (1/2)	192
Tabla C.2.22 Caracterización Ericsson (2/2)	193
Tabla C.2.23 Caracterización UB (1/2)	193
Tabla C.2.24 Caracterización UB (2/2)	194
Tabla C.2.25 Caracterización USD (1/2)	194
Tabla C.2.26 Caracterización USD (2/2)	195
Tabla C.2.27 Caracterización PAF (1/2)	195
Tabla C.2.28 Caracterización PAF (2/2)	196
Tabla C.2.29 Caracterización ETAPA (1/2)	196
Tabla C.2.30 Caracterización ETAPA (2/2)	197
Tabla C.2.31 Caracterización USD2	198
Tabla C.2.32 Caracterización OY2 (1/2)	199
Tabla C.2.33 Caracterización OY2 (2/2)	200
Tabla C.2.34 Caracterización NoviSad (1/2)	200
Tabla C.2.35 Caracterización NoviSad (2/2)	201
Tabla D.1.36 Modelo Mixto UPM-1	203
Tabla D.1.37 Resultados UPM-1 (1/2)	203
Tabla D.1.38 Resultados UPM-1 (2/2)	204
Tabla D.1.39 Modelo Mixto UPM-2	204
Tabla D.1.40 Resultados UPM-2	205
Tabla D.1.41 Modelo Mixto UPV05	206
Tabla D.1.42 Resultados UPV05	206
Tabla D.1.43 Resultados ORT05	207
Tabla D.1.44 Resultados ORT05	207
Tabla D.2.45 Resultados FSecure Helsinki	210
Tabla D.2.46 Resultados FSecure Kuala Lumpur	210
Tabla D.2.47 Resultados FSecure Oulu	211
Tabla D.2.48 Resultados Electrobit	211
Tabla D.2.49 Resultados OY1	211
Tabla D.2.50 Modelo Mixto UPM	212
Tabla D.2.51 Resultados UPM (1/2)	212
Tabla D.2.52 Resultados UPM	213
Tabla D.2.53 Modelo Mixto PlayTech	214
Tabla D.2.54 Resultados PlayTech	214
Tabla D.2.55 Resultados UPV	215
Tabla D.2.56 Modelo Mixto Ericsson	215
Tabla D.2.57 Resultados Ericsson	216
Tabla D.2.58 Modelo Mixto UB1	217
Tabla D.2.59 Resultados UB1	217
Tabla D.2.60 Resultados USD-UG	218
Tabla D.2.61 Resultados USD-G	218
Tabla D.2.62 Resultados USD	219
Tabla D.2.63 Modelo Mixto PAF	219
Tabla D.2.64 Resultados PAF	220
Tabla D.2.65 Resultados ETAPA	220
Tabla D.2.66 Resultados ETAPA	221
Tabla D.2.67 Modelo Mixto OY2	222
Tabla D.2.68 Resultados OY2	222
Tabla D.2.69 Modelo Mixto Novisad	223
Tabla D.2.70 Resultados Novisad	223
Tabla E.1.71 Resultados Modelo Base	225

Tabla E.1.72 Resultados Modelo Añadiendo las Variables Identificadas.....	225
Tabla E.1.73 Resultados Modelo Añadiendo las Variables Identificadas sin DOTSEM	226
Tabla E.1.74 Resultados Modelo Añadiendo la Técnica como Efecto Aleatorio	226
Tabla E.1.75 Resultados Modelo Añadiendo Interacciones	226
Tabla E.1.76 Resultados Modelo Final	227
Tabla E.2.77 Resultados Modelo Base	228
Tabla E.2.78 Resultados Modelo Añadiendo las Variables Identificadas.....	228
Tabla E.2.79 Resultados Modelo Añadiendo Enfoque de Desarrollo como Efecto Aleatorio.....	229
Tabla E.2.80 Resultados Modelo Añadiendo Interacciones (1/2)	229
Tabla E.2.81 Resultados Modelo Añadiendo Interacciones (2/2)	230
Tabla E.2.82 Resultados Modelo Final	230

Índice de Figuras

Figura 2.1 Distribución de las publicaciones identificadas en el mapeo sistemático de la literatura por año y procedencia	9
Figura 2.2 Número de experimentos en cada estudio primario	10
Figura 2.3 Frecuencia de estudios primarios según el número de experimentos.....	10
Figura 2.4 Cambios en las familias de experimentos de los estudios primarios	13
Figura 3.1 Problemas y Objetivos de Investigación	33
Figura 4.1 Proceso para la agregación de familias de experimentos propuesto.....	38
Figura 4.2 Pantalla para añadir un nuevo experimento.....	50
Figura 4.3 Superposición de 2 diseños de 2 factores, en el que se añaden nuevos niveles	56
Figura 4.4 Superposición de 2 diseños de un factor, en el que se crea un nuevo factor.....	56
Figura 4.5 Superposición de 2 diseños intra-sujetos, en el que se e crea un diseño crossover.....	56
Figura 4.6 Diseño de un factor incluido en un diseño de dos factores	57
Figura 4.7 Diseño intra-sujetos incluido en un diseño crossover	57
Figura 4.8 Diseño entre-sujetos interseca en un diseño intra-sujetos	58
Figura 4.9 Ejemplo de una familia de experimentos multinivel con diseño entre-sujetos.....	70
Figura 4.10 Ejemplo de una familia de experimentos multinivel con diseño intra-sujetos	70
Figura 4.11 Estructura jerárquica para la familia [EP14].....	72
Figura 4.12 Iteraciones para obtener un modelo de análisis final	76
Figura 5.1 Experiencia relativa de los participantes	98
Figura 5.2 Experiencia absoluta de los participantes.....	98
Figura 5.3 Jerarquía de los experimentos	99
Figura 5.4 Resultados de la interacción Técnica*Experiencia Relativa.....	104
Figura 6.1 Experiencia en Programación.....	129
Figura 6.2 Experiencia con el Lenguaje de Programación.....	130
Figura 6.3 Experiencia en Pruebas Unitarias	130
Figura 6.4 Experiencia con el framework de pruebas.....	131
Figura 6.5 Modelo jerárquico familia TDD (1/2)	132
Figura 6.6 Modelo jerárquico familia TDD (2/2)	133
Figura 6.7 Modelo 5: Incorporar interacciones de entre variables de distinto nivel.....	137
Figura 6.8 Resultados para la Experiencia en Programación	140
Figura 6.9 Comportamiento del enfoque de desarrollo a través de los distintos experimentos....	142
Figura 7.1 Caracterización de los participantes.....	151
Figura 7.2 Interacción Enfoque de Desarrollo*Tipo de Sujeto.....	158
Figura 7.3 Interacción Tipo de Sujeto*Enfoque de Desarrollo.....	158
Figura 7.4 Experiencia en programación según el tipo de sujeto.....	162
Figura A.1 Distribución de cambios para la Operacionalización	173
Figura A.2 Distribución de cambios para la Ejecución Experimental.....	174
Figura A.3 Distribución de cambios para la Instrumentación	174
Figura A.4 Distribución de cambios para la Población.....	175
Figura A.5 Distribución de cambios para los Experimentadores	175
Figura B.6 Listado de usuarios de la aplicación	177
Figura B.7 Creación de un nuevo usuario	177
Figura B.8 Listado de familias de experimentos	178
Figura B.9 Listado de los experimentos de una familia	178
Figura B.10 Detalle de la Operacionalización de un experimento.....	179
Figura B.11 Detalle de la Población de un experimento	179
Figura B.12 Edición de un experimento	180

Capítulo 1

Introducción

1.1 Área de Investigación

La Ingeniería de Software Empírica (ISE) es el área de la Ingeniería de Software (IS) cuyo objetivo consiste en la obtención de evidencia sobre los métodos, técnicas, procesos y herramientas involucrados en el desarrollo del software (Basili, et al., 1986). La evidencia se obtiene mediante la realización de estudios empíricos (Basili, et al., 1986), (Basili, 1996), (Fenton, 1993).

Existen distintos tipos de estudios empíricos. Entre los más utilizados (Wohlin, et al., 1999), se encuentran: las encuestas, los casos de estudio y los experimentos controlados. Cada tipo de estudio posee unas características diferentes (Wohlin, et al., 1999), (Pfleeger, 1994) en función de: la cantidad de control que el investigador tiene sobre el estudio, el grado en que el investigador puede decidir qué medidas se deben recopilar durante el estudio, el coste del estudio y la facilidad de replicación. De los distintos tipos de estudios empíricos existentes, **este trabajo se centra en los experimentos controlados** (también denominados simplemente, experimentos).

En un experimento, las variables clave del fenómeno bajo estudio se manipulan en un ambiente controlado con el fin de investigar el impacto de las variaciones en dichas variables clave (Juristo & Moreno, 2001). El entorno controlado tiene como objetivo garantizar que ninguna otra variable afecte los resultados y, por tanto, permitir revelar causalidad. En un experimento, el investigador tiene un control total sobre el estudio y las medidas que se deben recopilar, por su parte, la facilidad de replicación y el coste de la investigación son altas (Wohlin, et al., 1999).

1.2 Problema

En ISE, es bien sabido que los resultados de un único experimento muestran solamente una parte de la información. La obtención de evidencia fiable requiere de la combinación de resultados de múltiples experimentos (Gómez, et al., 2014), (Juristo, 2016), (Shull, et al., 2008), (Kitchenham, 2008).

La combinación de los resultados de varios experimentos requiere dos pasos:

- *Recopilar un conjunto de experimentos relacionados* que investiguen las mismas hipótesis.
- *Agregación de sus resultados* para obtener un hallazgo global.

Actualmente, hay dos enfoques para recopilar un conjunto de experimentos relacionados. El primer enfoque es el de *familia de experimentos*. El concepto de familia de experimentos aparece por primera vez en (Basili, et al., 1999), referido a la necesidad de realizar diferentes réplicas del mismo experimento para consolidar un cuerpo de conocimiento en IS. Una familia de experimentos (Santos, et al., 2018) es un conjunto de experimentos donde:

- Se garantiza el acceso a los datos sin procesar.
- Los investigadores conocen la configuración exacta de cada uno de los experimentos.
- Al menos tres experimentos evalúan los efectos de al menos dos tecnologías diferentes en la misma variable de respuesta.

El segundo enfoque es la revisión sistemática de la literatura (RSL). La RSL está asociada con el concepto de IS basada en evidencia (Kitchenham, 2004), (Dyba, et al., 2005), que proviene de una analogía con la medicina (medicina basada en evidencia). Consiste en la búsqueda de experimentos publicados sobre un determinado tema, para luego combinar sus resultados.

Hay claras diferencias entre los dos enfoques (Santos, et al., 2018). En las familias de experimentos existe una mayor libertad en la realización de análisis estadísticos (debido al conocimiento profundo de la configuración de cada experimento y al acceso a los datos recopilados), y se puede cambiar de forma planificada y controlada la configuración experimental para dar una respuesta global a la hipótesis. En las RSL, es posible tener acceso a una mayor cantidad de experimentos publicados, que han sido realizados de forma independiente por diferentes investigadores. Esta investigación se centra en cómo **agregar los resultados de una familia de experimentos**.

Un concepto inherente a las familias de experimentos es el del cambio. Típicamente, los experimentos de una familia no serán exactamente iguales (aunque si muy parecidos). La existencia de cambios en la configuración experimental de una familia de experimentos puede ser causa de resultados discrepantes. Por tanto, a la hora de agregar una familia de experimentos será necesario tener en cuenta los cambios que se han producido en la misma. Típicamente, la forma de incorporar estos cambios a la agregación es en forma de *variables moderadoras*. Un variable moderadora (Krein, et al., 2016) se define como: cualquier variable explicativa que podría modificar el efecto que las variables independientes tienen sobre una variable de respuesta.

En un estudio de mapeo sistemático de la literatura que se ha realizado sobre el estado de la práctica en lo que se refiere a la agregación de familias de experimentos en IS (y que se detalla en el Capítulo 2 de esta tesis) se ha descubierto que **no existe un proceso definido**. Un proceso para agregar familias de experimentos debe contemplar:

1. La identificación precisa de todas las similitudes y diferencias existentes entre los experimentos de los que consta la familia.
2. El establecimiento de manera sistemática de un conjunto de variables (tanto moderadoras como debidas al diseño experimental) que podrían estar afectando a los resultados de los experimentos de la familia.
3. El análisis de los datos de la familia, teniendo en cuenta de forma conjunta todas y cada una de las variables anteriormente identificadas.

Con respecto a estos tres aspectos, en el estudio de mapeo sistemático que se describe en el Capítulo 2, muestra el estado de la práctica en lo que se refiere al análisis de familias de experimentos. En él, se ha observado que:

1. *La identificación de similitudes y diferencias entre los experimentos de una familia no se realiza de forma precisa.* La identificación se realiza de manera informal. En el mejor de los casos se realiza la descripción del experimento base (experimento inicial) siguiendo las guías y recomendaciones de otros autores para el reporte de experimentos, y sus replicaciones se detallan en base a las diferencias existentes con respecto al experimento base. Por lo general, se suelen omitir detalles sobre los experimentos, lo que conlleva a una identificación de similitudes y diferencias imprecisa. En concreto, muchas diferencias pasan desapercibidas.
2. *El establecimiento del conjunto de variables no se realiza de forma sistemática.* El conjunto de variables a incluir en el análisis se crea a partir de las diferencias existentes entre los experimentos. Sin embargo, debido a que no todas las diferencias se transforman en variables, no se sabe bajo qué criterio se realiza la elección de unas frente a otras. En raras ocasiones, se introduce una diferencia premeditadamente, para tener una variable de la que se quiere conocer su posible impacto. Únicamente se identifican variables moderadoras, tanto de contexto como de participante, pero en ningún caso se identifican variables de diseño.
3. *El análisis de datos no incorpora de forma conjunta todas las variables identificadas.* En muchos casos, las variables identificadas ni siquiera llegan a analizarse nunca. En otros casos, se realizan análisis separados para cada una de las variables identificadas. Las técnicas de análisis utilizadas varían de una familia a otra, no existiendo un estándar.

Por su parte, un estudio del arte sobre el análisis de familias de experimentos arroja que no existe apenas investigación acerca de cómo deben analizarse las familias de experimentos en IS (Santos, et al., 2019), tal y como se muestra en el Capítulo 2.

El **objetivo principal** de esta tesis consiste en **proponer un proceso para la agregación de familias de experimentos en IS** que resuelva las carencias encontradas.

1.3 Aproximación a la Solución

La solución propuesta parte del estudio del estado de la práctica en lo que se refiere al proceso seguido en IS para la agregación de familias de experimentos. En concreto, se estudian los aspectos del proceso relativos a: 1) cómo se identifican las similitudes y diferencias entre los experimentos de la familia, 2) cómo se establece el conjunto de variables que deben incluirse en el análisis y 3) cómo se analizan los datos recogidos en los experimentos.

A partir de los resultados obtenidos en dicho estudio, se establecen los requisitos que debe cumplir el proceso de agregación de familias de experimentos y que giran en torno a los tres ejes anteriormente mencionados. A continuación, se diseña dicho proceso.

Para evaluar el proceso propuesto, se analizan dos familias de experimentos. La primera de ellas es una familia de experimentos que estudia la efectividad relativa de distintas técnicas de evaluación de código, que consta de 4 experimentos. La segunda es una familia de experimentos en la que se estudia

la calidad del código desarrollado utilizando desarrollo guiado por pruebas de software (de ahora en adelante, TDD, del inglés Test-driven development), frente al código desarrollado utilizando un enfoque de desarrollo tradicional, que consta de 16 experimentos.

Finalmente, en el estudio de mapeo sistemático se encontró que la variable moderadora más estudiada en las familias de experimentos en IS es el tipo de sujeto (el 39% de las familias de experimentos que identifican variables moderadoras estudian esta variable). Sin embargo, actualmente existe la hipótesis de que la etiqueta estudiante/profesional no sería la más adecuada para reflejar las diferencias entre los distintos tipos de sujetos empleados, y que podrían ser ciertas características de los sujetos las que estuviesen influyendo en los resultados experimentales (Feldt, et al., 2018). Por lo tanto, se estudia la idoneidad de la utilización de la etiqueta estudiante/profesional en la familia de experimentos de TDD analizada. Además, el estudio de la idoneidad se explorará a las familias de experimentos en IS identificadas en un mapeo sistemático, que utilizan tanto estudiantes como profesionales.

1.4 Publicaciones Derivadas

Hasta el momento, se ha conseguido una publicación corta (tipo póster) en congreso CORE A* (GGs A++), relacionada con el Capítulo 7 de la tesis:

- Sira Vegas, Patricia Riofrío, Natalia Juristo. Does Subject Type Influence Software Engineering Experiment Results? *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, Buenos Aires, Argentina, Mayo 2017, pp. 210-212.

Y una publicación en revista JCR Q1, relacionada con el Capítulo 5 de la tesis:

- Sira Vegas, Patricia Riofrío, Esperanza Marcos, Natalia Juristo. On (Mis)Perceptions of Testing Effectiveness: An Empirical Study. *Empirical Software Engineering Journal*, pp. 1-59. (En prensa)

Asimismo, se están finalizando dos publicaciones para revista JCR Q1 que serán enviadas en breve:

- Patricia Riofrío, Sira Vegas, Natalia Juristo. A Method for Aggregating Families of Experiments in Software Engineerings. *Empirical Software Engineering Journal*. Capítulos 4 y 5 de la tesis.
- Sira Vegas, Patricia Riofrío, Natalia Juristo. Exploring the Impact of Using Students in Software Engineering Experiments. *IEEE Transactions on Software Engineering*. Capítulo 7 de la tesis.

1.5 Organización del Documento

Esta investigación se desarrolla en 8 capítulos, que se describen a continuación:

- Los tres primeros capítulos realizan el planteamiento de la investigación. El Capítulo 1 explica brevemente el problema a tratar y cuál es la aproximación elegida para resolverlo.

El Capítulo 2 presenta el estado de la práctica, para ello se ha estudiado el método que siguen para agregar experimentos en las familias de 46 estudios primarios. Además, presenta en estado del arte, para ellos se han recopilado trabajos relacionados con esta investigación. El Capítulo 3 explica detalladamente el problema a resolver y se presentan y desarrollan cada uno de los objetivos de este trabajo.

- El Capítulo 4 detalla el proceso seguido para dar solución al problema planteado.
- Los dos siguientes capítulos se centran en la evaluación de la solución propuesta. Concretamente, en el Capítulo 5 se realiza la validación con una familia de tamaño pequeño (4 experimentos), mientras que en el Capítulo 6 la validación es realizada con una familia de tamaño grande (16 experimentos).
- El Capítulo 7, analiza la idoneidad de utilizar la etiqueta estudiante/profesional para la caracterización del tipo de sujeto en experimentos en IS.
- Finalmente, el Capítulo 8 muestra las conclusiones obtenidas durante la investigación y las futuras líneas de investigación.

Asimismo, los apéndices del documento muestran detalles relevantes de la investigación realizada.

Capítulo 2

Estado de la Cuestión

El objetivo de este capítulo consiste en conocer el estado de la práctica y del arte en lo relativo a la agregación de familias de experimentos. Para ello, se presenta en primer lugar un mapeo sistemático de la literatura cuyo objetivo consiste en estudiar cómo se lleva a cabo la agregación de experimentos en las familias de experimentos reportadas en el ámbito de la IS. A continuación, se estudia cómo debe realizarse la agregación según los investigadores en ISE. La conjugación de ambas vistas dará una visión completa del proceso de agregación.

2.1 Estado de la Práctica

2.1.1 Proceso de Búsqueda

Se quiere conocer cómo son los procedimientos de agregación de experimentos. Santos, et al. realizan un mapeo sistemático de la literatura sobre las técnicas de análisis estadísticos utilizados en familias de experimentos de IS (Santos, et al., 2018). Dado que el objetivo aquí es conocer el proceso de agregación que se sigue en las familias de experimentos en IS, se busca el mismo tipo de publicaciones. Por lo tanto, para este estado de la práctica se parte del conjunto de estudios primarios identificado en (Santos, et al., 2018). Dado que dicha publicación realiza la búsqueda hasta octubre del año 2016, esta se debe actualizar. Para ello se reproducirá la búsqueda inicial reportada en (Santos, et al., 2018) restringida al periodo 2016-2018¹.

Tal cual se realiza la búsqueda inicial, la búsqueda actualizada se lleva a cabo en las de bases de datos IEEE Xplore, ISI Web of Science, Science Direct y Scopus. La cadena utilizada es: (experiment*) AND (famil* OR serie* OR group*).

La tabla 2.1 muestra los resultados arrojados por la búsqueda actualizada. La segunda columna muestra el número total de artículos obtenidos. La tercera columna muestra el número de publicaciones duplicadas y la última columna muestra el número total de publicaciones relevantes para esta investigación. La búsqueda se realizó en el mismo orden que en el que aparecen las fuentes en la primera columna de la Tabla 2.1, es decir, Scopus fue la primera base de datos de búsqueda, mientras que Science Direct fue la última. Por lo tanto, las dos últimas columnas tienen en cuenta las fuentes anteriores. Por

¹ Se incluye el año 2016 porque la búsqueda inicial concluía en octubre de 2016 y por lo tanto no se disponía del año entero.

ejemplo, de las 22 publicaciones obtenidas en IEEE Xplore, 11 ya se habían obtenido previamente en Scopus, entre las 9 publicaciones relevantes. Por tanto, en esta base de datos no se obtiene ninguna publicación *nueva* relevante.

Fuente	Número de Publicaciones	Duplicados	Relevantes
Scopus	103	-	9
IEEE Xplore	22	11	0
ISI Web of Science	106	90	1
Science Direct	19	17	0
Total	250	118	10

Tabla 2.1 Número de publicaciones obtenidas durante el proceso de búsqueda

De los 10 estudios primarios relevantes obtenidos, uno ya había sido identificado en la búsqueda inicial por lo que debe descartarse. Además, otro de estos, coincide con otro ya identificado en la búsqueda inicial que, aunque se trata de publicaciones diferentes, reportan los mismos experimentos, y, por lo tanto, tampoco se va a incluir. Finalmente, se obtienen un total de 8 publicaciones relevantes. Posteriormente, con el proceso de bola de nieve hacia atrás no se obtuvo ningún estudio primario adicional.

Por otro lado, de los 39 estudios primarios identificados en la búsqueda inicial, dos estudios distintos reportaban los mismos experimentos, por lo que se descartó uno de estos, obteniendo un total de 38 estudios primarios.

Esto hace un total de 46 estudios primarios: 8 de la búsqueda actualizada y 38 de la búsqueda inicial.

2.1.2 Proceso de Selección

Asimismo, los 8 estudios primarios obtenidos en la búsqueda actualizada, deben que cumplir los criterios establecidos en el proceso de selección inicial:

- Todos los experimentos tienen que comparar *al menos dos tratamientos iguales* para una *variable de respuesta común*.
- Tienen que haber *al menos tres experimentos* dentro de la familia, de esta manera se evita el problema de considerar una replicación aislada de un experimento como una familia de experimentos y se asegura que las técnicas de agregación se puedan aplicar dentro de las familias, independientemente de si el objetivo es proporcionar resultados conjuntos o investigar moderadores.
- Los experimentos son realizados con *participantes humanos*.
- El artículo *no está duplicado*.
- Se trata de un artículo de *revisión por pares* (por ejemplo, lo que se conoce como una convocatoria de ponencias, discurso de apertura, prefacio).

Todas las publicaciones obtenidas cumplen estos criterios y, por lo tanto, no es necesario descartar ninguno. Finalmente, se tiene un total de 46 estudios primarios relevantes cuya distribución se muestra en la Figura 2.1: de los cuales 8 se corresponden con la búsqueda actualizada (representadas en color morado) y 38 de la búsqueda inicial (representadas en color negro).

En la Figura 2.1 se puede observar que a partir del año 2009 (excepto 2010 y 2017) el número de publicaciones incrementa considerablemente, siendo los años 2014 y 2015 (7 y 5 respectivamente), donde se ha realizado el mayor número de publicaciones. Además, se puede observar que la mayoría de las publicaciones fueron encontradas en revistas (83,33%). De hecho, la suma de artículos publicados en *Information and Software Technology (IST)* y *Empirical Software Engineering (EMSE)*, es más del doble que los encontrados en Congresos y Workshops.

Figura 2.1 Distribución de las publicaciones identificadas en el mapeo sistemático de la literatura por año y procedencia

Aunque, se han encontrado publicaciones en congresos desde el año 2005 al 2015, en los años 2006, 2007, 2009 y 2010 no hay ninguna. Además, solo hay una publicación en workshops (en el 2009). Finalmente, en los años 1999, 2000, 2002, 2003 y 2007 no se ha encontrado artículos.

Por otro lado, la Figura 2.2 el número total de experimentos que se han realizado en cada uno de los estudios encontrados. La Figura 2.3 muestra la frecuencia según el total de experimentos realizados en cada una de las familias, por tanto, se puede observar que la mayoría de las familias se componen de 3 experimentos y que es poco frecuente que haya familias con 5 o más experimentos.

Figura 2.2 Número de experimentos en cada estudio primario

Figura 2.3 Frecuencia de estudios primarios según el número de experimentos

2.1.3 Extracción de Datos

El objetivo principal de este mapeo sistemático de la literatura consiste en estudiar cuál es el proceso de agregación de experimentos seguido en las familias encontradas. Concretamente, se quiere responder a las siguientes 4 preguntas:

- P_1 : ¿Existen diferencias entre los experimentos que componen una familia?
Interesa conocer hasta que punto los experimentos de las familias son idénticos entre sí o por el contrario se realizan cambios entre ellos. Esta investigación tiene como objetivo

proponer un proceso de agregación que trata sistemáticamente la incorporación de cambios en el análisis.

- *P₂*: ¿Cómo se identifican las diferencias entre los experimentos de una familia?

Interesa conocer de qué forma se caracterizan/describen cada uno de los experimentos en las familias de experimentos de IS y cómo se identifican las diferencias y similitudes entre los experimentos de una misma familia.

El reconocimiento de diferencias y similitudes entre los experimentos tiene dos partes:

- Caracterización de los experimentos. La caracterización puede realizarse de distintos modos: de manera informal, mediante guías y recomendaciones para reportar experimentos y a través de una plantilla de caracterización experimental para identificar tipos de replicaciones.
- Identificación de diferencias. Se pueden dar los siguientes casos: se describen una por una las características de los elementos de la configuración experimental (participantes, materiales, procedimiento, etc.), se comparan los experimentos con el experimento base, se describe cada experimento por separado, se describen todos los experimentos como un único conjunto o existe una sección para tratar exclusivamente las diferencias.

- *P₃*: ¿Cómo se identifican las variables a incluir en el análisis en una familia?

Interesa conocer de qué forma se identifican las variables a incluir en el análisis en las familias de experimentos de IS y qué variables son.

La identificación o selección de variables de análisis tiene dos partes:

- Identificación de variables que podrían influir en los resultados. En todos los casos las variables identificadas representan diferencias entre los experimentos. Sin embargo, existen familias en las que no se identifica ninguna variable, y en las que se identifican no se sabe bajo qué criterio.
- El tipo de variables identificadas. Únicamente se identifican variables moderadoras, tanto de participante, como de contexto y en algunos casos se identifican ambos tipos de variables. Finalmente, nunca se identifican variables de diseño.

- *P₄*: ¿Cómo se analizan conjuntamente los experimentos de una familia?

Interesa conocer qué técnicas de agregación, así como el test estadístico utilizado para estudiar el impacto de las variables estudiadas, y la manera de incorporar en el análisis dichas variables.

El análisis de los experimentos se ha dividido en tres partes:

- La técnica de agregación utilizada. Existen 5 técnicas de agregación diferentes: Síntesis narrativa, meta-análisis de datos agregado, agregación de p-valores, meta-análisis de Datos Individuales de Participante y meta-análisis de Datos Individuales de Participante estratificado.
- La técnica de análisis asociada a la técnica de agregación. Aunque para cada técnica de agregación existen pruebas estadísticas específicas, únicamente se estudiarán las relacionadas con meta-análisis de Datos Individuales de Participante y meta-análisis de Datos Individuales de Participante estratificado y cuyos valores pueden ser: pruebas paramétricas, no paramétricas, correlación, bayesianos y la prueba del coeficiente Alfa de Cronbach.
- La estrategia de análisis utilizada. Finalmente, si se ha identificado más de una variable estas se analizan conjuntamente o una por una individualmente. Sin embargo, también

existen estudios primarios en los que se ha identificado una única variable o que no se ha identificado ninguna.

2.1.4 P₁: ¿Existen Diferencias Entre los Experimentos que Componen una Familia?

Para identificar si existen diferencias entre los experimentos de una familia, se ha utilizado la plantilla de caracterización definida en la Tabla 4.3 del Capítulo 4 de esta tesis; esta plantilla consta de 37 elementos que están agrupado en 5 dimensiones que son: Operacionalización, Ejecución experimental, Instrumentación, Población y Experimentadores. Dicha plantilla se ha instanciado para todos y cada uno de los experimentos de cada una de las 46 familias de experimentos identificadas, de acuerdo con la información proporcionada en cada estudio primario sobre la familia. Sin embargo, existe una limitación asociada a estos resultados y es que en ocasiones no se puede afirmar si un elemento ha cambiado o no, ya que los autores omiten algunos detalles de los experimentos.

Asimismo, para identificar cuantitativamente la magnitud de los cambios se ha utilizado una matriz de distancia. La distancia es un número comprendido entre 0 y 1 y mide el ratio de elementos de la plantilla de caracterización que han cambiado en 2 experimentos de una familia.

La Tabla 2.2 muestra un ejemplo de la matriz generada para la familia descrita en [EP14], la parte de arriba de la diagonal de esta matriz no está rellena porque la matriz es simétrica. En ella se puede comprobar que los experimentos más parecidos entre sí son UPV2 y UPV3, ya que son los que tienen el valor mínimo de cambios (0,08), mientras que el experimento más diferente a los demás es UNIBAS ya que su valor es el más alto (0,43).

Total	UPV1	UPV2	UPV3	UNIBAS
UPV1	0	-	-	-
UPV2	0,16	0	-	-
UPV3	0,14	0,08	0	-
UNIBAS	0,43	0,43	0,43	0

Tabla 2.2 Matriz de distancia para la familia [EP14]

En la Figura 2.4 se muestra el número mínimo, máximo y medio de variaciones entre los experimentos de cada una de las familias encontradas, esta gráfica está ordenada de manera creciente según la media. Dichos números se han calculado a partir de las matrices correspondientes. Por ejemplo, en el caso de la familia [EP14] el valor mínimo es 0,08, el valor máximo es 0,43 y la media es 0,28 sacada de realizar la media entre todos los valores de debajo de la diagonal de la matriz.

A partir de los resultados obtenidos en la Figura 2.4, se han definido 3 grupos de familias de experimentos atendiendo al valor medio de cambios entre los experimentos, para ello se identifica la familia ([EP8]) que tiene la media más grande (0,32) y luego se obtienen 3 grupos iguales cuyos valores son: poco ($\text{media} \leq 0,11$), medio ($0,11 < \text{media} \leq 0,22$) o mucho ($0,22 < \text{media} \leq 0,32$). La Tabla 2.3 muestra los grupos con los respectivos estudios primarios. Finalmente, se puede concluir que **en todas las familias hay variación** y en la mayoría de estas los cambios toman el valor medio.

El mismo procedimiento se ha realizado para cada una de las cinco dimensiones de la configuración experimental definidos en la plantilla de caracterización. Las figuras de valor mínimo, máximo y media y tablas de los grupos generados para cada caso se encuentran en el Apéndice A.

Figura 2.4 Cambios en las familias de experimentos de los estudios primarios

Grupo	Estudio Primario	Total
Poco	[EP7], [EP24], [EP26], [EP27], [EP30], [EP32], [EP33], [EP35], [EP44], [EP45], [EP46]	11 (24%)
Medio	[EP1], [EP2], [EP3], [EP5], [EP6], [EP9], [EP11], [EP12], [EP13], [EP15], [EP16], [EP18], [EP19], [EP20], [EP21], [EP22], [EP23], [EP25], [EP28], [EP29], [EP34], [EP37], [EP38], [EP39], [EP40], [EP42], [EP43]	27 (59%)
Mucho	[EP4], [EP8], [EP10], [EP14], [EP17], [EP31], [EP36], [EP41]	8 (17%)

Tabla 2.3 Grupos de variación: todos los elementos de la configuración experimental

La Tabla 2.4 muestra un resumen de los resultados obtenidos, en ella se reporta el número total de elementos cuyo valor mínimo es:

- Mayor que 0, es decir, hay diferencias entre los elementos de la dimensión correspondiente)
- Igual a 0, es decir, no hay diferencias o no existe información sobre dichos elementos.

Por lo tanto, de la información mostrada en la Tabla 2.4, se puede concluir que la *Población* es la dimensión en la que se han realizado más cambios. Aunque la dimensión *Experimentador* parece ser la que menos cambios presenta en realidad es la dimensión que más se omite.

	Operacionalización	Ejecución Experimental	Instrumentación	Población	Experimentador
Mayor que 0	20 (43%)	31 (67%)	33 (72%)	39 (85%)	14 (30%)
Igual a 0	26 (57%)	15 (33%)	13 (28%)	7 (15%)	32 (70%)

Tabla 2.4 Cambios para cada uno de los elementos de la configuración experimental

2.1.5 P₂: ¿Cómo se identifican las diferencias entre los experimentos de una familia?

2.1.5.1 Caracterización de las replicaciones

La Tabla 2.5 muestra los instrumentos utilizados en cada una de las familias para la caracterización de los experimentos de la familia.

En la mayoría de los estudios primarios (61%) la caracterización de los experimentos se hace siguiendo las guías y recomendaciones de otros autores para reportar experimentos. Es decir, el instrumento de caracterización utilizado es el mero reporte de experimentos. Por su parte, en un 37% de las familias no se especifica el instrumento utilizado para la descripción. Finalmente, existe un único estudio primario que utiliza una plantilla para determinar el tipo de replicación propuesta por (Gómez, et al., 2014); merece la pena mencionar que varios de los autores forman parte del grupo de investigación en el que se realiza esta tesis.

Instrumento	Estudio Primario	Total
Siguiendo guías y recomendaciones	[EP1], [EP2], [EP3], [EP4], [EP5], [EP6], [EP7], [EP9], [EP10], [EP11], [EP14], [EP15], [EP18], [EP20], [EP24], [EP25], [EP26], [EP30], [EP31], [EP32], [EP33], [EP35], [EP36], [EP38], [EP40], [EP41], [EP42], [EP45]	28 (61%)
Informal	[EP8], [EP12], [EP13], [EP16], [EP17], [EP19], [EP21], [EP22], [EP23], [EP28], [EP29], [EP34], [EP37], [EP39], [EP43], [EP44], [EP46]	17 (37%)
Plantilla de replicaciones	[EP27]	1 (2%)

Tabla 2.5 Instrumentos utilizados para la caracterización de familias de experimentos

En la Tabla 2.6 se muestra a qué autores corresponden las guías y recomendaciones utilizadas en las respectivas familias. Se puede observar que, 7 estudios primarios utilizan más de una guía a la vez; por otro lado, las guías más utilizadas son las de (Wohlin, et al., 1999).

Estudios Primarios	Total	(Carver, 2010)	(Ciolkowski, et al., 2002)	(Jedlitschka, et al., 2008)	(Juristo & Moreno, 2001)	(Kitchenham, et al., 2002)	(Wohlin, et al., 1999)
[EP1], [EP4], [EP9], [EP11], [EP14], [EP26]	6 (13,04%)	-	x	-	-	-	-
[EP2], [EP18], [EP38], [EP40], [EP41]	5 (10,87%)	-	-	x	-	-	-
[EP3], [EP5], [EP6], [EP15], [EP25], [EP36], [EP42]	7 (15,22%)	-	-	-	-	-	x
[EP7], [EP32], [EP33]	3 (6,52%)	-	-	-	x	-	x
[EP10]	1 (2,17%)	x	-	x	x	-	x
[EP20]	1 (2,17%)	-	x	x	-	-	x
[EP24], [EP45]	2 (4,35%)	-	-	-	-	x	-
[EP30]	1 (2,17%)	-	x	-	-	-	x
[EP31]	1 (2,17%)	x	-	-	-	-	-
[EP35]	1 (2,17%)	-	-	x	-	-	x

Tabla 2.6 Siguiendo guías y recomendaciones de otros autores

De los resultados obtenidos, se puede concluir que **el instrumento de caracterización utilizado no está específicamente diseñado para identificar diferencias y similitudes** entre los experimentos de una familia.

2.1.5.2 Identificación de Diferencias

En la Tabla 2.7 se puede observar cómo se identifican las diferencias entre los experimentos de las familias de cada un de los estudios primarios. Nótese que en todos los casos la identificación de diferencias se realiza a partir del instrumento correspondiente utilizado en esa familia.

Se ha encontrado que hay cinco formas de identificar las diferencias entre los experimentos:

Diferencias	Estudio Primario	Total
Descripción	[EP2], [EP5], [EP6], [EP7], [EP10], [EP11], [EP12], [EP18], [EP19], [EP20], [EP21], [EP24], [EP25], [EP29], [EP32], [EP34], [EP35], [EP36], [EP37], [EP38], [EP40], [EP41], [EP44], [EP45], [EP46]	25 (54%)
Experimento base	[EP1], [EP4], [EP9], [EP14], [EP15], [EP22], [EP27]	7 (15%)
Cada experimento por separado	[EP3], [EP8], [EP17], [EP23], [EP26], [EP42], [EP43]	7 (15%)
Conjunto	[EP13], [EP16], [EP30], [EP33]	4 (9%)
Sección	[EP28], [EP31], [EP39]	3 (7%)

Tabla 2.7 Identificación de diferencias entre experimentos

- *Descripción.* Se describe una por una las características de los elementos utilizados en la configuración experimental (participantes, material, procedimiento, etc.) y cuando uno es diferente se menciona el experimento en el que varía.
- *Descripción en función de otro experimento.* Se describe un experimento, normalmente el base, que es el que se realiza primero. Posteriormente para el resto de los experimentos únicamente se comentan las diferencias existentes con respecto al mismo. Nótese que pueden existir diferencias entre los otros experimentos de la familia; sin embargo, estas diferencias rara vez son identificadas. Por ejemplo, para una familia de 3 experimentos (E1, E2 y E3), cuya descripción global se realiza para E1, y luego se menciona que E2 difiere con respecto a E1 en los elementos *a*, *b* y que E3 difiere con respecto a E1 en los elementos *c*, *d*; por tanto, se desconoce si los cambios (*a*, *b*) realizados en E2, se mantienen y por lo tanto también se realizan en E3, es decir, no se menciona que E2 y E3 podrían diferir entre ellos en *a*, *b*, *c*, *d*.
- *Cada experimento por separado.* Se describe cada experimento por separado, con lo cual en ciertas ocasiones se pueden identificar similitudes entre ellos. Adicionalmente, en [EP17] y [EP43] los autores proporcionan un resumen de las diferencias.
- *Conjunto.* Se describen todos los experimentos como si fuera uno único y en caso de existir diferencias, se las menciona durante la descripción.
- *Sección.* Existe una sección en el artículo para tratar las diferencias entre los experimentos. Al igual que el segundo caso, normalmente se empieza describiendo un experimento (base) y luego en la sección de diferencias, se tratan una a una las diferencias, pero esta vez se realiza en función de los elementos utilizados en el experimento. Por ejemplo, para una familia de 3 experimentos (E1, E2 y E3), cuya descripción global se realiza para E1,

posteriormente en la Sección de diferencias, se dice que el elemento *a* cambia para E2 y que el elemento *c* cambia para E3.

Una de las principales limitaciones de estos procesos es que hay diferencias de las que los autores no son conscientes ya que por ejemplo para [EP10], [EP34] y [EP38] los autores mencionan previamente que existe solo una diferencia, pero posteriormente se identifican más:

- En [EP10], los autores afirman que la única diferencia es el tipo de sujetos utilizados, sin embargo, existen además 2 cambios:
 - El material que se escribió originalmente en inglés se tradujo al español y al italiano para las replicaciones correspondientes.
 - Los alumnos son supervisados por un instructor (monitor) diferente dependiendo de la replicación.
- En [EP34], los autores afirman que la única diferencia es el tipo de sujetos utilizados, sin embargo, existen además 5 cambios:
 - Para 3 de los 4 experimentos la asignación de los participantes a los grupos experimentales se realiza utilizando el GPA (Promedio de calificaciones) de los participantes, sin embargo, en el otro experimento la asignación de los participantes se realiza aleatoriamente.
 - Los participantes de uno de los experimentos no asistieron a una lección sobre el método SWEED (que sirve para definir diagramas de caso de uso, que son los objetos experimentales) porque ellos ya lo habían estudiado previamente en el curso de Ingeniería de Software al que asistieron.
 - En uno de los experimentos no se proporcionó el descanso de media hora entre las dos ejecuciones de laboratorio por falta de tiempo, ya que la disposición del laboratorio tenía un tiempo fijo.
 - Los cuestionarios para 3 de los experimentos tenía un número fijo de 10 elementos (preguntas) independientemente del sistema utilizado, sin embargo, para el otro experimento estos cuestionarios tienen un total de 7 elementos que con respuesta abierta en lugar de tipo test.
 - En uno de los experimentos no se pudo recoger la información demográfica sobre los participantes, debido a limitaciones de tiempo que no permitieron a los participantes completar un cuestionario.
- En [EP38], los autores afirman que la única diferencia es el tipo de sujetos utilizados, sin embargo, existen además 3 cambios:
 - El material y los objetos experimentales se renovaron y mejoraron.
 - El material experimental utilizado en el piloto y en uno de los experimentos estaba en italiano, sin embargo, para las replicaciones un hablante nativo de español tradujo todo el material.
 - Tres de los cuatro experimentos fueron realizados por un grupo diferente de experimentadores.

En la tabla 2.7 se puede observar que en la mayoría de los casos las diferencias se identifican a través de la descripción, los siguientes métodos utilizados, pero en menor proporción es mencionar diferencias con respecto al experimento base y describir cada experimento por separado que se usan por igual, el

siguiente es describir los experimentos como un conjunto y el menos utilizado es tener una sección específica para las diferencias.

De todo esto se puede concluir que independientemente de la manera de identificar las diferencias, en ningún caso **los experimentos se comparan sistemáticamente** y por lo tanto la **identificación de diferencias tampoco es sistemática**.

2.1.6 P₃: ¿Cómo se Identifican las Variables a Incluir en el Análisis en una Familia?

A la hora de describir el proceso de identificación de variables de análisis se tendrá en cuenta por una parte cómo se identifican las variables y por otro lado qué variables se identifican.

2.1.6.1 Cómo se Identifican las Variables a Incluir en el Análisis

La forma de identificar las variables de análisis típicamente se realiza en función de las diferencias existentes entre los experimentos o entre los participantes. En la Tabla 2.8 se puede observar cada uno de los estudios en función de cómo se han identificado estas variables.

Identificación de variables de análisis	Estudio Primario	Total
Sin definir	[EP2], [EP3], [EP4], [EP7], [EP8], [EP9], [EP12], [EP13], [EP14], [EP16], [EP17], [EP20], [EP21], [EP22], [EP23], [EP24], [EP26], [EP29], [EP30], [EP31], [EP35], [EP36], [EP37], [EP40], [EP42], [EP43], [EP44], [EP45]	28 (61%)
Ad-hoc	[EP1], [EP5], [EP6], [EP10], [EP11], [EP15], [EP18], [EP19], [EP25], [EP27], [EP28], [EP32], [EP33], [EP34], [EP38], [EP39], [EP41], [EP46]	18 (39%)

Tabla 2.8 Método para la identificación de variables moderadoras

- *No se identifican.* Aunque como ya se ha mencionado anteriormente, siempre existen diferencias, no se identifica ningún tipo de variable.
- *Ad-hoc.* Aunque estas variables representan las diferencias existentes entre los experimentos o participantes de los experimentos no se sabe bajo qué criterio se eligen unas diferencias y otras no.

En la Tabla 2.8 se puede observar que tan solo en un 39% de los estudios primarios se identifican variables. En los estudios que aparecen en azul (15%) desde un principio los autores tenían interés en conocer el posible impacto de una diferencia, por ejemplo, en [EP1] los autores mencionan que su objetivo era analizar el uso de modelos dinámicos con el fin de evaluar si mejora la comprensión de los requisitos funcionales para diferentes categorías de usuarios en términos de experiencia y habilidad.

Por otro lado, **existe un mayor número de diferencias que de variables identificadas**. Por lo tanto, en ninguno de los casos, se transforman todas las diferencias en variables, lo cual era de esperar. Sin embargo, en ningún caso se menciona cuál es el criterio seguido para elegir cuál de estas diferencias se transforman en variables moderadoras. Por lo tanto, **la identificación de variables de análisis no es sistemática ni exhaustiva**.

2.1.6.2 Qué Variables se Identifican

La Tabla 2.9 muestra las variables identificadas en las familias de experimentos. En ninguno de los estudios primarios se identifican variables de diseño y, por tanto, solo se identifican variables moderadoras tanto contextuales (diferencias en la configuración experimental) como variables de participante (diferencias existentes entre los participantes de un mismo experimento). Además, la Tabla 2.9 muestra el conjunto de variables identificadas según su tipo. Nótese que los porcentajes están dados en función de los 18 estudios que identifican variables de análisis, con excepción de la última fila, que se corresponde con los estudios primarios en los que no se identifica ninguna variable y cuyo porcentaje coincide con el mostrado en la Tabla 2.8.

Tipo de Variable	Variable	Estudios Primarios	Total	
De participante	Habilidad (Alta-Baja)	[EP1], [EP10], [EP11], [EP38]	10 (56%)	
	Confianza, Dificultad percibida	[EP15]		
	Experiencia con el dominio	[EP18]		
	Habilidad, Cobertura de código	[EP25]		
	Presión de tiempo, Dificultad de la tarea	[EP27]		
	Experiencia, Percepción de la presión de tiempo, Diferencia en la evaluación de los tratamientos	[EP28]		
	Experiencia, Motivación, Tiempo de estudio	[EP46]		
De contexto	Experiencia (Tipo de sujeto)	[EP5], [EP33]	4 (22%)	
	Perfil del participante (Tipo de sujeto)	[EP34]		
	Lenguaje utilizado	[EP41]		
Ambas	De contexto	Experiencia (Tipo de sujeto)	4 (22%)	
	De participante	Habilidad (Alta-Media-Baja),		[EP6]
		Campo de investigación (pruebas de software u otro)		
	De contexto	Tipo de sujeto		[EP19]
	De participante	Motivación, Conocimiento del patrón, Experiencia en desarrollo		
	De contexto	Experiencia (Tipo de sujetos)		[EP32]
	De participante	Habilidad (Ata-Baja)		
	De contexto	Experiencia (Tipo de Sujeto)		[EP39]
De participante	Tipo de tarea, Conocimiento del dominio, Tamaño del problema			
Ninguna		[EP2], [EP3], [EP4], [EP7], [EP8], [EP9], [EP12], [EP13], [EP14], [EP16], [EP17], [EP20], [EP21], [EP22], [EP23], [EP24], [EP26], [EP29], [EP30], [EP31], [EP35], [EP36], [EP37], [EP40], [EP42], [EP43], [EP44], [EP45]	28 (61%)	

Tabla 2.9 Identificación de variables moderadoras

Se puede observar que en la mayoría de los estudios primarios (56%) solo se identifican variables de participante, por otro lado, en 4 estudios (22%) se identifican variables contextuales y finalmente, en 4 estudios (22%) se identifican tanto variables de contexto como de participante.

Por lo tanto, se puede concluir que **las variables que más se identifican son aquellas que representan diferencias entre los participantes**, sin embargo, es importante recordar que la mayoría de los estudios no identifican ningún tipo de variable.

Por otro lado, en todos los estudios en los que se identifican variables de contexto, excepto uno ([EP41]), la variable identificada es el tipo de sujeto. Por lo tanto, **es una variable de gran interés** en las familias de experimentos de IS.

2.1.7 P₄: ¿Cómo se Analizan Conjuntamente los Experimentos de una Familia?

Finalmente, se quiere saber como se analiza el conjunto de datos obtenido a ejecutar los experimentos que componen una familia. En relación con este punto existen 3 aspectos importantes: 1) La técnica de agregación utilizada, o, cómo se estudian los resultados conjuntamente; 2) la técnica de análisis estadístico utilizada y 3) la estrategia de análisis seguida, o, cómo se incorporan al análisis las variables moderadoras identificadas.

2.1.7.1 Técnica de Agregación

Las distintas técnicas de agregación ya se han estudiado previamente en (Santos, et al., 2018). Sin embargo, en esta tesis se toma en consideración este aspecto, debido a que se ha actualizado la búsqueda. En (Santos, et al., 2018) se identifican 5 técnicas de agregación de experimentos:

- *Síntesis narrativa*. Proporciona un resumen textual de los resultados como una conclusión conjunta. Es difícil distinguir si los autores están agregando los resultados de los experimentos o simplemente comparándolos.
- *Meta-análisis de AD (del inglés Aggregated Data)*. Es la técnica de facto para agregar resultados de experimentos en RSLs. Consiste en combinar los tamaños de efecto ya que este cuantifica la relación entre de dos variables, una dependiente y otra independiente.
- *Meta-análisis de IPD (del inglés Individual Participant Data)*. Consiste en juntar los datos de un conjunto de experimentos y analizarlos conjuntamente como si los datos proviniesen de un único gran experimento.
- *IPD estratificado*. Al igual que meta-análisis de IPD, se juntan los datos de todos los experimentos en un único conjunto de datos, salvo que en este caso se añade una variable que diferencia a los experimentos.
- *Agregación de p-valores*. Agrupa los p-valores de todos los experimentos mediante un método estadístico como Fisher o de Stouffer (Borenstein, et al., 2011).

La Tabla 2.10 muestra las técnicas de agregación utilizadas en cada uno de los estudios primarios. En ella se puede observar que se siguen manteniendo las conclusiones que en (Santos, et al., 2018):

- En más de la mitad (61%) de los estudios primarios se realiza más de una técnica de análisis muchas veces con el objetivo de contrastar los resultados obtenidos.
- La técnica más utilizada es la agregación narrativa, se utiliza en 34 estudios primarios (74%) y 26 de los mismos se complementa con las otras de las técnicas de agregación.
- La siguiente técnica más utilizada es Meta-análisis, se utiliza en 17 estudios primarios (35%) y en 13 los mismos, se complementó con otras técnica agregación.
- La siguiente técnica más utilizada es mega-análisis de IPD, se utiliza en 16 estudios primarios (35%) y en 9 estudios se complementa otras técnicas de agregación. En [EP18]

existen 3 tipos de experimentos, de acuerdo con el dominio estudiado (Diagrama de funciones, Descripciones de gráficos e Interfaz gráfica de usuario), y cada tipo de experimento se realizó dos veces. En ambos casos se realiza agregación narrativa, las replicas de los “tipos” de experimentos se juntan sin distinción y se analizan conjuntamente, por lo tanto, se considera que para cada “tipo” se realiza mega-análisis de IPD.

Estudios Primarios	Total	Síntesis Narrativa	Meta-análisis	Agregación de p-valores	Mega-análisis IPD	IPD estratificado
[EP1], [EP8], [EP9], [EP10], [EP11], [EP14], [EP22], [EP38]	8 (17%)	x	x	-	-	-
[EP2], [EP40], [EP42]	3 (7%)	-	x	-	-	-
[EP3], [EP4], [EP17], [EP23], [EP26], [EP30], [EP37], [EP43]	8 (17%)	x	-	-	-	-
[EP5], [EP18], [EP25], [EP27], [EP32]	5 (11%)	x	-	-	x	-
[EP7], [EP15], [EP44]	3 (7%)	x	x	-	x	-
[EP12], [EP13], [EP20], [EP45], [EP46]	5 (11%)	-	-	-	x	-
[EP16]	1 (2%)	x	-	x	-	-
[EP19]	1 (2%)	-	-	-	x	x
[EP21], [EP28]	2 (4%)	x	x	x	-	-
[EP24]	1 (2%)	-	x	-	x	-
[EP6], [EP29], [EP31], [EP34], [EP36], [EP39]	6 (14%)	x	-	-	-	x
[EP33]	1 (2%)	x	-	-	x	x
[EP35], [EP41]	2 (4%)	-	-	-	-	x

Tabla 2.10 Técnica de agregación utilizada en cada familia

- La siguiente técnica más utilizada es IPD estratificado, se utiliza en 10 estudios primarios (22%) y en 8 de los mismos se complementa con otras técnicas de agregación.
 - En el caso de [EP6], se realizan 3 tipos de experimentos MvR, MvR y MvO, a su vez para los dos primeros existen dos replicas y para estas replicas se añade una variable (el tipo de sujetos) que en su caso distingue entre los experimentos, por lo tanto, en estos dos grupos de experimentos se realiza IPD estratificado.
 - Por otro lado, en [EP29], se realizan 3 experimentos, en los dos primeros los sujetos son estudiantes, mientras que en el tercero son profesionales, se juntan los experimentos con estudiantes y en este caso se añade el experimento como variable, finalmente se realiza agregación narrativa como si fueran dos grupos de experimentos: estudiantes y profesionales.
 - Finalmente, merece la pena mencionar que en los estudios [EP33], [EP34] y [EP39] se realiza IPD estratificado porque cuando se analiza alguna de las variables moderadoras identificadas esta está confundida con la variable experimento ya que toman un valor diferente para cada experimento. De hecho, en [EP39] los autores mencionan que los niveles de la variable tienen el mismo nombre que los experimentos {E1-Prof, E2-UniSA, E3-UniBAS, E4-UniBZ}.
- Finalmente, la técnica menos utilizada es la agregación de p-valores, se utiliza en 3 (7%) estudios primarios, en todos los estudios siempre se complementa con alguna otra técnica de agregación.

De los resultados obtenidos se puede concluir que a pesar de que IPD estratificado es una técnica confiable **en pocos estudios se utiliza**, a pesar de que según (Santos, et al., 2018):

- Las técnicas de agregación utilizadas dentro de las familias pueden afectar la confiabilidad de los resultados conjuntos.
- Las técnicas de agregación confiables, como AD o IPD estratificado, parecen ser adecuadas para analizar familias SE. Sin embargo, su uso tiene un costo: los supuestos estadísticos deben verificarse antes de interpretar los resultados conjuntos.

2.1.7.2 Técnica de Análisis utilizadas

Una vez conocidas las técnicas de agregación utilizadas, en esta sección solo se van a estudiar las técnicas estadísticas asociadas a mega-análisis de IPD e IPD-estratificado ya que en estas dos técnicas se basa el método propuesto en esta tesis.

En la Tabla 2.11 se muestran las técnicas estadísticas utilizadas para cada una de las dos técnicas de agregación, es importante tener en cuenta que existen estudios primarios en los que se realiza más de una prueba estadística. Por ejemplo, de los 16 estudios, cuya técnica de agregación es mega-análisis de IPD, en 10 estudios primarios (63%) se realizan más de una técnica estadística. Y, por otro lado, de los 10 estudios primarios, cuya técnica de agregación es IPD estratificado en 6 de estos (60%) se realiza más de una técnica estadística.

Teniendo en cuenta lo antes mencionado, ahora se va a detallar el uso de cada una de las pruebas estadística para cada una de las técnicas de agregación.

Por un lado, para la técnica de agregación mega-análisis de IPD:

- Las pruebas más utilizadas son las no paramétricas ya que se usan en 12 estudios primarios (75%), de estos en uno se utiliza más de una prueba no paramétrica y en otros 7 se utilizan otras pruebas estadísticas.
 - Mann-Whitney y Kruskal-Wallis son las más utilizadas dentro de las pruebas no paramétricas que se usan por igual en 5 de los estudios primarios (31%). Comparada con la utilización de todas las técnicas estadísticas serían las segundas más utilizadas.
- Las pruebas paramétricas son las siguientes más usadas que se usan en 9 de los estudios primarios (56%), de estos en 8 se utilizan otras técnicas estadísticas.
 - ANOVA es la más utilizada dentro de las pruebas paramétricas que se utiliza en 6 de los estudios primarios (38%), y también es la técnica más utilizada en relación con todas las técnicas estadísticas.
- Finalmente, correlación de Spearman, correlación de Pearson, métodos bayesianos, la prueba del coeficiente Alfa de Cronbach y el Procedimiento de Holm, se usan por igual en un estudio (6%) cada uno de ellos.
 - Correlación de Spearman, Correlación de Pearson y la prueba del coeficiente Alfa de Cronbach se usan en el mismo estudio primario ([EP13]).
 - Tanto en el estudio en el que se han utilizado métodos bayesianos ([EP19]) como en el que se utiliza el procedimiento de Holm ([EP25]), además se usan otras pruebas estadísticas.

Por otro lado, para la técnica de agregación IPD estratificado:

- Al contrario que para el caso anterior, las pruebas paramétricas son las más utilizadas que se usan en 7 estudios primarios (70%), de estos en 3 estudios se utilizan otras pruebas estadísticas.
 - ANOVA es la técnica estadística más utilizada tanto dentro de las pruebas paramétricas, como a nivel general ya que se usa en 4 estudios primarios (40%).
- Las pruebas no paramétricas son las siguientes más utilizadas y se usan en 5 estudios primarios (50%), de estos en dos se usan otras pruebas estadísticas.
 - Las pruebas de permutación son las más utilizadas dentro las pruebas no paramétricas y las segundas más utilizadas a nivel global, se usan en 4 estudios primarios (40%).
- Finalmente, los métodos bayesianos son los menos utilizados ya que se usan solamente en el estudio primario [EP19] (10%). Además, en ese mismo estudio se utiliza otra prueba paramétrica y también se ha utilizado la técnica de agregación mega-análisis de IPD.

En la Tabla 2.11, además se pueden observar las siguientes situaciones:

- Si en un estudio existen varias técnicas de agregación en todas se realizan las mismas pruebas estadísticas ([EP19]).
- La utilización de una prueba estadística depende de si los datos cumplen las restricciones de esta ([EP25]).
- Se contrastan los resultados de las pruebas paramétricas con las no paramétricas o viceversa ([EP32]).

Finalmente, independientemente de la técnica de Agregación, ANOVA es la técnica estadística más utilizada, ya que la misma se utiliza en 10 estudios primarios (42%). Nótese que existen dos estudios primarios ([EP19] y [EP33]) en los que se realizan las dos técnicas de agregación: Mega-análisis de IPD e IPD estratificado.

De los resultados obtenidos se puede concluir que, **no existe una única técnica estadística para la agregación de resultados** y que, aunque se deberían usar **los modelos mixtos muy pocos estudios los utilizan**.

2.1.7.3 Estrategia de Análisis

A la hora de estudiar cómo se analizan las variables moderadoras que se han identificado previamente, se han encontrado 4 formas distintas:

- *Se analiza una única variable.* Solamente se ha identificado una variable y esta se incorpora en el análisis, es importante distinguir este caso, ya que podría decirse que está contenido en cualquiera de los dos siguientes.
- *Se analizan todas las variables identificadas individualmente.* Se realiza un análisis por separado para cada de las variables identificadas.
- *Se analizan todas las variables identificadas conjuntamente.* Se incorporan todas las variables identificadas en un único análisis.

Tipo de Análisis	Estudios Primarios	No paramétricos				Paramétricos				Correlación		Bayesiano (Efectos aditivos)	Prueba del Coeficiente Alfa de Cronbach	Procedimiento de Holm
		Wilcoxon	Mann- Whitney	Kruskal -Wallis	Pruebas de permutación	T-test	GLM	ANOVA	Modelo lineal mixto	Spearman	Pearson			
IPD	[EP5]	-	-	-	X	-	-	-	-	-	-	-	-	-
IPD-E	[EP6]	X	X	-	-	-	X	-	-	-	-	-	-	-
IPD	[EP7]	-		X	-	-	-	-	-	-	-	-	-	-
IPD	[EP12], [EP15], [EP24], [EP44]	-	-	X	-	-	-	X	-	-	-	-	-	-
IPD	[EP13]	-	-	-	-	-	-	-	-	X	X	-	X	-
IPD	[EP18]	X	-	-	-	-	-	-	-	-	-	-	-	-
IPD, IPD-E	[EP19]	-	-	-	-	-	-	-	X	-	-	X	-	-
IPD	[EP20]	-	X	-	-	X	-	-	-	-	-	-	-	-
IPD	[EP25]	X	-	-	-	X	-	X	-	-	-	-	-	X
IPD	[EP27]	-	-	-	-	-	-	-	X	-	-	-	-	-
IPD-E	[EP29], [EP31]	-	-	-	-	-	-	X	-	-	-	-	-	-
IPD	[EP32]	X	X	-	-	X	-	X	-	-	-	-	-	-
IPD	[EP33]	X	X	-	-	-	-	-	-	-	-	-	-	-
IPD-E		-	-	-	-	-	-	X	-	-	-	-	-	-
IPD-E	[EP34], [EP35], [EP39]	-	-	-	X	-	-	-	-	-	-	-	-	-
IPD-E	[EP36]	-	X	-	-	-	-	X	-	-	-	-	-	-
IPD-E	[EP41]	-	-	-	-	-	-	-	X	-	-	-	-	-
IPD	[EP45], [EP46]	-	X	-	-	-	-	-	-	-	-	-	-	-

Tabla 2.11 Técnica de análisis estadístico (IPD e IPD-E)

- *Sin análisis.* No se realiza ningún tipo de análisis solo se muestran los estadísticos descriptivos de las variables identificadas.

En la Tabla 2.12 se puede observar que la mayoría de los estudios primarios (56%) solo identifican una única variable. Nótese que en [EP39] aunque en un principio se habían identificado varias variables moderadoras, finalmente, solo se ha analizado una. La siguiente estrategia más utilizada es el análisis de cada variable por separado, que se realiza en 5 estudios primarios (28%). Finalmente, existen más estudios primarios en los que no se analiza ninguna variable (11%) que los que realizan un análisis conjunto de todas las variables identificadas.

Variables	Estudio Primario	Total
Una única variable	[EP1], [EP5], [EP10], [EP11], [EP18], [EP33], [EP34], [EP38], [EP39], [EP41]	10 (56%)
Individualmente	[EP15], [EP19], [EP25], [EP27], [EP32]	5 (28%)
Conjuntamente	[EP6]	1 (5%)
Sin análisis	[EP28], [EP46]	2 (11%)

Tabla 2.12 Identificación de variables moderadoras

De esto se puede concluir que, **aunque las variables deberían analizarse conjuntamente, ya que de lo contrario se podría estar perdiendo información con respecto a las posibles interacciones entre las mismas**, únicamente un estudio realiza este tipo de estrategia de análisis.

2.1.8 Conclusiones generales

A continuación, en la Tabla 2.13 y la Tabla 2.14 se presenta un resumen de todos los valores obtenidos para las preguntas planteadas anteriormente:

- *P₂: ¿Cómo se identifican las diferencias entre los experimentos de una familia?*
 - Caracterización de los experimentos. El instrumento de caracterización utilizado no está específicamente diseñado para identificar diferencias y similitudes entre los experimentos de una familia.
 - Identificación de diferencias. En ningún caso los experimentos se comparan sistemáticamente y por lo tanto la identificación de diferencias tampoco es sistemática.
- *P₃: ¿Cómo se identifican las variables a incluir en el análisis en una familia?*
 - Identificación de variables que podrían influir en los resultados. La identificación de variables moderadoras no es sistemática ni exhaustiva.
 - El tipo de variables identificadas. En la mayoría de los estudios no se identifica ninguna variable. Las variables identificadas representan diferencias entre los participantes y entre la configuración de los experimentos. Por otro lado, la variable Tipo de Sujeto es una variable de gran interés en las familias de experimentos de IS. En ningún caso se identifican variables de diseño.
- *P₄: ¿Cómo se analizan conjuntamente los experimentos de una familia?*
 - La técnica de agregación utilizada. A pesar de que IPD estratificado es una técnica confiable en pocos estudios se utiliza.

Estudio Primario	Identificación de Diferencias		Identificación de Variables		Análisis		
	Caracterización	Diferencias	Identificación de Variables	Tipo de Variables	Técnica de Agregación	Técnica de Análisis	Estrategia de Análisis
[EP1]	Guías	Experimento base	Ad-hoc	Participante	Síntesis Narrativa, AD	-	Una variable
[EP2], [EP40]	Guías	Descripción	Sin definir	Ninguna	AD	-	N/A
[EP3], [EP26]	Guías	Experimentos por separado	Sin definir	Ninguna	Síntesis Narrativa	-	N/A
[EP4]	Guías	Experimento base	Sin definir	Ninguna	Síntesis Narrativa	-	N/A
[EP5]	Guías	Descripción	Ad-hoc	Contexto	Síntesis Narrativa, IPD	Permutación	Una variable
[EP6]	Guías	Descripción	Ad-hoc	Participante, Contexto	IPD-E	Wilcoxon, Mann-Whitney, GLM	Conjunto
[EP7]	Guías	Descripción	Sin definir	Ninguna	Síntesis Narrativa, AD, IPD	Kruskal-Wallis	N/A
[EP8]	Informal	Experimentos por separado	Sin definir	Ninguna	Síntesis Narrativa, AD	-	N/A
[EP9], [EP14]	Guías	Experimento base	Sin definir	Ninguna	Síntesis Narrativa, AD	-	N/A
[EP10], [EP11], [EP38]	Guías	Descripción	Ad-hoc	Participante	Síntesis Narrativa, AD	-	Una variable
[EP12]	Informal	Descripción	Sin definir	Ninguna	IPD	Kruskal-Wallis, ANOVA	N/A
[EP13]	Informal	Conjunto	Sin definir	Ninguna	IPD	Spearman, Pearson, Coeficiente Alfa de Cronbach	N/A
[EP15]	Guías	Experimento base	Ad-hoc	Participante	Síntesis Narrativa, AD, IPD	Kruskal-Wallis, ANOVA	Individual
[EP16]	Informal	Conjunto	Sin definir	Ninguna	Síntesis Narrativa, Agregación de p-valores	-	N/A
[EP17], [EP23], [EP43]	Informal	Experimentos por separado	Sin definir	Ninguna	Síntesis Narrativa	-	N/A
[EP18]	Guías	Descripción	Ad-hoc	Participante	Síntesis Narrativa IPD	Wilcoxon	Una variable
[EP19]	Informal	Descripción	Ad-hoc	Participante, Contexto	IPD, IPD-E	Modelo lineal mixto, Bayesiano	Individual
[EP20]	Guías	Descripción	Sin definir	Ninguna	IPD	Mann-Whitney, T-test	N/A
[EP21]	Informal	Descripción	Sin definir	Ninguna	Síntesis Narrativa, AD, Agregación de p-valores	-	N/A

Tabla 2.13 Resumen de los valores para las preguntas P₂, P₃ y P₄ (1/2)

Estudio Primario	Identificación de Diferencias		Identificación de Variables		Análisis		
	Caracterización	Diferencias	Identificación de Variables	Tipo de Variables	Técnica de Agregación	Técnica de Análisis	Estrategia de Análisis
[EP21]	Informal	Descripción	Sin definir	Ninguna	Síntesis Narrativa, AD, Agregación de p-valores		N/A
[EP22]	Informal	Experimento base	Sin definir	Ninguna	Síntesis Narrativa, AD	-	N/A
[EP24]	Guías	Descripción	Sin definir	Ninguna	AD, IPD	Kruskal-Wallis, ANOVA	N/A
[EP25]	Guías	Descripción	Ad-hoc	Participante	Síntesis Narrativa, IPD	Wilcoxon, T-test, ANOVA	Individual
[EP27]	Plantilla	Experimento base	Ad-hoc	Participante	Síntesis Narrativa, IPD	Modelo lineal mixto	Individual
[EP28]	Informal	Sección	Ad-hoc	Participante	Síntesis Narrativa, AD, Agregación de p-valores	-	Sin análisis
[EP29]	Informal	Descripción	Sin definir	Ninguna	IPD-E	ANOVA	N/A
[EP30]	Guías	Conjunto	Sin definir	Ninguna	Síntesis Narrativa	-	N/A
[EP31]	Guías	Sección	Sin definir	Ninguna	IPD-E	ANOVA	N/A
[EP32]	Guías	Descripción	Ad-hoc	Participante, Contexto	Síntesis Narrativa, IPD	Wilcoxon, Mann-Whitney, T-test, ANOVA	Individual
[EP33]	Guías	Conjunto	Ad-hoc	Contexto	Síntesis Narrativa, IPD, IPD-E	Wilcoxon, Mann-Whitney, ANOVA	Una variable
[EP34]	Informal	Descripción	Ad-hoc	Contexto	IPD-E	Permutación	Una variable
[EP35]	Guías	Descripción	Sin definir	Ninguna	IPD-E	Permutación	N/A
[EP36]	Guías	Descripción	Sin definir	Ninguna	IPD-E	Mann-Whitney, ANOVA	N/A
[EP37]	Informal	Descripción	Sin definir	Ninguna	Síntesis Narrativa	-	N/A
[EP39]	Informal	Sección	Ad-hoc	Participante, Contexto	IPD-E	Permutación	Una variable
[EP41]	Guías	Descripción	Ad-hoc	Contexto	IPD-E	Modelo lineal mixto	Una variable
[EP42]	Guías	Experimentos por separado	Sin definir	Ninguna	AD	-	N/A
[EP44]	Informal	Descripción	Sin definir	Ninguna	Síntesis Narrativa, AD, IPD	Kruskal-Wallis, ANOVA	N/A
[EP45]	Guías	Descripción	Sin definir	Ninguna	IPD	Mann-Whitney	N/A
[EP46]	Informal	Descripción	Ad-hoc	Participante	IPD	Mann-Whitney	Sin análisis
Proceso Propuesto	Esquema	Sistemática	Sistemática	Diseño, Participante, Contexto	IPD + IPD-E	Modelo lineal mixto	Conjunto

Tabla 2.14 Resumen de los valores para las preguntas P₂, P₃ y P₄ (2/2)

- La técnica de análisis asociada a la técnica de agregación. No existe una única técnica estadística para la agregación de resultados y aunque se deberían usar los modelos mixtos muy pocos estudios los utilizan.
- La estrategia de análisis utilizada. La estrategia de análisis más adecuada es analizar las variables conjuntamente, sin embargo, únicamente un estudio la utiliza.

En la Tabla 2.14 además de los valores que toman cada uno de los estudios, la última fila de la muestra los valores que se proponen en el proceso descrito en este documento.

Por lo tanto, tal y como se ha podido observar:

- **No existe un método que resuelva de manera genérica y sistemática** el problema de agregar los experimentos de una familia.
- Los experimentos que conforman una familia no son replicaciones exactas, **siempre hay diferencias entre los mismos**, por lo tanto, es importante identificar y explorar la existencia de variables moderadoras que podrían influir en los resultados.
- **No existe un marco formal para identificar las similitudes y diferencias entre experimentos.** Aunque en todos los estudios primarios se empieza con la descripción o caracterización de los experimentos, esta no sigue un patrón específico y esto hace que se omitan varios detalles de los experimentos. A través del mero reporte de los experimentos se pueden identificar similitudes y diferencias entre los experimentos, aunque este no sigue un enfoque formalizado. Por lo tanto, es necesario sistematizar la identificación de diferencias y similitudes entre los experimentos.
- **No todos los estudios tienen en cuenta variables**, y en los estudios en los que se identifican, se considera un menor número de variables y no se sabe qué criterio se ha seguido para descartar las diferencias que finalmente no aparecen como variables a incluir en el análisis. Por lo tanto, hay que formalizar y sistematizar la extracción de variables.
- **Existen variables identificadas que finalmente no son analizadas**, en algunas ocasiones, solo se proporciona los estadísticos descriptivos de las mismas. Cuando se realiza el análisis de dichas variables si existe más de una variable se realizan análisis independientes para cada una o se analizan todas conjuntamente.
- **Estas estrategias no son las más adecuadas para la agregación.** Por lo tanto, hay que definir una estrategia de análisis que contemple la adecuada introducción de todas las variables moderadoras y utilizar una técnica que de soporte a la estrategia de análisis.
- De los resultados obtenidos del estado de la práctica se puede concluir que **no existe una solución apropiada para el problema descrito.**

2.2 Trabajos Relacionados

Se han encontrado únicamente dos trabajos que traten sobre el tema de la agregación de experimentos

2.2.1 Agregación Cuantitativa de Evidencia de Experimentos Controlados

Ciolkowski presenta un enfoque para la agregación cuantitativa de agregar resultados experimentales en IS. En él, adapta el meta-análisis de AD a la IS, donde la mayoría de los estudios presentan gran heterogeneidad en los datos y no se dispone de toda la información (Ciolkowski, 2009) (Ciolkowski, 2012). Este enfoque permite la agregación de la evidencia obtenida con las revisiones sistemáticas de la literatura y consta de cinco pasos:

1. *Definición*: Definir los objetivos, variables dependientes e independientes, criterios de inclusión para estudios.
2. *Selección de estudios*: recopilar estudios primarios de forma sistemática, de acuerdo con objetivos definidos; evaluar la idoneidad con respecto a los criterios de inclusión, extraer datos.
3. *Evaluación de la calidad del estudio*: además de los criterios generales de inclusión y exclusión, generalmente se considera importante evaluar la "calidad" de los estudios primarios.
4. *Extracción de datos*: diseñar formularios de extracción de datos para registrar con precisión la información que los investigadores obtienen de los estudios primarios; principalmente la información necesaria para calcular el tamaño del efecto de los estudios individuales.
5. *Síntesis de datos*: recopilar y resumir los resultados de los estudios primarios incluidos; se utiliza metanálisis para combinar cuantitativamente los tamaños de los efectos de los estudios primarios. Los subpasos importantes son:
 - a. *Calcular un único tamaño del efecto* de los estudios individuales.
 - b. *Agregar o combinar tamaños de efectos*: calcular el efecto combinado, examine los niveles de significancia y evaluar la importancia del tamaño del efecto combinado
 - c. *Evaluar la heterogeneidad*: evaluar si los efectos son consistentes en dirección y magnitud.
 - d. *Tratar con la heterogeneidad*: la heterogeneidad indica que puede haber influencias adicionales en el efecto. Las opciones son ignorar, incorporar o explicar las diferencias.
 - e. *Visualizar y evaluar información*: Finalmente, la información agregada debe ser interpretada, visualizada y evaluada para varios interesados.

2.2.2 Agregación de Familias de Experimentos

Únicamente existe un estudio que investiga específicamente la agregación de familias de experimentos (Santos, et al., 2019) en el que proponen un método para analizar familias de experimentos compuesto de 4 fases:

1. *Descripción de los participantes*. Su objetivo es informar sobre la población y conocer las posibles fuentes de heterogeneidad.
2. *Analizar las replicaciones individuales*. Ayuda a identificar patrones y evita la heterogeneidad debida a las diferentes técnicas de análisis utilizadas
3. *Agregar resultados*. Aumenta la confiabilidad de las conclusiones conjuntas

4. Conducir un análisis exploratorio. Su objetivo es identificar variables moderadoras tanto a nivel de experimento como a nivel de participante.

Además (Santos, et al., 2019), han establecido 6 guías para los dos últimos pasos. Las guías para agregar resultados son:

- Guía 1. Indica que se debe evitar el uso de agregación narrativa y la combinación de p-valores para proporcionar conclusiones conjuntas
- Guía 2. Recomendamos evitar Mega-análisis de IPD, es decir, juntar los datos como si se tratase de un único gran experimento.
- Guía 3. Recomendamos fomentar el uso de Meta-análisis de AD e IPD estratificado (teniendo el experimento como una variable en el análisis).

Relacionadas con el análisis exploratorio las pautas son:

- Guía 4. Indica el estudio de variables moderadoras a nivel de experimento mediante el uso de Meta-análisis de AD e IPD estratificado.
- Guía 5. Indica el estudio de variables moderadoras a nivel de participante mediante el uso de IPD estratificado.
- Guía 6. Indica que se deben reconocer las limitaciones de los análisis exploratorios.

2.2.3 Resumen

La Tabla 2.15 realiza una comparativa entre el proceso propuesto por Ciolkowski (Ciolkowski, 2012) y Santos (Santos, et al., 2019), y el que aquí se propone.

	Identificación de Diferencias		Identificación de Variables		Análisis		
	Caract.	Diferencias	Identificación de Variables	Tipo de Variables	Técnica de Agregación	Técnica de Análisis	Estrategia de Análisis
(Ciolkowski, 2012)	No existe	No existe	No existe	Contexto	AD	Efectos Aleatorios	Conjunto
(Santos, et al., 2019)	No existe	No existe	No existe	Contexto	AD	Efectos Aleatorios	Conjunto
				Participante	IPD-E	Modelo lineal mixto	Conjunto
Proceso Propuesto	Esquema	Sistemática	Sistemática	Diseño, Contexto, Participante.	IPD + IPD-E	Modelo lineal mixto	Conjunto

Tabla 2.15 Resultados para el estado del arte

Por lo tanto, tal y como se ha podido observar:

- **No existe un método que resuelva de manera genérica y sistemática** el problema de agregar los experimentos de una familia.
- **No existe un marco formal para identificar las similitudes y diferencias entre experimentos.** En ningún caso los experimentos se caracterizan y por lo tanto no se identifican diferencias.

- **No se identifican todos los tipos de variables**, el procedimiento de (Santos, et al., 2019) solo permite agregar experimentos con el mismo diseño, ya que no se identifican este tipo de variables. En ambos casos, se indica como analizar las variables moderadoras, pero no se sabe como se seleccionan.
- **Se realizan análisis separados para cada tipo de variables identificadas**, Santos identifica dos tipos de variables y propone dos análisis distintos para cada tipo, por lo tanto, no se analizan todas las variables identificadas conjuntamente.
- De los resultados obtenidos del estado del arte se puede concluir que **no existe una solución apropiada para el problema descrito**.

Finalmente, dado que ni el estado de la práctica ni el estado del arte proporcionan una solución apropiada para este problema, el problema que se pretende resolver está abierto y tiene sentido.

Capítulo 3

Planteamiento del Problema

En este capítulo se van a estudiar las principales dificultades encontradas en el Capítulo 2 y a partir de ellas se plantean las hipótesis y objetivos de investigación de esta tesis, así como la metodología de trabajo a seguir.

3.1 Descripción del Problema

Según se ha visto en el Capítulo 2, el problema a resolver *es la falta de un proceso para la agregación de familias de experimentos en IS que permita el estudio sistemático de variables moderadoras*. Este problema se ha dividido en 3 subproblemas, los cuales se describen en detalle a continuación:

1. La identificación de similitudes y diferencias entre los experimentos de una familia no es precisa.
2. El establecimiento del conjunto de variables a incluir en el análisis (tanto moderadoras como de diseño) no se realiza de forma sistemática.
3. El análisis de datos no tiene en cuenta todas las variables y se realizan análisis separados para cada una de ellas.

Por otra parte, en el Capítulo 2 se ha identificado como variable moderadora más estudiada el tipo de sujeto. Por lo tanto se va a realizar un *estudio de la adecuación de la etiqueta estudiante/profesional para distinguir entre distintos tipos de sujetos en experimentos*.

En lo que se refiere a la **identificación imprecisa de diferencias y similitudes entre los experimentos de una familia**, en el Capítulo 2 se ha visto que no existe un instrumento que permita la caracterización exhaustiva de los experimentos que conforman una familia y, por lo tanto, sea posible identificar de manera precisa las similitudes y diferencias entre ellos. Hasta la fecha, la identificación se realiza de manera informal. En el mejor de los casos se realiza la descripción del experimento base (experimento inicial) siguiendo las guías y recomendaciones de otros autores para el reporte de experimentos, y sus replicaciones se detallan en base a las diferencias existentes con respecto al experimento base. Las guías más utilizadas por las familias son las de (Wohlin, et al., 1999). Debido a que el instrumento utilizado no es el adecuado, en ciertas ocasiones se omiten detalles entre los experimentos que pueden ser clave a la hora de identificar diferencias y similitudes. Como consecuencia de la omisión de información, en muchos

de los casos lo autores consideran menos cambios entre los experimentos de los que han realizado. Es decir, muchas veces ni siquiera son conscientes de todas las diferencias que existen entre los experimentos.

En lo que se refiere al **establecimiento sistemático de un conjunto de variables de análisis (moderadoras y de diseño)** se ha podido observar que no siempre los cambios identificados entre los experimentos de cada familia pasan a ser considerados como variables moderadoras. Las variables identificadas representan algunos de los cambios existentes entre los experimentos o participantes de cada experimento y en ningún caso se identifican variables de diseño. Sin embargo, esta identificación se realiza de forma circunstancial. Como consecuencia se están perdiendo posibles variables que pueden ser relevantes en los resultados.

En lo que se refiere al **análisis conjunto de las variables identificadas**, la agregación debería realizarse mediante IPD-E y utilizando modelos lineales mixtos. No obstante, una cantidad ínfima (0,04%) de familias de experimentos realiza el análisis de este modo. Asimismo, el análisis realizado a menudo no incluye las variables moderadoras identificadas previamente. Finalmente, se realizan análisis separados para cada variable moderadora identificada. Como consecuencia de todo esto, se desconoce el impacto de las variables moderadoras no estudiadas, así como de las posibles interacciones entre las variables tanto las estudiadas como las no estudiadas.

Finalmente, en lo que se refiere al **estudio de la etiqueta estudiante/profesional** para identificar tipos de sujetos en los experimentos, la variable tipo de sujeto podría ser muy genérica y no ser capaz de recoger todas las características de un participante. Como consecuencia se puede estar perdiendo información sobre las características o habilidades de los participantes que podrían estar influyendo en los resultados. Al considerar estas características los resultados que se obtengan pueden ser más exactos.

3.2 Hipótesis y Objetivos de la Investigación

El objetivo del presente trabajo consiste en definir un proceso de agregación de los experimentos pertenecientes a una familia, en el que se consideren las variaciones o diferencias entre los mismos ya que estas pueden afectar en los resultados. Para conseguir este objetivo, se han definido 4 subobjetivos:

O₁: Creación de un modelo de caracterización de experimentos.

O₂: Identificación de variables de análisis.

O₃: Análisis conjunto de los experimentos.

O₄: Estudio del impacto del uso de estudiantes y profesionales en las familias de experimentos.

La hipótesis de esta tesis es:

Se puede formalizar y crear un proceso de agregación de familias de experimentos, de modo que facilite y permita la inclusión de las modificaciones o alteraciones realizadas entre los experimentos de la familia.

En la Figura 3.1 se puede observar que los objetivos planteados ayudan a la resolución de los principales problemas identificados.

Figura 3.1 Problemas y Objetivos de Investigación

3.2.1 Creación de un Modelo de Caracterización de Experimentos

La caracterización de los experimentos de una familia permite identificar de manera sencilla las diferencias existentes entre los experimentos; por otro lado, permite conocer si los experimentos son comparables entre sí y si se pueden agregar. Por lo tanto, es necesario crear un modelo o esquema que permita incluir todas las características que conlleva la ejecución de un experimento. Este esquema deberá identificar cada experimento de la familia y tiene que ser exactamente el mismo para todos los experimentos de la familia. De esta premisa se desprende el primer objetivo de este trabajo:

O₁: Crear un modelo de caracterización de experimentos.

Este objetivo dará solución al problema de la no existencia de un método único de caracterización y a la dificultad inherente de identificar las diferencias entre los experimentos.

3.2.2 Identificación de un Conjunto de Variables de Análisis

Existen 2 tipos de variables de análisis: de diseño y moderadoras. A su vez existen dos tipos de variables moderadoras: de contexto y de participante. Cada una de ellas aglutina diferentes tipos de información que debe ser tenida en cuenta.

Con respecto a las variables de diseño, debido a que puede ocurrir que el diseño experimental cambie entre los experimentos que conforman una familia, ya sea por corregir algún problema (por ejemplo, que los tratamientos estén confundidos con la sesión) o por adaptación del experimento al contexto (por ejemplo, no es posible estudiar todos los tratamientos por limitaciones de tiempo). Por lo

tanto, es necesario estudiar la manera de identificar las diferencias entre los diseños como variables de diseño.

Con la caracterización de experimentos propuesta en el objetivo O₁, se podrán identificar las variables contextuales, las cuales representan las similitudes y diferencias entre los experimentos debidas a los cambios en el contexto. Sin embargo, es necesario estudiar cuales de sus características podrían tener un impacto en los resultados, para de esta manera transformar dichas diferencias a variables que se puedan estudiar estadísticamente.

Por otro lado, gracias a la información demográfica de los participantes recopilada mediante cuestionarios, durante la ejecución de los experimentos se pueden conocer las principales diferencias entre participantes. Ya que, por ejemplo, muchas veces se cree que, los estudiantes son parecidos entre sí y los profesionales son parecidos entre sí, lo que podría ser un error. Al poder estudiar las diferencias entre cada uno de los participantes se podrá saber cuan similares son en realidad y cuales de estas diferencias se podrán considerar porque tengan un impacto en los resultados de la familia de experimentos para poder incluirlas posteriormente en el análisis.

De todo lo mencionado se desprende el segundo objetivo:

O₂: Identificación de variables de análisis

Este objetivo dará solución al problema de que se identifican menos variables que las que realmente podrían influir en los resultados.

3.2.3 Análisis Conjunto de los Experimentos

Se necesita una técnica de agregación que permita la inclusión de las variables de diseño y moderadoras identificadas. Como se ha mencionado en el Capítulo 2, y se plantea en los objetivos O₂ las variables moderadoras pueden representar diferencias entre participantes o entre experimentos. Por lo tanto, este objetivo es:

O₃: Analizar los experimentos conjuntamente

Este objetivo permitirá comparar todos los experimentos de una familia y nos ayudará a identificar si las diferencias tanto a nivel de participante como las de experimento están influyendo en los resultados, es decir, se podrá saber si los experimentos son realmente diferentes entre sí o si estas diferencias son propias de los participantes.

3.2.4 Estudio del impacto del tipo de sujeto

Se necesita conocer si la variable moderadora tipo de sujeto, es suficiente para reflejar toda la información relevante de los participantes. Esto da lugar el siguiente objetivo:

O₄: Estudiar el impacto del uso de estudiantes y profesionales en las familias de experimentos

Este objetivo permitirá comprender si se debe utilizar la etiqueta estudiantes y profesionales o las características inherentes entre los participantes.

3.3 Metodología de Trabajo

A continuación, se describe la metodología de trabajo seguida para poder alcanzar los objetivos previamente mencionados.

Para la caracterización de cada uno de los experimentos de la familia se necesita un esquema de caracterización que permita reflejar la mayor cantidad de detalles sobre los experimentos ya que esto ayudará a su descripción y facilitará la posterior comparación entre los mismos. Se parte del esquema para caracterizar una situación experimental propuesta por Gómez (Gómez, 2012). Dicho esquema está pensado para identificar tipos de replicaciones y así determinar cual es el tipo de estas (exacta, diferenciada, interna, etc.). Este esquema se va a utilizar para caracterizar tanto las 46 familias de experimentos encontradas en el mapeo sistemático de la literatura descrito en el Capítulo 2 de este trabajo. Gracias a esta instanciación va a ser posible modificar el esquema de caracterización para adaptarlo a su nuevo uso: identificar similitudes y diferencias entre los experimentos de una familia.

Una vez creado el esquema, es necesario establecer la forma en la que se van a identificar las variables que se incluirán en el análisis. El primer paso consiste en identificar las variables asociadas al diseño experimental. Todos los experimentos de la familia pueden tener el mismo diseño, en cuyo caso las variables de diseño para la agregación coincidirán con las de cada diseño individual. Para el caso de que los experimentos de la familia tuviesen distinto diseño, se define un diseño para la agregación a partir de la superposición de todos los diseños de cada experimento de la familia. Se crean una serie de reglas que permiten identificar las variables de diseño a incluir en el análisis conjunto de los experimentos, a partir del diseño agregado.

Una vez identificadas las variables asociadas al diseño experimental, se debe definir la forma de identificar sistemáticamente las variables moderadoras, tanto a nivel de experimento como a nivel de participante.

Finalmente, para el análisis estadístico del conjunto de todos los experimentos mediante IPD-E y modelos mixtos se propone el uso de modelos jerárquicos. Para ello se utiliza el método propuesto en (Heck, et al., 2014).

A continuación, es necesario validar el método propuesto. Esto se hace con dos familias de experimentos diferentes. La primera familia, la más pequeña, consta de 4 experimentos que estudian la efectividad de tres técnicas de evaluación de código y la segunda familia consta de 16 experimentos que estudian la calidad del código escrito por los desarrolladores de software al utilizar distintos métodos de desarrollo: TDD e ITL.

Finalmente, las variables moderadoras más estudiadas en las familias de experimentos del Capítulo 2 son el tipo de sujeto o alguna de las características de los participantes y rara vez se añaden otras variables, por lo tanto, es necesario estudiar si la etiqueta estudiantes profesionales es apropiada para caracterizar a los participantes en experimentos de IS. Para ello se va a realizar un mapeo sistemático de la literatura donde se mira la frecuencia con la que se usa cada tipo de participante, la manera de caracterizar a los participantes (variable Tipo de Sujeto o conjunto de características) y finalmente, se va a comprobar el impacto de sus características en función del análisis realizado. Además, en la segunda familia de experimentos utilizada para evaluar el método propuesto, se va a estudiar la influencia en los resultados de la variable tipo de sujeto y las características de los participantes.

Capítulo 4

Propuesta de Proceso para la Agregación de Familias de Experimentos

En este capítulo se explica el proceso que se propone para agregar los experimentos de una familia. Los contenidos de este capítulo aparecen en un artículo que se está preparando para enviar a una revista JCR Q1:

- P. Riofrío, S. Vegas, N. Juristo. “A Method for Aggregating Families of Experiments in Software Engineerings”. Para enviar a Empirical Software Engineering Journal. Springer.

4.1 Visión General del Proceso Propuesto

El proceso propuesto, ilustrado en la Figura 4.1, consta de cinco pasos:

1. *Identificación de los grupos de agregación.* El proceso de agregación definido requiere que los constructos de los experimentos a agregar sean similares (debido a la técnica de análisis estadístico empleada). El objetivo de este paso consiste, por tanto, en identificar dentro de la familia de experimentos aquellos que tengan constructos similares. Como resultado, se obtienen uno o más grupos de replicaciones a agregar. Este paso consta de dos actividades:
 - 1.1. *Caracterización de los experimentos de la familia.* Todos y cada uno de los experimentos de la familia se caracterizan siguiendo el mismo esquema. Esto tiene un doble objetivo: 1) permitir conocer todos los detalles de cada experimento y 2) facilitar la comparación entre ellos.
 - 1.2. *Identificación de grupos de experimentos a agregar.* Dado que la agregación requiere que los experimentos a agregar posean constructos similares, en esta actividad se comparan los constructos de todos los experimentos de la familia, y se determina si el nivel de similitud entre ellos es suficiente. Aquellos experimentos que posean constructos similares serán asignados al mismo grupo; aquellos que no lo sean, formarán parte de grupos distintos.

Figura 4.1 Proceso para la agregación de familias de experimentos propuesto

2. *Identificación de variables de diseño.* El proceso permite la agregación de experimentos con distintos diseños. El objetivo de este paso consiste en la identificación de las variables que se deben incorporar al modelo de análisis, y que vienen impuestas por el diseño experimental (variables de diseño). Este paso consta de las siguientes actividades:

2.1. *Creación de un diseño agregado.* Si todos los experimentos del grupo poseen el mismo diseño, el diseño agregado coincide con el diseño de cada experimento individual. No obstante, si hay experimentos con distinto diseño, debe establecerse el diseño global o agregado. Este diseño se obtendrá a partir de la superposición de los distintos diseños y a partir de unas reglas que se han definido con tal fin.

2.2. *Exploración de los resultados de análisis individuales.* Cada experimento debe ser analizado de forma individual, y los resultados de todos ellos comparados en unos términos precisos definidos a tal efecto.

2.3. *Identificación del conjunto de variables de diseño.* Las variables de diseño vienen definidas por el diseño agregado definido en la actividad 2.1. No obstante, dado que el proceso de agregación incorpora variables en distintos momentos y podemos terminar analizando un gran número de variables, es necesario estudiar si es posible reducir el número de variables planteadas inicialmente. Esto tiene especial sentido en el caso de grupos de experimentos con distintos diseños. La posible reducción se realizará siguiendo una serie de reglas preestablecidas teniendo en cuenta los resultados obtenidos en la actividad 2.2.

3. *Identificación de variables contextuales.* El proceso permite identificar de manera sistemática las diferencias existentes entre los experimentos de una familia. El objetivo de

este paso consiste en la identificación de variables a incorporar al modelo de análisis, y que son resultado de las diferencias contextuales entre los distintos experimentos. Este paso consta de las siguientes actividades:

- 3.1. *Identificación de diferencias contextuales.* Se representan las diferencias contextuales de los experimentos de un mismo grupo considerando todos los elementos que componen una situación experimental (excepto los elementos del diseño).
- 3.2. *Creación de un conjunto inicial de variables contextuales.* Se crea un conjunto inicial de variables en función de si los cambios realizados en las características de los experimentos de la familia puedan tener impacto en los resultados de la agregación. Luego, se transforman las características a variables, considerando que no necesariamente va a existir una correspondencia uno a uno entre características y variables.
- 3.3. *Refinamiento del conjunto inicial de variables contextuales.* A partir del conjunto de variables identificado en la actividad 3.2, se evalúa si pueden estudiarse todas las variables ya que estás podrían estar confundidas con otras. Además, se estudia el balanceo de las variables seleccionadas. En caso de no resultar balanceadas, se puede considerar su eliminación del modelo.
4. *Identificación de variables de participantes.* Mediante la información recogida durante la ejecución cada experimento sobre los participantes se identifican las variables a considerar a nivel de participante. Este paso consta de las siguientes actividades:
 - 4.1. *Identificación de diferencias entre participantes.* Se recopila toda la información recogida sobre los participantes en cada uno de los experimentos.
 - 4.2. *Creación de un conjunto inicial de variables de participantes.* Se crea un conjunto inicial de variables. Únicamente se podrán considerar aquellas que se han medido en todos los experimentos con la misma escala.
 - 4.3. *Refinamiento del conjunto inicial de variables de participantes.* Se refina este conjunto dependiendo de si no están confundidas entre sí o con las variables contextuales.
5. *Realización del análisis conjunto.* El análisis de cada uno de los grupos de experimentos se lleva a cabo siguiendo la teoría de modelado multinivel (reconocen la existencia de una jerarquía en los datos). El objetivo consiste en obtener resultados conjuntos para cada grupo de experimentos. Este paso consta de las siguientes actividades:
 - 5.1. *Identificación del modelo de análisis.* Se establece la jerarquía del grupo de experimentos tomando como base sus características. A continuación, se asignan a cada nivel cada una de las variables identificadas en los pasos 2, 3 y 4 de este proceso. Asimismo, se establece la escala en la que se han de medir las variables.
 - 5.2. *Análisis del modelo.* Siguiendo un enfoque iterativo basado en las técnicas de análisis multinivel, se procede al análisis estadístico de cada uno de los grupos de experimentos identificados.
 - 5.3. *Interpretación de los resultados.* Se procede a la interpretación de los resultados obtenidos por el modelo final alcanzado en la actividad anterior. Esta interpretación es similar a la que se lleva a cabo cuando se analizan experimentos aislados, aunque tiene ciertas peculiaridades (debido al análisis conjunto de varios experimentos) que deben ser tenidas en cuenta.

- 5.4. *Interpretación conjunta de los resultados de los distintos grupos de experimentos.* En caso de que se hayan obtenido distintos grupos de experimentos, será necesario hacer una interpretación conjunta de los mismos.

4.2 Identificación de los Grupos de Agregación

4.2.1 Caracterización de los Experimentos de la Familia

Para caracterizar los experimentos de una familia, es necesario utilizar un esquema de caracterización. La sección 4.2.1.1 presenta el esquema utilizado. La sección 4.2.1.2 incluye un ejemplo de uso del esquema. Finalmente, la sección 4.2.1.3 introduce una herramienta que se ha desarrollado con el fin de facilitar la caracterización.

4.2.1.1 Esquema de Caracterización Utilizado

La caracterización de cada uno de los experimentos de la familia ayuda a su descripción con el fin de facilitar su posterior comparación. Por tanto, es esencial que el esquema de caracterización utilizado permita reflejar la mayor cantidad de detalles sobre los experimentos.

En su tesis doctoral, Gómez (Gómez, 2012) propone un esquema para caracterizar la situación experimental de una serie de replicaciones con el fin de identificar el tipo de replicaciones que se están realizando. El esquema propuesto por Gómez aparece ilustrado en la Tabla 4.1.

Se utiliza como punto de partida el esquema propuesto por Gómez para caracterizar cada una de las 46 familias de experimentos que se han discutido en el Capítulo 2 de esta tesis. Los resultados de dicha caracterización aparecen en la carpeta resultados del repositorio de GitHub <https://github.com/GRISE-UPM/Tesis-Riofrio>.

La caracterización de estas familias ha mostrado que la propuesta de Gómez necesita ser ampliada para la caracterización de familias de experimentos. La Tabla 4.2 resume los ajustes que se han producido en el esquema propuesto por Gómez.

Por una parte, ha sido necesario añadir nuevos elementos para acomodar características de los experimentos que eran necesarias. Por otra parte, se ha producido un reajuste de los contenidos de las dimensiones, moviendo elementos entre ellas. La cuarta columna de la Tabla 4.2 representa la acción realizada en cada uno de los elementos. Para los elementos que no son nuevos, en la última columna se indica la dimensión a la que pertenecían inicialmente (nótese que el protocolo experimental se ha renombrado y pasa a llamarse ejecución experimental). Finalmente, aparecen casos en los que se renombró el elemento; en este caso, el antiguo nombre aparece tachado y el nuevo a su lado. Por lo tanto, los cambios realizados son:

- *Añadir nuevos elementos.* Se ha observado que cierta información que aparecía en los experimentos no podía acomodarse en el esquema. En concreto, ha ocurrido esto con los elementos:
 - *Sitio.* Permite conocer, si los experimentos se han realizado en el ámbito académico o industrial.
 - *Publicación.* Permite conocer si un experimento se ha publicado individualmente.

Dimensión		Elemento
Operacionalización	Causa	Tratamiento Tratamiento acorde a la literatura de SE Procedimiento de aplicación del tratamiento Tratamiento como instrucciones Recursos del tratamiento Aspectos de transmisión del tratamiento
	Efecto	Variable respuesta Operacionalizaciones de efecto Métrica Procedimiento de medición (criterio de corrección)
Protocolo experimental	Análisis	Técnica de análisis de datos
	Diseño	Diseño experimental Orden de aplicación de los tratamientos Combinación entre el entrenamiento y las sesiones Tratamientos por factor por sesión Sesiones requeridas por el sujeto para aplicar el tratamiento Tareas por sujeto por sesión
	Instrumentación	Instrumentos o equipo de medición
	Objetos experimentales	Programas Faltas
	Materiales	Formularios del experimento Instrucciones del experimento Guías suplementarias
Población	Propiedades de los sujetos	Tipo de sujeto Experiencia con el tratamiento Otras condiciones de los sujetos
	Propiedades de los objetos experimentales	Lenguaje de programación Tamaño Complejidad ciclomática Tipo de faltas
Experimentadores y Sitio		Diseñador Entrenador Monitor Medidor Analista Sitio

Tabla 4.1 Propuesta de esquema de caracterización para replicaciones (Gómez, 2012)

- *Factores*. Permite conocer el factor principal de la familia y cuando hay más de un factor permite diferenciar que tratamientos corresponden a cada uno.
- *Factores por sesión*. Permite diferenciar los factores estudiados en cada sesión.
- *Variables de bloque*. Permite conocer si en el diseño se han utilizado variables de bloque con el fin de estudiar la posible influencia de otras variables en la variable respuesta.
- *N. de sujetos*. Permite conocer cuantos participantes hay en cada experimento.
- *Unidades experimentales*. Permite conocer las unidades experimentales que se podrán añadir al agregar los experimentos.
- *Duración de la tarea*. Permite conocer si hubo restricciones de tiempo durante la realización de la tarea experimental.

- *Duración del entrenamiento.* Permite conocer si hubo restricciones de tiempo durante el entrenamiento.
- *Transmisión de los tratamientos.* Permite conocer si se han realizado cambios en el material utilizado para la enseñanza de los tratamientos.

Dimensión	Elemento	Estado	Estructura inicial
Identificación	Sitio Institución	Redefinido	Experimentadores y sitio
	Sitio	Reorganizado	-
	Publicación	Añadido	-
Operacionalización	Causa	Factor	Añadido
		Tratamiento	-
		Tratamiento acorde a la literatura de SE	-
		Procedimiento de aplicación del tratamiento	-
		Tratamiento como instrucciones	-
	Efecto	Variable respuesta	-
		Op. de efecto Elaboración de la VR	-
		Métrica	-
	Medición	Instrumentos o equipo de medición	P. Exp: Instrumentación
		Procedimiento de medición	Op.: Efecto
Protección Ejecución experimental	Diseño	Diseño experimental	Protocolo Experimental
		Número de sesiones requeridas por el sujeto para aplicar el tratamiento	Protocolo Experimental
		Número de factores por sesión	-
		N. de tratamientos por factor por sesión	Protocolo Experimental
		Variables de bloque	-
		Orden de aplicación de los tratamientos	Protocolo Experimental
		Tarea Objetos por sujeto por sesión	Protocolo Experimental
		Número de sujetos	-
		Número de unidades experimentales	-
		Duración de la tarea	-
	Entrenamiento	Combinación entre entrenamiento y sesiones	P. Exp.: Diseño
		Aspectos de transmisión del tratamiento	Op.: Causa
		Duración del entrenamiento	-
	Análisis	Técnica de análisis de datos	P. Exp.: Análisis
Instrumentación		Guías suplementarias	P. Exp.: Materiales
		Formularios	P. Exp.: Materiales
		Transmisión de los tratamientos	-
		Objetos experimentales	P. Exp.: Objetos exp.
		Recursos del trat. Infraestructura tecnológica	Op: Causa
Población	Propiedades de los sujetos	Características Tipo de sujetos	-
		comunes Experiencia tratamiento	-
		Otras condiciones Características variables	-
	Propiedades de los objetos experimentales	Complejidad	-
		Tamaño Dominio	-
		Lenguaje de programación	-
Experimentadores y sitio		Diseñador	-
		Entrenador	-
		Monitor	-
		Medidor	-
		Analista	-

Tabla 4.2 Diferencias con la estructura inicial

- *Redefinir elementos.* Algunos de los elementos fueron redefinidos, para que fueran más genéricos.
 - *Institución.* Sustituye Sitio.
 - *Elaboración de la variable respuesta.* Sustituye a Operacionalización del efecto.
 - *Número de sesiones.* Sustituye a Número de sesiones requeridas por el sujeto para aplicar el tratamiento.
 - *Objetos experimentales por sesión.* Sustituye a Tareas por sujeto por sesión.
 - *Infraestructura tecnológica.* Sustituye a Recursos del tratamiento.
 - *Características variables.* Sustituye a Otras condiciones.
 - *Dominio.* Se cambió el tamaño de los objetos experimentales por el dominio, para diferenciar entre programas de juguete y programas reales. Dado que el tamaño, estaba confundido con la complejidad de los objetos experimentales.
 - *Ejecución experimental.* Esta dimensión sustituye a la dimensión Protocolo Experimental. Esta dimensión ahora solo contiene dos subdimensiones: Diseño y Entrenamiento; mientras que la dimensión inicial tenía cinco subdimensiones: Análisis, Diseño, Instrumentos, Objetos experimentales y Materiales. Las subdimensiones Protocolo experimental: Objetos experimentales y Protocolo experimental: Materiales, fueron reemplazadas por la dimensión Instrumentación y la subdimensión Protocolo experimental: Análisis por la dimensión Análisis.
 - *Experimentadores.* Sustituye a la dimensión Experimentadores y sitio.
- *Reorganizar elementos.* Permite que los elementos comunes se encuentren en la misma dimensión.
 - *Institución.* Se mueve de Experimentadores y Sitio a Identificación.
 - *Instrumentos o equipo de medición.* Se mueve de Protocolo Experimental: Instrumentación a Operacionalización: Medición.
 - *Procedimiento de medición.* Se mueve de Operacionalización: Efecto a Operacionalización: Medición.
 - *Combinación entre entrenamiento y sesiones.* Se mueve de Protocolo Experimental: Diseño a Ejecución experimental: Entrenamiento.
 - *Aspectos de transmisión del tratamiento.* Se mueve de Operacionalización: Causa a Ejecución experimental: Entrenamiento
 - *Técnica de análisis de datos.* Se mueve de Protocolo experimental: Análisis a Análisis.
 - *Guías suplementarias.* Se mueve de Protocolo experimental: Materiales a Instrumentación.
 - *Formularios.* Se mueve de Protocolo experimental: Materiales a Instrumentación.
 - *Objetos experimentales.* Se mueve de Protocolo experimental: Objetos experimentales a Instrumentación.
 - *Infraestructura tecnológica.* Se mueve de Operacionalización: Causa a Instrumentación.

Nótese que no se ha borrado ninguno de los elementos que aparecen en la propuesta original. Sin embargo, la dimensión Análisis y el analista en la dimensión Experimentadores, no se considerarán en este procedimiento de caracterización, debido a que los análisis estadísticos se realizan nuevamente en

la actividad 2.2 de este proceso y por lo tanto no influyen en la identificación de diferencias entre los experimentos de la familia.

Por lo tanto, el esquema de caracterización de experimentos que finalmente se usará aparece reflejado en la Tabla 4.3. A continuación, se describen cada uno de sus elementos.

Dimensión		Elemento
Identificación		Institución
		Sitio
		Publicación
Operacionalización	Causa	Factores
		Tratamientos
		Tratamiento acorde a la literatura de SE
		Procedimiento de aplicación del tratamiento
		Tratamiento como instrucciones
	Efecto	Variable respuesta
Elaboración de la VR		
Métrica		
Medición	Instrumentos o equipo de medición	
	Procedimiento de medición	
Ejecución experimental	Diseño	Diseño experimental
		Número de sesiones
		Número de factores por sesión
		Número de tratamientos por factor por sesión
		Variables de bloque
		Orden de aplicación de los tratamientos
		Objetos experimentales por sesión
		Número de sujetos
		Unidades experimentales
	Duración de la tarea	
	Entrenamiento	Combinación entre entrenamiento y sesiones
Aspectos de transmisión del tratamiento		
Duración del entrenamiento		
Instrumentación		Guías suplementarias
		Formularios
		Transmisión de los tratamientos
		Objetos experimentales
		Infraestructura tecnológica
Población	Propiedades de los sujetos	Características comunes
		Características variables
	Propiedades de los objetos experimentales	Complejidad
		Dominio
Lenguaje de programación		
Experimentadores		Diseñador
		Entrenador
		Monitor
		Medidor

Tabla 4.3 Plantilla para la caracterización

Hay 3 elementos de **identificación**: institución, sitio y publicación.

- **Institución.** Indica el nombre de la institución donde se realizó el experimento.
- **Sitio.** Indica si el experimento se realizó en la academia o en la industria.

- *Publicación.* Indica si el experimento se ha publicado individualmente.

Hay 11 elementos de **operacionalización** agrupados en 3 subdimensiones: cinco de causa, tres de efecto y dos de medición de datos.

Los elementos de causa son:

- *Factores.* Son los elementos principales de un experimento. Permiten evaluar la variable respuesta.
- *Tratamientos.* Son los distintos valores que puede tomar el factor. Se compone del grupo experimental, que es el tratamiento que se desea estudiar y el grupo de control, que es un elemento neutro con el que se compara el grupo experimental.
- *Tratamiento acorde a la literatura.* El investigador utiliza alguna fuente (por ejemplo, la literatura) para seleccionar los tratamientos que desea examinar.
- *Procedimiento de aplicación del tratamiento.* El investigador adapta de la literatura a su experimento el procedimiento de aplicación de los tratamientos.
- *Tratamiento como instrucciones.* Una vez adaptados los tratamientos, éstos son transformados en instrucciones que el sujeto realizará en las sesiones experimentales.

Los elementos de efecto son:

- *Variable respuesta.* Es la variable de interés, sobre la que se mide la influencia del factor.
- *Elaboración de la variable respuesta.* Es la definición de la variable respuesta, el paso previo para determinar su métrica. Por ejemplo: para la variable respuesta Calidad, su elaboración sería: “conocer como de bien está hecho un trabajo”.
- *Métrica.* Representa la operacionalización de efecto que el investigador define para medir el constructo de efecto. Por ejemplo, para medir la efectividad (constructo) de una técnica de inspección de código, el investigador puede operacionalizar este constructo a través de la métrica “porcentaje de faltas observadas”.

Los elementos de medida son:

- *Procedimiento de medición.* El investigador especifica el procedimiento de medición en el se describe cómo calcular y asignar a la variable de respuesta las observaciones obtenidas. Las variables respuesta se derivan de las métricas previamente definidas. Este procedimiento puede ser sencillo, por ejemplo, obtener de forma directa los valores de la variable respuesta o complejo como puede ser utilizar una serie de fórmulas para calcular a partir de las observaciones sus valores. En algunos casos este procedimiento puede automatizarse de tal forma que las observaciones recabadas durante el experimento se asignen directamente a las variables respuesta.
- *Instrumentos o equipos de medición.* Este elemento se refiere a los instrumentos o equipos de medición que pueden emplearse para medir la variable de respuesta del experimento. Por ejemplo, una variable de respuesta relacionada con el tiempo de duración de alguna tarea puede medirse de forma manual (el sujeto lleva un registro manual del tiempo empleado en alguna tarea) o automática se dispone de un instrumento que registra automáticamente el tiempo de la tarea.

Hay 13 elementos que componen la **ejecución experimental** agrupados en 2 subdimensiones: diez de diseño y tres de entrenamiento.

Los elementos de diseño son:

- *Diseño experimental*. El diseño experimental dicta la forma en cómo se deben organizar los tratamientos que se asignan a los participantes. Existen varios tipos de diseños experimentales y dependiendo del tipo, es posible examinar interacciones entre los tratamientos y otras condiciones o elementos de la situación experimental.
- *Sesiones experimentales*. Indica el número de sesiones para aplicar los tratamientos.
- *Factores por sesión*. Indica el número de factores que se estudian en cada sesión.
- *Tratamientos por factor y por sesión*. Indica el número de tratamientos que se estudian en cada sesión. Cuando hay más de un factor permite saber cuantos tratamientos de cada factor se usan en la misma sesión.
- *Variables de bloque*. Permite determinar grupos de participantes con el objetivo de disminuir la variabilidad de un modelo.
- *Orden de aplicación de los tratamientos*. Indica el orden en el que se aplican los tratamientos, puede ser: fijo, cuando los tratamientos se aplican siempre en el mismo orden (diseños Intra-sujetos); todos, cuando los tratamientos se aplican en todas sus combinaciones posibles (diseños crossover); o ninguno, cuando los participantes aplican los tratamientos una única vez (diseños Entre-sujetos).
- *Objetos experimentales por sesión*. Indica el total de objetos utilizados en cada sesión.
- *Número de participantes*. Indica el total de participantes de un experimento.
- *Unidades experimentales*. Indica el total de unidades experimentales del experimento.
- *Duración de la tarea*. Indica el tiempo para aplicar los tratamientos. Si los participantes no tienen límite de tiempo se considera Ilimitado en caso contrario Limitado.

Los elementos de entrenamiento son:

- *Combinación del entrenamiento con las sesiones*. Puede ser Secuencial si los participantes recibieron primero el entrenamiento completo y después se realizaron las sesiones del experimento; o Intercalado si los participantes reciben una parte del entrenamiento después realizan la sesión del experimento referente al entrenamiento dado, nuevamente reciben otra parte el entrenamiento y aplican otro tratamiento.
- *Aspectos de transmisión del tratamiento*. La forma en cómo se transmiten los tratamientos a los participantes puede influir en los resultados. Por ejemplo, si el investigador responsable de transmitir el tratamiento a los participantes no cuenta con suficiente experiencia en los tratamientos que se utilizan en el experimento.
- *Duración del entrenamiento*. Indica el tiempo utilizado en enseñar los tratamientos.

Los elementos que componen la **instrumentación** son los siguientes cinco:

- *Guías suplementarias*. Las guías suplementarias orientan al sujeto sobre algún aspecto relacionado con el tratamiento o con el objeto experimental. Por ejemplo, en un experimento donde se compara la efectividad de varias técnicas de inspección de

código, las guías suplementarias pueden contener información referente a alguna de las funciones o procedimientos implementados en los códigos fuente.

- *Formularios.* Los formularios se utilizan para recabar información de las tareas que realizan los participantes durante la sesión del experimento.
- *Transmisión de los tratamientos.* Es la documentación para enseñar los tratamientos.
- *Objetos experimentales.* Los objetos experimentales son los artefactos tales como especificaciones, diagramas, códigos fuente o programas que se entregan a los participantes durante el experimento para que completen una serie de tareas.
- *Infraestructura tecnológica.* La aplicación del tratamiento puede requerir de algún tipo de software o hardware. Dependiendo del experimento, los recursos pueden ser manuales i.e. no se requiere el uso de ordenador; o automatizados, como puede ser algún tipo de herramienta software, sistema operativo o tipo de ordenador.

Hay 5 elementos que componen la población agrupados en 2 subdimensiones: dos de propiedades de los sujetos y tres de propiedades de los objetos.

Los elementos de propiedades de los sujetos son:

- *Características comunes de los sujetos.* Son las características que no varían entre los participantes de un mismo experimento. Suelen ser: tipo (estudiantes, posgrado, profesionales, etc.); la motivación de los participantes, etc.
- *Características variables de los sujetos.* Sirven para diferenciar a los participantes de un experimento. Suelen estar relacionados con sus conocimientos y experiencia.

Los elementos de propiedades de los objetos son:

- *Complejidad.* Es una propiedad de los objetos experimentales, por ejemplo, el código fuente suele medirse por líneas de código de programa o la complejidad ciclomática. Tenga en cuenta que, aunque esta sea la misma en todos los experimentos, para algunos participantes los objetos podrían ser más/menos complejos que para otros.
- *Dominio.* Es el tipo de objetos experimentales, pueden ser de juguete si se han creado para el experimento o reales si los participantes trabajan con ellos en su día a día.
- *Lenguaje de programación.* Es una propiedad de los objetos experimentales, que indica en que lenguaje se han escrito los programas utilizados en los experimentos.

Hay 5 elementos que componen la dimensión de los **experimentadores** que son las personas encargadas del desarrollo y ejecución de los experimentos:

- *Diseñador.* Se encarga de definir y especificar la estructura de los tratamientos que se asignan a los participantes; Definir la operación del experimento y planificar la ejecución de éste; Así mismo elabora estrategias para aumentar el control sobre posibles variables extrañas mitigando las amenazas a la validez del experimento.
- *Entrenador.* Es responsable de transmitir el tratamiento en el entrenamiento, el tratamiento puede ser transmitido durante una asignatura, curso, seminario o taller.
- *Monitor.* Cuida a los participantes durante la sesión experimental, sus funciones son dar instrucciones a los participantes referentes a cómo se llevará a cabo la sesión,

entregarles los materiales (tratamientos, formularios, objetos experimentales y herramientas), recoger los materiales al final de la sesión, y aclarar dudas. Nótese que la aclaración de dudas puede alterar el concepto de los tratamientos.

- *Medidor.* Realiza el procedimiento de medición sobre los datos recabados.

4.2.1.2 Caracterización de una Familia de Experimentos

La Tabla 4.4 y la Tabla 4.5 muestran un ejemplo de la caracterización de una de las 46 familias que se han mencionado en el Capítulo 2 [EP14]. Cuando los autores no reportan algún elemento de la configuración experimental las celdas correspondientes contienen un N/A.

Dimensión		Elemento	Valor	UPV1	UPV2	UPV3	UNIBAS
Operacionalización	Causa	Factor	Método de Evaluación de Arquitectura	x	x	x	x
		Tratamiento	ATAM	x	x	x	x
			QuaDAI	x	x	x	x
		Tratamiento acorde a la literatura	Sección 3.1 y 3.2 del artículo	x	x	x	x
		Tratamiento como instrucciones	Sección 3.1 y 3.2 del artículo	x	x	x	x
		Procedimiento aplicación tratamiento	Sección 3.1 y 3.2 del artículo	x	x	x	x
	Efecto	Variable respuesta	Efectividad	x	x	x	x
			Eficiencia	x	x	x	x
			Facilidad de uso percibida	x	x	x	x
			Utilidad percibida	x	x	x	x
			Intención de uso	x	x	x	x
		Elaboración de la VR	Sección 5.1.1 del artículo	x	x	x	x
		Métrica	Sección 5.1.1 del artículo	x	x	x	x
			Nuevo NFR en efectividad	-	-	-	x
	Medición	Instrumentos de medición	N/A	N/A			
		Procedimiento de medición	N/A	N/A			
Ejecución experimental	Diseño	Diseño experimental	Intra-sujetos	x	x	x	x
		Sesiones	2	x	x	x	x
		Factores por sesión	1	x	x	x	x
		Tratamientos por factor por sesión	2	x	x	x	x
		Variables de bloque	-	-	-	-	-
		Orden aplicación tratamientos (gr)	Todos (4)	x	x	x	x
		Objetos experimentales por sesión	2	x	x	x	x
			28	x	-	-	-
		Número de sujetos	16	-	x	-	-
			36	-	-	x	-
			12	-	-	-	x
		Unidades experimentales	4*N. participantes	x	x	x	x
		Duración de la tarea	90 min (ilimitado)	x	x	x	x
	Entrenamiento	Combinación entrenamiento y sesiones	Intercalado	x	x	x	-
			Secuencial	-	-	-	x
		Aspectos de transmisión del trat.	N/A	N/A			
		Duración del entrenamiento	1 sesión (120), 2 sesiones (30). Total: 180 min	x	x	x	-
			3 sesiones (60). Total: 180 min	-	-	-	x

Tabla 4.4 Caracterización familia [EP14] (1/2)

Dimensión		Elemento	Valor		UPV1	UPV2	UPV3	UNIBAS
Instrumentación	Guías suplementarias		Apéndice A.1.1 del artículo		x	x	x	x
			Apéndice A.1.3 del artículo		x	x	x	x
			Apéndice A.1.4 del artículo		x	x	x	x
			Apéndice A.1.5 del artículo		x	x	x	x
			El material se tradujo del español al inglés		-	-	-	x
	Formularios		Apéndice A.2 del artículo		x	x	x	x
			Se incluyeron preguntas de control		-	x	x	x
	Transmisión de los tratamientos		El material se tradujo del español al inglés		-	-	-	x
			Diapositivas que contienen la introducción a arquitecturas de software, patrones arquitectónicos y evaluación de arquitectura de software		x	x	x	x
			Diapositivas que describen método QuaDAI		x	x	x	x
			Diapositivas que describen método ATAM		x	x	x	x
Población	Sujetos	Características comunes	Tipo de sujetos	Estudiantes de Grado	x	-	x	x
				Estudiantes de Máster	-	x	-	-
	Objetos	Complejidad	Características variables		N/A			
					N/A			
					N/A			
		Dominio	Se modificó un nivel de un NFR en el objeto experimental O2		-	x	x	x
			Se incluyó un tercer NFR y una nueva métrica de software		-	-	-	x
			El material se tradujo del español al inglés		-	-	-	x
Experimentador			N/A		N/A			
			Investigador externo		-	-	-	x
			Investigador externo		-	-	-	x
			N/A		N/A			
			N/A		N/A			

Tabla 4.5 Caracterización familia [EP14] (2/2)

4.2.1.3 Herramienta para Recoger la Información de Familias de Experimentos

Con el objetivo de facilitar el almacenamiento de la información asociada a los elementos de la configuración experimental, se ha desarrollado como parte de esta tesis una aplicación web. La aplicación, además de recoger la información asociada a las familias de experimentos, sirve como repositorio para almacenar tanto los ficheros de los datos obtenidos en los experimentos, como los ficheros de análisis de datos generados (tanto a nivel de experimento como de familia).

La Figura 4.2 muestra una de las pantallas para dar de alta un nuevo experimento de una familia, concretamente esta pantalla se corresponde con la dimensión de Identificación de la Tabla 4.3.

Figura 4.2 Pantalla para añadir un nuevo experimento

La aplicación permite la gestión (alta, baja, modificación y consulta) de usuarios, de familias de experimentos y de experimentos:

- **Gestión de usuarios.** El sistema permite el acceso mediante usuario y contraseña, que además está securizado mediante roles. Los roles existentes son: administrador, creador, visualizador.
 - *Alta de usuarios:* únicamente los usuarios con rol de “administrador” pueden crear nuevos usuarios.
 - *Baja de usuarios:* únicamente los usuarios con rol de “administrador” pueden dar de baja a otros usuarios, siempre y cuando el rol sea inferior.
 - *Modificación de usuarios:* cualquier usuario puede editar sus propios datos personales.
 - *Consulta de usuarios:* cualquier usuario puede consultar la información de los demás usuarios en el sistema.
- **Gestión de familias de experimentos.** El sistema permite la creación de familias a las que posteriormente se irán añadiendo experimentos.
 - *Alta de familias de experimentos:* únicamente los usuarios con rol de “administrador” pueden crear nuevas familias de experimentos.
 - *Baja de familias de experimentos:* únicamente los usuarios con rol de “administrador” pueden dar de baja nuevas familias, siempre y cuando se hayan eliminado previamente todos los experimentos asociados a la familia.
 - *Modificación de familias de experimentos:* únicamente los usuarios con rol de “administrador” pueden modificar los datos de la familia (tema bajo estudio y breve descripción).
 - *Consulta de familias de experimentos:* cualquier usuario puede consultar la información de las familias dadas de alta en el sistema.

- **Gestión de experimentos.** El sistema permite la creación de experimentos que necesariamente tienen que estar asociados a una familia.
 - *Alta de un experimento:* cualquier usuario con rol “creador” y que posea permisos de acceso a una familia puede añadir experimentos, esto requiere rellenar todos los elementos que se muestran en la Tabla 4.3.
 - *Baja de un experimento:* los usuarios con rol de “administrador” o aquellos con el rol “creador” que lo dieron de alta al experimento (“dueño”) pueden darlo de baja. Tenga en cuenta que esto eliminará toda la información relacionada con el mismo.
 - *Modificación de un experimento:* los usuarios con rol de “administrador” o aquellos con el rol “creador” que lo dieron de alta al experimento (“dueño”) pueden modificar los elementos que se muestran en la Tabla 4.3.
 - *Consulta de un experimento:* cualquier usuario puede consultar la información de los experimentos asociados a las distintas familias.

Más detalles sobre la aplicación y más capturas de pantalla se encuentran en el Apéndice B.

4.2.2 Identificación de los Grupos de Experimentos a Agregar

La dimensión de la *operacionalización* de la Tabla 4.3 describe la esencia del experimento. Los tratamientos que son investigados, las variables respuesta que se usan y cómo están siendo medidas. Como ya se ha mencionado, este proceso permite únicamente la agregación de los resultados de aquellos experimentos en la familia que poseen una operacionalización similar. Esto significa que se pueden identificar diferentes subgrupos de experimentos dependiendo de los tratamientos estudiados o la variable respuesta que se mide.

Los elementos de la caracterización que se deben tener en cuenta para la identificación de subgrupos aparecen reflejados en la Tabla 4.6 y son todos los de la operacionalización. Los experimentos que pertenezcan al mismo grupo deben cumplir que:

Dimensión	Elemento
Operacionalización	Factor
	Tratamiento
	Tratamiento acorde a la literatura de SE
	Procedimiento de aplicación del tratamiento
	Tratamiento como instrucciones
	Variable respuesta
Efecto	Elaboración de la VR
	Métrica
Medición	Instrumentos o equipo de medición
	Procedimiento de medición

Tabla 4.6 Elementos para la identificación de grupos de experimentos

- Todos los elementos de la dimensión de operacionalización del efecto sean los mismos. Es decir, todos los experimentos del mismo grupo deben examinar la misma variable respuesta, que a su vez debe tener la misma elaboración y estar medida con la misma métrica. Esto es un requisito que viene impuesto por la técnica de análisis estadístico.

- Todos los elementos de la dimensión de operacionalización de la medición sean, si no igual, al menos equivalentes. Es decir, tanto los instrumentos de medida utilizados para medir la variable respuesta, como el procedimiento de medida seguidos, deben ser, si no iguales, al menos lo suficientemente similares como para garantizar que los datos medidos tienen el mismo significado. Por ejemplo, se está midiendo la calidad de un programa utilizando como instrumento de medida un conjunto de casos de prueba generados. Y en dos experimentos distintos se utiliza distintos lenguajes de programación (C++ y Java). Se sabe que, el lenguaje de programación C++ no permite la asignación a una variable con el valor nulo (null), mientras que el lenguaje de programación Java sí (int i = null;). Por tanto, se pueden utilizar diferentes casos de prueba que contemplen esta restricción al medir la calidad del mismo programa en ambos lenguajes. Si esto no ocurriese, los datos serían “diferentes” y no se cumpliría el requisito impuesto por la técnica de análisis.
- Haya al menos un factor en común en la dimensión de operacionalización de la causa. Específicamente el que será considerado como el factor principal de la familia y que corresponde con la tecnología de software en estudio por la familia. Nótese que la agregación, independientemente de la técnica utilizada o el proceso seguido no tiene sentido si no hay un objeto de estudio común.

Continuando con el ejemplo de la familia [EP14], en la Tabla 4.7 se muestran los valores para los elementos de la Operacionalización. En ella se puede observar que se podría crear un único grupo de agregación ya que todos los elementos de causa son exactamente iguales. Sin embargo:

- El hecho de medir un nuevo requisito no funcional en la variable respuesta efectividad podría hacer que los elementos de la dimensión efecto no sean los mismos y que, por lo tanto, no se cumpla la primera restricción.

Dimensión		Elemento	Valor	UPV1	UPV2	UPV3	UNIBAS
Operacionalización	Causa	Factor	Método de Evaluación de Arquitectura	x	x	x	x
		Tratamiento	ATAM	x	x	x	x
			QuaDAI	x	x	x	x
			Tratamiento acorde a la literatura	x	x	x	x
			Tratamiento como instrucciones	x	x	x	x
			Procedimiento aplicación tratamiento	x	x	x	x
	Efecto	Variable respuesta	Efectividad	x	x	x	x
			Eficiencia	x	x	x	x
			Facilidad de uso percibida	x	x	x	x
			Utilidad percibida	x	x	x	x
			Intención de uso	x	x	x	x
		Elaboración de la VR	Sección 5.1.1 del artículo	x	x	x	x
			Sección 5.1.1 del artículo	x	x	x	x
	Medición	Métrica	Nuevo NFR en efectividad	-	-	-	x
		Instrumentos de medición	N/A	N/A			
		Procedimiento de medición	N/A	N/A			

Tabla 4.7 Identificación de grupos de experimentos, familia [EP14]

- No se han mencionado elementos de medición y por lo tanto no se sabe si se cumple la segunda restricción.

4.3 Identificación de las Variables de Diseño

Dado que el proceso aquí descrito permite la agregación de experimentos con distintos diseños, es necesario establecer cuáles serán las variables de diseño que han de incluirse en el análisis conjunto. En función de las características de los diseños individuales, podrán aparecer nuevas variables de análisis.

Para la identificación del diseño individual de cada uno de los experimentos se ha de utilizar la información que aparece en la Tabla 4.8.

Dimensión	Elemento
Ejecución experimental	Diseño experimental
	Número de sesiones
	Número de factores por sesión
	Número de tratamientos por factor por sesión
	Variables de bloque
	Orden de aplicación de los tratamientos
	Objetos experimentales por sesión
	Número de sujetos
	Unidades experimentales

Tabla 4.8 Elementos para la identificación de variables de experimento

No obstante, en el proceso aquí propuesto se propone utilizar una representación gráfica del diseño experimental en lugar de mirar los elementos que componen esta dimensión, ya que un gráfico expresa mejor el diseño y facilita su comprensión.

4.3.1 Creación de un Diseño Agregado

Para conocer las variables de análisis a considerar, se realizará una superposición de todos los experimentos del grupo. Mediante la superposición se creará lo que se ha denominado un **diseño agregado**.

La creación del diseño agregado se realiza de manera iterativa, empezando con cualquier pareja de diseños, e incorporando sucesivamente, uno a uno, el resto de los diseños. Es importante hacer notar que si todos los diseños del grupo de experimentos son exactamente iguales, el diseño agregado corresponderá con el mismo que tiene cada experimento individual. Con el objetivo de ilustrar cómo ha de realizarse la superposición, a continuación, se muestran una serie de heurísticas, asociadas a las diferentes situaciones que pueden surgir.

Previamente, es necesario recordar al lector, que en función de cómo se realice la asignación de los tratamientos a los participantes en un experimento, un diseño puede ser:

- *Entre-sujetos*. En este diseño, cada participante aplica un único tratamiento. Se establecen tantos grupos de participantes como tratamientos. La Tabla 4.9 muestra un ejemplo de este tipo de diseño. El factor *Factor* posee los niveles *Tratamiento₁*, *Tratamiento₂*, *Tratamiento₃* y *Tratamiento₄* y utiliza un único objeto experimental.

E	S1
G ₁	Tratamiento ₁ -Objeto ₁
G ₂	Tratamiento ₂ -Objeto ₁
G ₃	Tratamiento ₃ -Objeto ₁
G ₄	Tratamiento ₄ -Objeto ₁

Tabla 4.9 Ejemplo de un diseño Entre-sujetos

- **Intra-sujetos.** En este diseño todos los participantes aplican todos los tratamientos. Estos pueden ser de medidas repetidas o crossover.
 - **Medidas repetidas.** Los tratamientos se aplican en el mismo orden. En cada sesión se usa un único tratamiento, sin embargo, dependiendo de la manera en que se reparten los objetos experimentales, este tipo de diseño puede tener las tareas cruzadas si diferentes grupos experimentales aplican cada tratamiento sobre diferentes objetos experimentales. La Tabla 4.10 y 4.11 muestran un ejemplo de este tipo de diseño, sin cruzar y cruzando tareas. En ambos casos el factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y utiliza dos objetos experimentales *Objeto₁* y *Objeto₂*.

E	S1	S2
G ₁	Tratamiento ₁ -Objeto ₁	Tratamiento ₂ -Objeto ₁
G ₂	Tratamiento ₁ -Objeto ₂	Tratamiento ₂ -Objeto ₂

Tabla 4.10 Ejemplo de un diseño Intra-sujetos sin cruzar tareas

E	S1	S2
G ₁	Tratamiento ₁ -Objeto ₁	Tratamiento ₂ -Objeto ₂
G ₂	Tratamiento ₁ -Objeto ₂	Tratamiento ₂ -Objeto ₁

Tabla 4.11 Ejemplo de un diseño Intra-sujetos cruzando tareas

- **Crossover.** A diferencia de los diseños de medidas repetidas aquí los tratamientos se aplican en distinto orden. Se establecen tantos grupos experimentales como posibles órdenes de aplicación de los tratamientos. La Tabla 4.12 muestra un ejemplo de este tipo de diseño. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y utiliza dos objetos experimentales *Objeto₁* y *Objeto₂*.

E	S1	S2
G ₁	Tratamiento ₁ -Objeto ₁	Tratamiento ₂ -Objeto ₂
G ₂	Tratamiento ₂ -Objeto ₂	Tratamiento ₁ -Objeto ₁

Tabla 4.12 Ejemplo de un diseño Crossover

Por su parte en función del número de factores que haya en el diseño y la posible existencia de variables de bloque, un diseño puede ser:

- **Un factor.** Existe una única fuente de variación en el experimento, correspondiente al factor. El factor puede tener dos o más tratamientos. La Tabla 4.13 muestra un ejemplo de este tipo de diseño. El factor *Factor* posee los niveles *Tratamiento₁*, *Tratamiento₂*, *Tratamiento₃* y *Tratamiento₄* y utiliza un único objeto experimental.

E	S1	S2
G ₁	Tratamiento ₁ -Objeto ₁	Tratamiento ₂ -Objeto ₁
G ₂		
G ₃		Tratamiento ₄ -Objeto ₁
G ₄		

Tabla 4.13 Ejemplo de un diseño de Un factor

- *Un factor con variables de bloque.* Existe una única fuente de variación en el experimento, correspondiente al factor, pero además se pueden establecer bloques de participantes en los que cada nivel de esta variable aplica a cada uno de los niveles del factor principal. Es importante tener en cuenta que en este tipo de diseño se puede conocer la influencia del factor y la variable de bloque, pero no se puede examinar su interacción. La Tabla 4.14 muestra un ejemplo de este tipo de diseño. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y utiliza dos objetos experimentales *Objeto₁* y *Objeto₂*, en este caso el objeto experimental actúa como una variable de bloque.

E	S1	S2
G ₁	Tratamiento ₁ -Objeto ₁	Tratamiento ₂ -Objeto ₂
G ₂		

Tabla 4.14 Ejemplo de un diseño de Un factor con variables de bloque

- *Factorial.* Existen dos o más fuentes de variación en el experimento, correspondientes a diferentes factores. Donde además de conocer la influencia de cada factor, también se puede examinar una posible interacción entre ellos. La Tabla 4.15 muestra un ejemplo de este tipo de diseño. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y el factor *Objeto experimental* posee los niveles *Objeto₁* y *Objeto₂*. Como hay dos factores y cada uno tiene dos niveles el diseño mostrado es un factorial 2x2.

E	S1
G ₁	Tratamiento ₁ -Objeto ₁
G ₂	Tratamiento ₁ -Objeto ₂
G ₃	Tratamiento ₂ -Objeto ₁
G ₄	Tratamiento ₂ -Objeto ₂

Tabla 4.15 Ejemplo de un diseño Factorial

Para simplificar, las heurísticas definidas consideran diseños con una variable de bloque o dos factores como máximo, y cada factor con dos niveles. No obstante, la extrapolación a otros casos es directa. Cada heurística se ilustra con un ejemplo:

1. *Los diseños se complementan entre sí y se añaden nuevos niveles a un factor ya existente.* La Figura 4.3 muestra una situación donde hay dos experimentos con dos factores: *Factor* y *Objeto experimental*, cada uno de ellos con dos niveles. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂*. En el caso del factor *Objeto experimental* posee los niveles *Objeto₁* y *Objeto₂* en el Experimento₁ y los niveles *Objeto₃* y *Objeto₄* en el Experimento₂. En este caso la superposición trae como consecuencia que se crea un diseño agregado en el que el factor *Objeto experimental* tiene 4 niveles.

E ₁	S1		E ₂	S1	=		S1
G _{1.1}	Tratamiento ₁ -Objeto ₁		G _{2.1}	Tratamiento ₁ -Objeto ₃		G _{1(1.1)}	Tratamiento ₁ -Objeto ₁
G _{1.2}	Tratamiento ₁ -Objeto ₂	+	G _{2.2}	Tratamiento ₁ -Objeto ₄		G _{2(1.2)}	Tratamiento ₁ -Objeto ₂
G _{1.3}	Tratamiento ₂ -Objeto ₁		G _{2.3}	Tratamiento ₂ -Objeto ₃		G _{3(1.3)}	Tratamiento ₂ -Objeto ₁
G _{1.4}	Tratamiento ₂ -Objeto ₂		G _{2.4}	Tratamiento ₂ -Objeto ₄		G _{4(1.4)}	Tratamiento ₂ -Objeto ₂
						G _{5(2.1)}	Tratamiento ₁ -Objeto ₃
						G _{6(2.2)}	Tratamiento ₁ -Objeto ₄
						G _{7(2.3)}	Tratamiento ₂ -Objeto ₃
						G _{8(2.4)}	Tratamiento ₂ -Objeto ₄

Figura 4.3 Superposición de 2 diseños de 2 factores, en el que se añaden nuevos niveles

2. Los diseños se complementan entre sí. En función del tipo de diseño puede ocurrir que:

- Dos diseños de un factor (entre sujetos, medidas repetidas o crossover) formen un diseño de dos factores, porque se crea un nuevo factor. La Figura 4.4 muestra una situación donde hay dos experimentos de un factor: *Factor* y este posee dos niveles *Tratamiento₁* y *Tratamiento₂*. En el Experimento₁ se usa un objeto experimental Objeto₁ y en el Experimento₂ se usa un objeto experimental Objeto₂ (diferente al del Experimento₁). En este caso la superposición trae como consecuencia que se crea el factor *Objeto experimental*.

E ₁	S1		E ₂	S1	=		S1
G _{1.1}	Tratamiento ₁ -Objeto ₁		G _{2.1}	Tratamiento ₁ -Objeto ₂		G _{1(1.1)}	Tratamiento ₁ -Objeto ₁
G _{1.2}	Tratamiento ₂ -Objeto ₁	+	G _{2.2}	Tratamiento ₂ -Objeto ₂		G _{2(1.2)}	Tratamiento ₂ -Objeto ₁
						G _{3(2.1)}	Tratamiento ₁ -Objeto ₂
						G _{4(2.2)}	Tratamiento ₂ -Objeto ₂

Figura 4.4 Superposición de 2 diseños de un factor, en el que se crea un nuevo factor

- Dos diseños intra-sujetos forman un diseño crossover. La Figura 4.5 muestra una situación donde hay dos experimentos con diseño intra-sujetos donde se aplican los tratamientos en distinto orden. Ambos experimentos tienen dos factores: *Factor* y *Objeto experimental*, cada uno de ellos con dos niveles. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y el factor *Objeto experimental* posee los niveles *Obj₁* y *Obj₂*. En el Experimento₁ se aplica el *Tratamiento₁* en la primera sesión y el *Tratamiento₂* en la segunda sesión, mientras que en el Experimento₂ los tratamientos se aplican en orden inverso, el *Tratamiento₂* en la primera sesión y el *Tratamiento₁* en la segunda sesión. En ambos casos el factor *Objeto experimental* se aplica en el mismo orden, es decir, *Obj₁* en la primera sesión y *Obj₂* en la segunda sesión. En este caso la superposición trae como consecuencia que se crea un diseño crossover.

E ₁	S1	S2		E ₂	S1	S2
G _{1.1}	Tratamiento ₁ -Obj ₁	Tratamiento ₂ -Obj ₂	+	G _{2.1}	Tratamiento ₂ -Obj ₁	Tratamiento ₁ -Obj ₂
			=			
					S1	S2
				G _{1(1.1)}	Tratamiento ₁ -Obj ₁	Tratamiento ₂ -Obj ₂
				G _{2(2.1)}	Tratamiento ₂ -Obj ₁	Tratamiento ₁ -Obj ₂

Figura 4.5 Superposición de 2 diseños intra-sujetos, en el que se crea un diseño crossover.

3. *Un diseño está contenido en otro diseño.* Nuevamente, en función del tipo de diseño puede ocurrir que:

- *El diseño de un factor está incluido en un diseño de dos factores.* La Figura 4.6 muestra un ejemplo de un diseño de un factor incluido dentro de un diseño de dos factores. El Experimento₁ tiene dos factores: *Factor* y *Objeto experimental*, cada uno de ellos con dos niveles. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y el factor *Objeto experimental* posee los niveles *Objeto₁* y *Objeto₂*. Por su parte, el Experimento₂ tiene un único factor: *Factor* que posee los niveles *Tratamiento₁* y *Tratamiento₂* y además usa un único Objeto experimental el *Objeto₁*. En este caso la superposición trae como consecuencia que el Experimento₂ está incluido en Experimento₁ y por tanto se añaden participantes a los grupos correspondientes, concretamente el grupo 1 es el mismo en ambos experimentos (G_{1.1} y G_{2.1}) y grupo 2 del Experimento₂ (G_{2.2}) es igual al grupo 3 del Experimento₁ (G_{1.3}).

E ₁	S1		E ₂	S1	=		S1
G _{1.1}	Tratamiento ₁ -Objeto ₁		G _{2.1}	Tratamiento ₁ -Objeto ₁		G _{1(1.1 y 2.1)}	Tratamiento ₁ -Objeto ₁
G _{1.2}	Tratamiento ₁ -Objeto ₂		G _{2.2}	Tratamiento ₂ -Objeto ₁		G _{2(1.2)}	Tratamiento ₁ -Objeto ₂
G _{1.3}	Tratamiento ₂ -Objeto ₁					G _{3(1.3 y 2.2)}	Tratamiento ₂ -Objeto ₁
G _{1.4}	Tratamiento ₂ -Objeto ₂					G _{4(1.4)}	Tratamiento ₂ -Objeto ₂

Figura 4.6 Diseño de un factor incluido en un diseño de dos factores

- *El diseño intra-sujetos está incluido en un diseño crossover.* La Figura 4.7 muestra un ejemplo de un diseño intra-sujetos incluido dentro de un diseño crossover. Ambos experimentos tienen dos factores: *Factor* y *Objeto experimental*, cada uno de ellos con dos niveles. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y el factor *Objeto experimental* posee los niveles *Obj₁* y *Obj₂*. En este caso la superposición trae como consecuencia que el Experimento₂ está incluido en Experimento₁ y por tanto se añaden participantes al grupo correspondiente, concretamente el grupo 1 es el mismo en ambos experimentos (G_{1.1} y G_{2.1}).

E ₁	S1	S2		E ₂	S1	S2
G _{1.1}	Tratamiento ₁ -Obj ₁	Tratamiento ₂ -Obj ₂		G _{2.1}	Tratamiento ₁ -Obj ₁	Tratamiento ₂ -Obj ₂
G _{1.2}	Tratamiento ₂ -Obj ₁	Tratamiento ₁ -Obj ₂				

	S1	S2
G _{1(1.1 y 2.1)}	Tratamiento ₁ -Obj ₁	Tratamiento ₂ -Obj ₂
G _{2(1.2)}	Tratamiento ₂ -Obj ₁	Tratamiento ₁ -Obj ₂

Figura 4.7 Diseño intra-sujetos incluido en un diseño crossover

4. *Dos diseños tienen intersección entre sí.* La figura 4.8 muestra un diseño entre-sujetos que interseca parcialmente en un diseño intra-sujetos. El Experimento₁ tiene dos factores: *Factor* y *Objeto experimental*, cada uno de ellos con dos niveles. El factor *Factor* posee los niveles *Tratamiento₁* y *Tratamiento₂* y el factor *Objeto experimental* posee los niveles *Objeto₁* y *Objeto₂*. Por su parte, el Experimento₂ tiene un único factor: *Factor* que posee los niveles *Tratamiento₁* y *Tratamiento₂* y además usa un único Objeto experimental el

Objeto₁. En este caso la superposición trae como consecuencia que el Experimento₂ interseca con el Experimento₁ y por tanto se añaden nuevos participantes (grupos) al diseño correspondiente, estos grupos se corresponden con participantes para los que se ha perdido alguna de las medidas.

E ₁	S1	S2	+	E ₂	S1
G _{1.1}	Tratamiento ₁ -Objeto ₁	Tratamiento ₂ -Objeto ₂		G _{2.1}	Tratamiento ₁ -Objeto ₁
				G _{2.2}	Tratamiento ₂ -Objeto ₁

=		S1	S2
	G _{1(1.1)}	Tratamiento ₁ -Objeto ₁	Tratamiento ₂ -Objeto ₂
	G _{2(2.1)}	Tratamiento ₁ -Objeto ₁	
	G _{3(2.2)}	Tratamiento ₂ -Objeto ₁	

Figura 4.8 Diseño entre-sujetos interseca en un diseño intra-sujetos

En la Tabla 4.16 se muestra el diseño que tienen todos los experimentos de la familia [EP14]. Dado que todos los experimentos tienen el mismo diseño, el diseño agregado coincide con el diseño individual de cada experimento.

	S1	S2
G _{1(1.1, 2.1, 3.1 y 4.1)}	ATAM-Objeto ₁	QuaDAI-Objeto ₂
G _{2(1.2, 2.2, 3.2 y 4.2)}	ATAM-Objeto ₂	QuaDAI -Objeto ₁
G _{3(1.3, 2.3, 3.3 y 4.3)}	QuaDAI-Objeto ₁	ATAM-Objeto ₁
G _{4(1.4, 2.4, 3.4 y 4.4)}	QuaDAI-Objeto ₂	ATAM-Objeto ₂

Tabla 4.16 Diseño de los experimentos de la familia [EP14]

4.3.2 Exploración de los Resultados de los Análisis Individuales

El diseño agregado proporciona el conjunto completo de variables de diseño que debe tenerse en cuenta a la hora de analizar cada grupo de experimentos. No obstante, en algún caso puede ser conveniente la eliminación de alguna variable de diseño. Esto puede ocurrir por varios motivos:

1. No es una variable de interés.
2. La variable no es significativa consistentemente a lo largo de los experimentos del grupo.

Nótese que en ningún caso se podrá eliminar el factor principal estudiado en la familia, por lo tanto, esto no aplica para él.

Para el segundo caso se utilizarán los resultados de los análisis individuales de cada experimento del grupo. El objetivo es obtener conclusiones iniciales sobre el comportamiento del (los) factor(es) del experimento a través de todos los experimentos en el grupo. Este análisis viene a desempeñar el mismo papel que los estadísticos descriptivos cuando se analizan experimentos individuales. Los parámetros que se explicarán son:

- *Resultados de las pruebas de hipótesis.* Cada experimento debe ser analizado como corresponda según su diseño, independientemente del análisis realizado para el resto de los experimentos del grupo. Hay que tener en cuenta que se pueden analizar diferentes modelos estadísticos en los distintos experimentos. Por ejemplo, los

experimentos de un factor con dos niveles se analizarán con un t-test, mientras que los experimentos de dos factores se analizarán con un ANOVA factorial. Del mismo modo, un experimento podría requerir la transformación de la variable respuesta debido a problemas de no normalidad o heterocedasticidad, o incluso realizar una prueba no paramétrica, mientras que otro podría analizarse con un ANOVA de un factor porque sí cumple las restricciones del modelo estadístico. En todos los casos, se debe seguir la estrategia de análisis adecuada, independiente de la estrategia seguida en otros experimentos en el grupo.

- *Intervalos de confianza.* Especialmente cuando se usan para la construcción de gráficos de perfil, son especialmente útiles para observar el comportamiento relativo de los tratamientos tanto de los efectos principales como de las interacciones.
- *Tamaño del efecto.* Es importante analizar la misma medida de tamaño de efecto para todos los experimentos del grupo bien con modelos paramétricos (como por ejemplo la *d* de Cohen) o no paramétricos. De modo que se puedan comparar todos los experimentos del grupo en los mismos términos.

Por lo tanto, para cada experimento se registra: la significación estadística, la dirección y tamaño del efecto, y los intervalos de confianza de los efectos principales y las interacciones que se estudian en un experimento específico. Con esta información se desarrolla una tabla, donde se resaltarán (en diferentes colores):

- Los términos que son estadísticamente significativos y el tamaño del efecto es relevante.
- Los términos que no son estadísticamente significativos, pero tienen un tamaño del efecto relevante.

La información recopilada sobre los resultados individuales de cada experimento incluye, además de las pruebas de hipótesis, los intervalos de confianza para la media y el tamaño del efecto. Esto se debe fundamentalmente a que los experimentos en IS suelen tener un tamaño muestral pequeño, lo que conduce a que en ocasiones la falta de significación se deba a ello.

Teniendo en cuenta la información sobre el tamaño del efecto y los intervalos de confianza se pueden obtener decisiones más informadas sobre la relevancia de una variable en el análisis estadístico.

La Tabla 4.17, muestra mediante un ejemplo de la información que se recopilará para cada uno de los experimentos. Se obtendrá la significancia estadística y práctica para cada uno de los factores, así como, las interacciones entre los mismos que procedan. Además, se recogerá su valor medio y la dirección del efecto. Si existe significancia estadística el p-valor aparecerá en rojo y si existe significancia práctica la celda estará marcada en fondo gris claro o gris oscuro, de esta manera es más sencillo identificar, los factores relevantes bajo estudio.

Nótese que cuando la variable estudiada tiene dos niveles el p-valor de la dirección del efecto no tiene sentido ya que este coincide con el p-valor de la variable.

Variable	p-valor	niveles	Intervalo de Confianza	Dirección efecto	p-valor	Tamaño efecto
Factor	Sig	Trat ₁ Trat ₂ Trat ₃	Valor ₁ Valor ₂ Valor ₃	Trat ₁ <Trat ₂ Trat ₁ <Trat ₃ Trat ₂ <Trat ₃	Sig Sig Sig	Despreciable [0-0,09]
Objeto Experimental	No sig	Obj ₁ Obj ₂	Valor ₃ Valor ₄	Obj ₁ <Obj ₂	No sig	Pequeño [0,1-0,23]
Factor * Objeto Experimental	Sig	Trat ₁ =>Obj ₁ Trat ₁ =>Obj ₂ Trat ₂ =>Obj ₁ Trat ₂ =>Obj ₂	Valor ₅ Valor ₆ Valor ₇ Valor ₈	Trat ₁ =>Obj ₁ <Obj ₂ Trat ₂ =>Obj ₁ <Obj ₂	Sig No sig	Mediano [0,24-0,36] Grande [0,37-1]

Tabla 4.17 Ejemplo de los resultados de los análisis individuales

Finalmente, con los resultados obtenidos para cada uno de los experimentos se creará una tabla como la Tabla 4.18 que reúna todas las variables estudiadas en cada experimento y si se ha obtenido significancia práctica y/o estadística, para facilitar la selección de variables de diseño en la actividad siguiente. Los casos en los que no existe ni significación estadística ni práctica se representarán con un guión, mientras que, si las variables no se han considerado en ese análisis, se representarán con un N/A.

Variable	Experimento ₁	Experimento ₂	Experimento ₃
Factor	Estadística Práctica	Estadística	-
Objeto Experimental	-	Práctica	N/A
Factor * Objeto Experimental	N/A	Estadística Práctica	N/A

Tabla 4.18 Ejemplo de obtención de significancia práctica y/o estadística de todos los experimentos

En la Tabla 4.19, Tabla 4.20, Tabla 4.21 y Tabla 4.22 se muestran los resultados de cada uno de los 4 experimentos que componen la familia [EP14], para las cinco variables respuesta que se estudian en la misma. Nótese que estos resultados son los proporcionados por los autores en el propio estudio, es decir, no se ha realizado ningún análisis ya que no se disponía de los datos crudos.

Variable Respuesta	Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Efectividad	Método	0,906	QuaDAI ATAM	0,684 0,626	ATAM<QuaDAI	N/A	No calculado
Eficiencia	Método	0,030	QuaDAI ATAM	0,029 0,019	ATAM<QuaDAI	N/A	No calculado
Facilidad de uso percibida	Método	0,000 0,003	QuaDAI ATAM	3,98 3,50	ATAM<QuaDAI	N/A	No calculado
Utilidad percibida	Método	0,000 0,000	QuaDAI ATAM	3,804 3,723	ATAM<QuaDAI	N/A	No calculado
Intención de uso	Método	0,000 0,000	QuaDAI ATAM	3,654 3,547	ATAM<QuaDAI	N/A	No calculado

Tabla 4.19 Resultados experimento UPV1 [EP14]

En estas tablas se muestran los resultados únicamente para el factor principal *Método de Evaluación de Arquitectura* y no se han proporcionado resultados ni para los objetos experimentales utilizados, ni para ninguna interacción, ni tampoco en ningún caso se ha calculado el tamaño del efecto. Finalmente, cuando la variable respuesta es subjetiva (facilidad de uso percibida, utilidad percibida e intención de uso) los autores las analizan a través de preguntas en escala de Likert que se realizan para cada uno de los niveles, es por ello, que existen dos p-valores para cada nivel.

Variable Respuesta	Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Efectividad	Método	0,036	QuaDAI ATAM	0,797 0,461	ATAM<QuaDAI	N/A	No calculado
Eficiencia	Método	0,043	QuaDAI ATAM	0,023 0,014	ATAM<QuaDAI	N/A	No calculado
Facilidad de uso percibida	Método	0,001 0,000	QuaDAI ATAM	4,312 4,020	ATAM<QuaDAI	N/A	No calculado
Utilidad percibida	Método	0,001 0,001	QuaDAI ATAM	4,167 3,927	ATAM<QuaDAI	N/A	No calculado
Intención de uso	Método	0,001 0,001	QuaDAI ATAM	4,063 3,906	ATAM<QuaDAI	N/A	No calculado

Tabla 4.20 Resultados experimento UPV2 [EP14]

Variable Respuesta	Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Efectividad	Método	0,003	QuaDAI ATAM	0,708 0,462	ATAM<QuaDAI	N/A	No calculado
Eficiencia	Método	0,006	QuaDAI ATAM	0,025 0,015	ATAM<QuaDAI	N/A	No calculado
Facilidad de uso percibida	Método	0,000 0,000	QuaDAI ATAM	3,59 3,59	ATAM<QuaDAI	N/A	No calculado
Utilidad percibida	Método	0,001 0,000	QuaDAI ATAM	4,07 3,85	ATAM<QuaDAI	N/A	No calculado
Intención de uso	Método	0,000 0,000	QuaDAI ATAM	3,935 3,740	ATAM<QuaDAI	N/A	No calculado

Tabla 4.21 Resultados experimento UPV3 [EP14]

Variable Respuesta	Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Efectividad	Método	0,008	QuaDAI ATAM	0,631 0,297	ATAM<QuaDAI	N/A	No calculado
Eficiencia	Método	0,045	QuaDAI ATAM	0,011 0,006	ATAM<QuaDAI	N/A	No calculado
Facilidad de uso percibida	Método	0,297 0,023	QuaDAI ATAM	2,667 2,444	ATAM<QuaDAI	N/A	No calculado
Utilidad percibida	Método	0,002 0,001	QuaDAI ATAM	3,569 3,458	ATAM<QuaDAI	N/A	No calculado
Intención de uso	Método	0,003 0,059	QuaDAI ATAM	3,458 3,417	ATAM<QuaDAI	N/A	No calculado

Tabla 4.22 Resultados experimento UNIBAS [EP14]

La Tabla 4.23 muestra la significancia obtenida para cada una de las variables estudiadas en los distintos experimentos que componen la familia [EP14]. Dado que los autores no proporcionan el tamaño del efecto, solo se muestra la significancia estadística. En UPV1 como el método de evaluación de arquitectura en el caso de la variable respuesta Efectividad no es significativo aparece un guión en la celda correspondiente. Finalmente, en esta familia se han estudiado las mismas variables en todos los experimentos.

Variable Respuesta	Variable	UPV1	UPV2	UPV3	UNIBAS
Efectividad	Método	-	Estadística	Estadística	Estadística
Eficiencia	Método	Estadística	Estadística	Estadística	Estadística
Facilidad de uso percibida	Método	Estadística	Estadística	Estadística	Estadística
Utilidad percibida	Método	Estadística	Estadística	Estadística	Estadística
Intención de uso	Método	Estadística	Estadística	Estadística	Estadística

Tabla 4.23 Resumen de la significancia para todos los experimentos de la familia [EP14]

4.3.3 Identificación del Conjunto de Variables de Diseño

El modelo del diseño agregado define las variables asociadas al diseño que se van a incluir en el análisis conjunto, y se crea a partir del diseño agregado. Este modelo parte de los factores que aparecen en el diseño agregado, junto con sus niveles. Hay que tener en cuenta que el modelo será intra-sujetos siempre y cuando exista un experimento que siga ese diseño, ya que los diseños intra-sujetos prevalecen entre los diseños entre-sujetos por ser estos últimos un caso especial de los primeros.

En este punto, se debe sopesar la posible simplificación del modelo del diseño agregado. La simplificación se hará en base a los resultados de los análisis individuales:

- Si hay experimentos crossover y no hay evidencia de la influencia del periodo, la secuencia o el carryover (al comprobar tanto la significación estadística como el tamaño del efecto) el diseño agregado se puede analizar como un diseño no crossover de medidas repetidas. En el caso de que exista evidencia de la influencia del carryover, se deben eliminar los datos asociados al segundo periodo y sucesivos en todos los experimentos intra-sujetos y el diseño agregado será un diseño entre-sujetos.
- Se puede considerar la posibilidad de eliminar los datos asociados a tratamientos u objetos experimentales que se estudian específicamente en un único experimento, ya que añadirán muy poco a la agregación. Por ejemplo, si un grupo de experimentos explora dos tratamientos, pero hay un único experimento que explora tres, los investigadores deben considerar la posibilidad de eliminar los datos asociados a este tercer tratamiento. O, si un grupo de experimentos trabaja con dos objetos experimentales diferentes, pero uno de los experimentos trabaja con cuatro (dos más que el resto de los experimentos), se debe considerar la posibilidad de eliminar los datos asociados a estos dos nuevos tratamientos. Es importante tener en cuenta que, esto debe hacerse con especial precaución y, en caso de duda, mantener todos los datos.

Nótese que en ningún caso pueden eliminarse el factor principal del experimento.

Continuando con el ejemplo de la familia [EP14], debido a que en los análisis individuales los autores solo proporcionaron los resultados para el factor principal, no se puede concluir si eliminar o añadir

alguna variable. De haberlo hecho se tendría que determinar si estudiar los objetos experimentales, la sesión, el grupo experimental e incluso interacciones de estas variables entre ellas y también con el factor principal. Pero en ningún caso se podría eliminar el método de evaluación de arquitectura ya que es el factor principal de la familia.

4.4 Identificación de las Variables Contextuales

4.4.1 Identificación de las Diferencias Contextuales

Para identificar las diferencias contextuales entre los experimentos del mismo grupo, se tienen en cuenta los elementos que aparecen reflejados en la Tabla 4.24. Estos son: todos los elementos de la población (excepto las características variables de los sujetos), la instrumentación (excepto los objetos experimentales), los experimentadores y de la dimensión de la ejecución experimental los elementos de la dimensión entrenamiento, más el elemento duración de la tarea de la dimensión del diseño.

Dimensión		Elemento
Ejecución Experimental	Diseño	Duración de la tarea
	Entrenamiento	Combinación entre entrenamiento y sesiones
		Aspectos de transmisión del tratamiento
		Duración del entrenamiento
Instrumentación		Guías suplementarias
		Formularios
		Transmisión de los tratamientos
		Infraestructura tecnológica
Población	Propiedades de los sujetos	Características comunes
	Propiedades de los objetos experimentales	Complejidad
		Dominio
		Lenguaje de programación
Experimentadores		Diseñador
		Entrenador
		Monitor
		Mecedor

Tabla 4.24 Elementos para la identificación de variables contextuales

La identificación de las diferencias contextuales se hace basándose en seleccionar aquellos elementos cuyos valores se cree podrían influir en los resultados experimentales. Es importante recalcar que esta actividad no se reduce a una mera identificación de aquellos elementos que tienen valores distintos en el grupo de experimentos. Ni la existencia de un cambio implica que se deba identificar automáticamente el cambio como una diferencia contextual, ni la ausencia del cambio implica que se identifique automáticamente que no hay diferencia contextual. Como se explica a continuación, puede haber situaciones en las que la ausencia de cambio conduzca a la identificación de diferencias contextuales, y la existencia de cambio conduzca a descartar una posible diferencia contextual:

- Elementos que tienen el mismo valor en diferentes experimentos podrían conducir a la identificación de diferencias contextuales. Por ejemplo, una familia de experimentos usa el idioma inglés en todos los experimentos, pero en alguno de ellos, los participantes no dominan el idioma. En este caso, la no existencia de cambio podría tener

implicaciones en los resultados experimentales, ya que los experimentos en los que los participantes no dominan el idioma inglés podrían tener diferentes resultados. Por lo tanto, aquí se debería identificar una diferencia contextual.

- Elementos que no tienen el mismo valor en diferentes experimentos podrían conducir a descartar diferencias contextuales. Por ejemplo, en un experimento la duración del entrenamiento podría ser diferente de otra. Pero el cambio ha sido hecho porque la experiencia y conocimientos de los participantes es diferente en ambos experimentos, y el cambio se ha adaptado a las necesidades de los participantes, y no se esperan implicaciones en los resultados experimentales. Por lo tanto, aquí no se debería identificar una diferencia contextual.

Por tanto, la identificación de diferencias contextuales entre experimentos se hará atendiendo al estudio de las posibles implicaciones en los resultados del cambio/no cambio.

La Tabla 4.25 muestra un ejemplo de la identificación de diferencias contextuales para la familia [EP14]. A continuación, se va a determinar si estos elementos podrían generar variables contextuales, de acuerdo con la información proporcionada por los autores:

Dimensión		Elemento	Valor	UPV1	UPV2	UPV3	UNIBAS
Ejecución Experimental	Diseño	Duración de la tarea	90 min (ilimitado)	x	x	x	x
	Entrenamiento	Combinación entre entrenamiento y sesiones	Secuencial	x	x	x	-
			Intercalado	-	-	-	x
		Duración del entrenamiento	1 sesión (120), 2 sesiones (30). Total: 180 min	x	x	x	-
			3 sesiones (60). Total: 180 min	-	-	-	x
Instrumentación		Guías suplementarias	Traducción	-	-	-	x
		Formularios	Preguntas de control	-	x	x	x
		Transmisión de los tratamientos	Traducción	-	-	-	x
			Traducción	-	-	-	x
			Electrónica	-	-	-	x
Población	Propiedades de los sujetos	Características comunes	Tipo de sujetos	Grado	x	-	x
				Máster	-	x	-
	Propiedades de los Objetos Experimentales	Complejidad	Modificación un nivel de un NFR	-	x	x	x
			Un tercer NFR y nueva métrica	-	-	-	x
			Traducción	-	-	-	x
Experimentadores		Entrenador	Investigador externo	-	-	-	x
		Monitor	Investigador externo	-	-	-	x

Tabla 4.25 Identificación de diferencias contextuales familia [EP14]

- *Duración de la tarea.* Este elemento no cambia en ninguno de los experimentos, sin embargo, no se puede determinar si esto podría afectar de manera distinta en algún experimento.
- *Combinación entre entrenamiento y sesiones.* Los autores no dan ninguna explicación del porqué de este cambio. A pesar de ello, se puede considerar esta diferencia como variable ya que los participantes que recibieron el entrenamiento de manera secuencial podrían haber olvidado el primer tratamiento enseñado, provocando un conocimiento desigual de los mismos.

- *Duración del entrenamiento.* Los autores no dan ninguna explicación del porqué de este cambio. Sin embargo, esta diferencia no se puede considerar como variable ya que, la duración total del training es la misma, la diferencia está en el tiempo empleado en cada sesión y ese valor estaría confundido con el elemento anterior.
- *Guías suplementarias y Transmisión del tratamiento.* En estos dos elementos, el único cambio existente, es la traducción del material, con respecto a esto los autores mencionan: “El material fue traducido del español al inglés, que era el idioma en el que se realizó la replicación. El inglés no era la lengua materna de los participantes. Esto puede haber dado lugar a amenazas de validez externa y de construcción debido a la familiaridad de los participantes con el idioma inglés. Sin embargo, todos los participantes tuvieron que aprobar un examen de idioma inglés para inscribirse en el segundo año de su programa de Grado, y el conocimiento de los participantes sobre el inglés fue, por lo tanto, casi homogéneo, lo que mitigó las posibles amenazas mencionadas anteriormente”. Por lo tanto, esta diferencia no se considerará como una variable contextual.
- *Formularios.* Este elemento presenta dos cambios, la traducción (ya mencionado anteriormente) y la inclusión de preguntas de control, con respecto a esto los autores mencionan: “Incluimos un conjunto de preguntas de control en el material experimental para analizar la comprensión de los patrones y las métricas que se aplican. Estas preguntas ayudaron a los participantes a centrarse en comprender los patrones y las métricas y nos permitieron controlar su comprensión del problema. Estas preguntas no influyeron en la ejecución del experimento y los resultados; su propósito era únicamente controlar la comprensión de los patrones y las métricas.” Por lo tanto, esta diferencia no se considerará como una variable contextual.
- *Infraestructura tecnológica.* Se utiliza la versión electrónica del material en lugar de las versiones impresas de los documentos. Los autores no dan ninguna explicación del porqué de este cambio. Sin embargo, esta diferencia no se puede considerar como variable ya que al parecer son las mismas versiones, pero no se han impreso.
- *Tipo de sujetos.* Hay dos tipos de sujetos utilizados, los autores mencionan: “nos centramos en el perfil de evaluadores novatos, ya que uno de nuestros objetivos es proporcionar un método de evaluación de arquitectura de software que ayudará a los evaluadores menos experimentados a realizar evaluaciones de arquitectura, reutilizando el conocimiento del experto en el dominio”. Por lo tanto, los dos tipos son considerados evaluadores novatos y esta diferencia no se considerará como una variable contextual.
- *Complejidad.* Este elemento presenta tres cambios: la modificación del NFR, inclusión de un tercer NFR y métrica software y la traducción (ya mencionado anteriormente). Con respecto a la modificación los autores mencionan: “También se cambió un nivel de un NFR en el objeto experimental O2 ya que en este objeto experimental fue más fácil encontrar la mejor solución (el 100% de los participantes) en comparación con el experimental objeto O1 (solo el 71% de los participantes, independientemente del método).” Y con respecto al tercer NFR y la métrica software ellos comentan: “también incluimos un tercer NFR y una nueva métrica de software que se aplicará para hacer que la compensación sea más compleja y menos obvia”. Por lo tanto, esta diferencia se considerará como una variable contextual.

- *Entrenador y monitor.* En ambos casos se utiliza un investigador externo, los autores mencionan: “para verificar los hallazgos obtenidos en los tres experimentos anteriores y para evitar cualquier sesgo de autor que pueda haber estado presente en los estudios anteriores”. Por lo tanto, esta diferencia se considerará como una variable contextual.

4.4.2 Creación de un Conjunto Inicial de Variables Contextuales

Las diferencias contextuales identificadas en la actividad anterior son el punto de partida para identificar las variables contextuales. Sin embargo, no necesariamente habrá una correspondencia uno a uno entre ellos. A veces, un grupo de elementos puede conducir a una única variable moderadora, ya que un grupo de cambios puede identificarse como uno solo. En concreto:

- Un único cambio puede causar una cascada de cambios no deseados. Por ejemplo, un cambio en el elemento del *entrenador* puede hacer que la *transmisión del tratamiento* de dimensión de la instrumentación también cambie.
- Varios cambios podrían estar interrelacionados (esto es, hechos a la vez por alguna razón común). Por ejemplo, todo el conjunto de experimentadores podría cambiar de un experimento a otro porque hay un cambio en el grupo de investigación.
- Finalmente, se pueden identificar variables que estén confundidas:
 - Con otras de las variables identificadas, es decir, variables que tienen la misma distancia o distribución en su conjunto de valores.
 - Con la variable Experimento, es decir, variables que toman un valor diferente para cada uno de los experimentos de la familia.

En la Tabla 4.26 se puede observar que las variables “Variable₁” y la “Variable₂” están confundidas entre sí, por lo tanto, estas se agruparán en una única variable. Es importante tener en cuenta que, de resultar significativas, no se sabrá cual de las dos variables iniciales está influyendo en los resultados. Por otro lado, la “Variable₃” está confundida con la variable experimento y no se podrá incluir en el modelo. Por lo tanto, las posibles variables a incluir serían: “Variable₁₋₂” (agrupación de 2 variables) y la “Variable₄”.

Estas cuestiones deben tenerse en cuenta como ya se ha mencionado, la complejidad involucrada en los modelos multinivel, así como la necesidad de que el modelo definido sea compatible con un número adecuado de puntos de datos en todos los niveles de la jerarquía, resultará en la selección del menor número posible de variables contextuales.

		Experimento ₁	Experimento ₂	Experimento ₃
Variable ₁	Valor 1.1	x	-	x
	Valor 1.2	-	x	-
Variable ₂	Valor 2.1	x	-	x
	Valor 2.2	-	x	-
Variable ₃	Valor 3.1	x	-	-
	Valor 3.2	-	x	-
	Valor 3.3	-	-	-
Variable ₄	Valor 4.1	x	x	-
	Valor 4.2	-	-	x

Tabla 4.26 Ejemplo de distribución de variables

Continuando con el ejemplo anterior, en la Tabla 4.27 se muestra el conjunto inicial de variables contextuales. En ella se puede observar que, la combinación entre entrenamiento y sesiones, el entrenador y el monitor, tienen la misma distribución y por lo tanto se agruparán en una única variable.

		UPV1	UPV2	UPV3	UNIBAS
Combinación entrenamiento y sesiones	Secuencial	x	x	x	-
	Intercalado	-	-	-	x
Complejidad	Sin modificación	x	-	-	-
	Modificación leve	-	x	x	-
	Modificación fuerte	-	-	-	x
Entrenador	Investigador interno	x	x	x	-
	Investigador externo	-	-	-	x
Monitor	Investigador interno	x	x	x	-
	Investigador externo	-	-	-	x

Tabla 4.27 Ejemplo de conjunto inicial de variables familia [EP14]

4.4.3 Refinamiento del Conjunto Inicial de Variables Contextuales

La última actividad de este paso consiste en conocer como de balanceadas están las variables seleccionadas, ya que esto también afectará al rendimiento del modelo multinivel. Para ello, es necesario observar como se distribuyen los valores de cada variable contextual en cada uno de los experimentos del grupo. Si las variables no están suficientemente balanceadas, es posible, que haya que considerar su eliminación del modelo por problemas en análisis.

Además, es conveniente conocer cuántos participantes hay en cada experimento, de esta manera, se puede conocer, por ejemplo, si existen valores perdidos en los experimentos con diseño Intra-sujetos.

La Tabla 4.28, muestra un ejemplo de la distribución y balanceo de variables. En ella se puede observar que la Variable₄ está bastante desbalanceada, ya que para uno de sus valores existen 230 observaciones, pero para el otro valor, solo hay 20 observaciones.

		Experimento	Sujetos	Observaciones
Variable ₁₋₂	Valor 1.1	Experimento ₁	45	90
		Experimento ₃	20	20
		Total	65	110
	Valor 1.2	Experimento ₂	70	140
Variable ₄	Valor 4.1	Experimento ₁	45	90
		Experimento ₂	70	140
		Total	115	230
	Valor 4.2	Experimento ₃	20	20

Tabla 4.28 Ejemplo de distribución de variables

Nuevamente, retomando el ejemplo anterior el de la familia [EP14] en la Tabla 4.29 se puede observar que la variable *ComEntMon*, que es la combinación de las variables que estaban confundidas, está bastante desbalanceada ya que para el valor *CEM₁* existen 320 observaciones, mientras que para el valor *CEM₂* existen tan solo 48 observaciones, es decir, 6 veces menos observaciones que para el valor *CEM₁*. Por otro lado, se puede observar que la variable *Complejidad* también está

desbalanceada, ya que nuevamente las observaciones para el valor *Modificación fuerte* tiene menos de la mitad de las observaciones que *Sin Modificación* y 4 veces menos observaciones que *Modificación leve* y, por lo tanto, *Sin Modificación* tiene poco más de la mitad de las observaciones que *Modificación leve*. Por lo tanto, en caso de ocurrir problemas en el análisis podría considerarse su eliminación.

		Experimentos	Sujetos	Observaciones
ComEntMon	CEM ₁	UPV1	28	112
		UPV2	16	64
		UPV3	36	144
		Total	80	320
	CEM ₂	UNIBAS	12	48
Complejidad	Sin modificación	UPV1	28	112
	Modificación leve	UPV2	16	64
		UPV3	36	144
		Total	52	208
	Modificación fuerte	UNIBAS	12	48

Tabla 4.29 Ejemplo de conjunto final de variables familia [EP14]

4.5 Identificación de las Variables de Participantes

4.5.1 Identificación de las Diferencias entre los Participantes

Muy a menudo, en experimentos en Ingeniería del Software, se recoge información sobre los participantes (típicamente experiencia o conocimientos). Los elementos que se deben tener en cuenta para determinar sus diferencias aparecen reflejados en la Tabla 4.30. Toda la información disponible de los participantes está contenida en el elemento *características variables de los sujetos* de la dimensión de la población. En esta actividad se han de recopilar las variables medidas para todos los experimentos.

Dimensión	Elemento
Población	Propiedades de los sujetos Características variables

Tabla 4.30 Elementos para la identificación de variables de participantes

Debido a que estas variables pueden tomar valores distintos para cada uno de los participantes, independientemente del experimento, la identificación de estas solo incluirá la escala en la que se miden y los experimentos en los que aparecen. La Tabla 4.31, muestra un ejemplo de esta situación.

		Exp ₁	Exp ₂	Exp ₃
Acuerdo Técnica	Escala de Likert (con 5 respuestas)	x	x	x
Experiencia Programación	Número de años	x	x	x
Profesión	Nominal	-	x	x
Asistencia Clases	Valor Binario	x	-	-
Puntaje	Calificación del curso	x	x	-

Tabla 4.31 Ejemplo de diferencias entre participantes

En este paso no se puede continuar con el ejemplo, proporcionando valores para la familia [EP14], ya que no han identificado diferencias entre los participantes, como ellos afirman, dado que los participantes no tenían experiencia de evaluación de arquitectura, no hubo diferencias entre ellos.

4.5.2 Creación de un Conjunto Inicial de Variables de Participantes

Se observan si todas las características de los participantes se han medido en todos los experimentos del grupo, en caso contrario, se las tienen que eliminar del modelo. Por otro lado, es necesario asegurarse que todas estas variables hayan sido medidas de la misma forma (escala), en caso contrario, se las pueden transformar para poder compararlas bajo las mismas condiciones, si no se pueden transformar se tienen que eliminar del modelo.

La Tabla 4.32 muestra un ejemplo de esta actividad, en ella se puede observar una situación en la que la variable que representa los conocimientos de los participantes con el lenguaje de programación Java, en el Experimento₁ ha sido medida en escala de Likert y podría tomar por ejemplo los valores “Alto”, “Medio” y “Bajo”, mientras que en el Experimento₂ y Experimento₃ esta variable se ha medido en número de años, por lo tanto, se podrían transformar la variable de estos dos experimentos para que se corresponda con la misma escala del Experimento₁. Por otro lado, la variable Profesión se eliminará ya que solo se ha medido en un experimento y no se puede transformar.

		Experimento ₁	Experimento ₂	Experimento ₃
Conocimiento de Java	Escala de Likert (A, M, B)	x	-	-
Conocimiento de Java	Número de años	-	x	x
Profesión	Nominal	-	x	-

Tabla 4.32 Ejemplo de conjunto inicial de variables de participantes

4.5.3 Refinamiento del Conjunto Inicial de Variables de Participantes

En esta actividad se quiere conocer como se distribuyen las variables en los distintos experimentos, debido a que, si las variables no están suficientemente equilibradas, es posible, que los resultados obtenidos no sean válidos. Debido a que estas variables pueden tomar un valor diferente para cada participante, se estudiarán a través de gráficos o diagramas:

- Variables discretas: se usarán diagramas de barras o histogramas
- Variables continuas: se usarán diagramas de dispersión

Además, estos diagramas permiten observar posibles conflictos con las otras variables de participantes o con las variables contextuales seleccionadas en el paso anterior. En caso de que las variables de participantes estén confundidas con las variables contextuales, en el modelo se mantendrán únicamente las variables contextuales.

Finalmente, se puede determinar si las variables de participantes identificadas deben tenerse en cuenta a este nivel, o si deben ser variables contextuales. Por ejemplo, si la variable experiencia de programación tiene un rango diferente de valores para diferentes experimentos, quizás tenga sentido

crear una nueva variable contextual y considerar la necesidad de eliminarla del conjunto de variables de participantes.

Por tanto, el conjunto inicial de variables de participantes se refina en función de su distribución, disponibilidad y posibles conflictos con otras variables.

4.6 Realización del Análisis

Los datos multinivel se diferencian de los datos de un único nivel, en que las observaciones individuales están anidadas dentro de los grupos de nivel más alto. La agregación de familias de experimentos en IS sigue un modelo multinivel (jerárquico) en naturaleza, ya que los participantes (sobre los que a menudo se toman medidas repetidas) están anidados dentro de los experimentos.

La Figura 4.9 muestra la estructura jerárquica de 2 niveles de una familia de experimentos con diseños entre-sujetos, donde el nivel 1 corresponde a los participantes en el experimento y el nivel 2 a los distintos experimentos. Por su parte la Figura 4.10 muestra la estructura jerárquica de 3 niveles de una familia de experimentos con diseños intra-sujetos, donde el nivel 1 corresponde a las medidas repetidas tomadas sobre los distintos participantes, el nivel 2 a los participantes del experimento y el nivel 3 a los distintos experimentos.

Figura 4.9 Ejemplo de una familia de experimentos multinivel con diseño entre-sujetos

Figura 4.10 Ejemplo de una familia de experimentos multinivel con diseño intra-sujetos

Los modelos multinivel reciben varios nombres en la literatura: modelos de coeficiente aleatorio, modelos de efecto mixto, modelos de regresión multinivel, modelos lineales jerárquicos y modelos de estructura de covarianza (o ecuación estructural) multinivel. Esto es un reflejo de la diversidad de campos en los que son utilizados.

En esta investigación se cree que el modelado multinivel es un enfoque adecuado para el análisis de familias de experimentos, por dos razones:

- Es importante reconocer la naturaleza jerárquica de los datos, porque la variable que representa cada nivel introduce dependencia en las observaciones. Esto significa que las unidades en un nivel dado serán más parecidas entre sí si pertenecen a la misma unidad de nivel superior. Por ejemplo, los participantes que pertenecen a un mismo experimento serán más homogéneos que los participantes que pertenecen a otros experimentos distintos. Como consecuencia, los residuos estarán correlacionados y, por lo tanto, no se cumplirá el supuesto principal de independencia de errores que usualmente utilizan los modelos lineales (como ANOVA o regresión). El modelado multinivel es capaz de examinar datos correlacionados, con heterocedasticidad, diseños no balanceados y datos perdidos.
- El núcleo de este tipo de procedimientos es la descomposición de la varianza en los resultados a través de los distintos niveles de la jerarquía y la explicación de esta varianza con las variables especificadas en cada nivel. Por lo tanto, permiten la incorporación de teorías sustanciales sobre variables asociadas a participantes y experimentos. Esto tiene un gran potencial para los análisis exploratorios (que es el tipo de análisis que se suelen hacer durante la agregación de experimentos), ya que los investigadores pueden examinar las relaciones entre la variable respuesta y las variables pertenecientes a distintos niveles (por ejemplo, factores, que son variables típicamente intra-experimentos y variables contextuales, que son variables entre-experimentos) y especificar interacciones entre niveles superiores e inferiores (por ejemplo, si la relación encontrada en un experimento entre la variable respuesta y el factor se mantiene para los otros experimentos).

Sin embargo, el modelado multinivel tiene algunas limitaciones. La principal desventaja es que los modelos pueden ser difíciles de (Goldstein, 1995):

- *Especificar*: Hay muchas variables que explorar y los modelos pueden complicarse bastante muy rápidamente.
- *Interpretar*: Debido a su complejidad, los modelos se pueden mal interpretar fácilmente.
- *Estimar*: Se necesitan tamaños de muestra adecuados en cada nivel para garantizar una potencia suficiente para detectar los efectos.

La Figura 4.11 muestra la estructura jerárquica que tiene la familia [EP14] en ella se ve que es una estructura de 3 niveles, un nivel superior podría ser el sitio, sin embargo, para UPV habría tres valores distintos mientras que para UNIBAS solo uno y, por lo tanto, esa jerarquía no estaría equilibrada.

Figura 4.11 Estructura jerárquica para la familia [EP14]

4.6.1 Identificación del Modelo de Análisis

Una vez identificadas las variables que van a incluirse en el análisis, es necesario construir el modelo jerárquico que soporte a la teoría bajo investigación y permita realizar los análisis. Como ya se ha comentado, en el caso de diseños entre-sujetos la jerarquía más sencilla tiene 2 niveles, donde el nivel 1 corresponde a los participantes y el nivel 2 a los experimentos. En el caso de experimentos con diseño intra-sujetos, la jerarquía tendría un nivel más, correspondiente a las medidas repetidas.

Asimismo, deben estudiarse las particularidades de la familia a la hora de establecer la jerarquía. La primera actividad para desarrollar este modelo consiste en identificar la estructura jerárquica de los datos.

En función de la complejidad estructural de la familia de experimentos, podrían añadirse nuevos niveles a la jerarquía. Por ejemplo, en el caso de una familia de experimentos llevados a cabo en varias universidades durante años académicos sucesivos. En este caso, por encima de los participantes habría un nivel que corresponde a los cursos académicos y por encima de este, uno que representaría a las universidades.

Una vez desarrollada la jerarquía, se deben establecer las variables que formarán parte del modelo de análisis. Dicho modelo está compuesto por las variables de diseño, las variables contextuales y las variables de participantes. Para completar el desarrollo del modelo, es necesario asociar cada variable del modelo a su nivel correspondiente: participante, experimento, sitio, etc. Tal y como se muestra en la Tabla 4.33.

Variable	Nivel
Variables de diseño	Intra-participante
Variables de participante	Participante
Variables contextuales	Experimento

Tabla 4.33 Ejemplo de asociación de variables a sus respectivos niveles

Asimismo, debe desarrollarse la teoría que subyace al modelo desarrollado, y las preguntas de investigación/hipótesis asociadas al mismo. Dicha teoría debe incluir preguntas relacionadas con:

- El comportamiento de las variables de diseño.
- El comportamiento de las variables de participantes.
- El comportamiento de las variables contextuales.
- El comportamiento del factor principal a lo largo de los distintos experimentos.

- El comportamiento de las posibles interacciones entre las variables de participante y el factor principal y las variables contextuales y el factor principal.

Las preguntas de investigación vienen impuestas por la manera en la que se analizan los modelos jerárquicos y, por lo tanto, siempre se estudian el mismo tipo de preguntas.

La Tabla 4.34 muestra la asociación de las variables identificadas en la familia [EP14] a sus respectivos niveles. Nótese que se podrían incluir más variables de diseño tales como la sesión o los programas utilizados, además como la familia no ha identificado ninguna variable de participante no existe ese nivel en la asociación.

Tipo	Variable	Nivel	Valores	Escala
Diseño	Efectividad	Intra-participante	0-1	Intervalo
	Eficiencia	Intra-participante	0-1	Intervalo
	Facilidad de uso percibida	Intra-participante	Likert (con 5 respuestas)	Nominal
	Utilidad percibida	Intra-participante	Likert (con 5 respuestas)	Nominal
	Intención de uso	Intra-participante	Likert (con 5 respuestas)	Nominal
	Método de Evaluación de Arquitectura	Intra-participante	QuaDAI ATAM	Intervalo
	Otras variables de diseño	Intra-participante	N/A	N/A
Contextual	ComEntMon	Experimento	CEM ₁ CEM ₂ Sin modificación	Nominal
	Complejidad	Experimento	Modificación leve Modificación fuerte	Nominal

Tabla 4.34 Asociación de variables a sus respectivos niveles para la familia [EP14]

Por lo tanto, las preguntas de investigación asociadas a esta familia serían:

- El comportamiento de las variables de diseño.
 - ¿Depende la efectividad del método de evaluación de arquitectura?
 - ¿Depende la eficiencia del método de evaluación de arquitectura?
 - ¿Depende la facilidad de uso percibida del método de evaluación de arquitectura?
 - ¿Depende la utilidad percibida del método de evaluación de arquitectura?
 - ¿Depende la intención de uso del método de evaluación de arquitectura?
- El comportamiento de las variables de participantes.
 - No aplica, porque no se han identificado variables de participante.
- El comportamiento de las variables contextuales.
 - ¿Depende la efectividad de las diferencias entre los experimentos representadas por ComEntMon y Complejidad?
 - ¿Depende la eficiencia de las diferencias entre los experimentos representadas por ComEntMon y Complejidad?
 - ¿Depende la facilidad de uso percibida de las diferencias entre los experimentos representadas por ComEntMon y Complejidad?
 - ¿Depende la utilidad percibida de las diferencias entre los experimentos representadas por ComEntMon y Complejidad?
 - ¿Depende la intención de uso de las diferencias entre los experimentos representadas por ComEntMon y Complejidad?

- El comportamiento del factor principal a lo largo de los distintos experimentos.
 - ¿Se comportan de forma distinta el método de evaluación de arquitectura en los diferentes experimentos?
- El comportamiento de las posibles interacciones entre las variables de participante y el factor principal y las variables contextuales y el factor principal.
 - ¿Interactúan las características de los experimentos con el método de evaluación de arquitectura?
 - No se han identificado variables de participante, por tanto, no se puede estudiar dicha interacción.

4.6.2 Análisis del Modelo

La heurística seguida a la hora de analizar modelos jerárquicos consiste en evolucionar desde un modelo muy simple a uno más complicado, hasta que se alcance el modelo completo especificado en el paso anterior. Para ello se sigue el proceso propuesto en (Heck, et al., 2014). El cual realiza el análisis utilizando Modelos Lineales Mixtos.

A continuación, se describe el proceso para realizar el análisis. La actividad de análisis en sí comienza con un modelo inicial que explica:

- *Variables de diseño.* Factor(es) y variable(s) de respuesta del diseño agregado. El factor(es) se definirá en este modelo como fijo.
- *La existencia de medidas repetidas.* Esto es necesario solo en el caso de grupos de experimentos que siguen un modelo intra-sujeto.

Este modelo de análisis corresponde a lo que se denomina comúnmente mega-análisis de IPD (Santos, et al., 2018), donde los datos de todos los experimentos en la familia se analizan conjuntamente como si provinieran de un solo experimento.

El ajuste general del modelo se evalúa mediante el AIC (del inglés, Akaike's information criterion). El AIC es un estimador de la calidad relativa de los modelos estadísticos para un conjunto dado de datos². No tiene sentido hablar de que su valor sea grande o pequeño per se, pero sus valores se pueden comparar. Por lo tanto, dada una colección de modelos para un conjunto de datos, el AIC proporciona un medio para la selección de modelos.

Se pueden utilizar otros criterios en lugar del AIC. Pero su principal ventaja con respecto al resto es la compensación que realiza entre la bondad de ajuste del modelo y su simplicidad. En otras palabras, el AIC se ocupa tanto del riesgo de sobreajuste como de la insuficiencia. Es importante tener en cuenta que los criterios de diferentes modelos se pueden comparar siempre que (Field, 2013):

1. Se use la opción de estimación de máxima verosimilitud completa para el análisis (en lugar de la probabilidad máxima restringida).
2. El nuevo modelo contiene todos los efectos del modelo anterior (es decir, es una extensión) o un subconjunto de ellos (es una particularización).

² Cuando se usa un modelo estadístico para representar un conjunto de datos, la representación casi nunca será exacta y se perderá cierta información. El AIC calcula la cantidad relativa de información perdida para un modelo dado: cuanto menos información pierde un modelo, mayor es su calidad.

Finalmente, hay que tener en cuenta que un modelo lineal mixto es válido cuando los residuos siguen una distribución normal. Además, para cada modelo, se utilizarán distintas estructuras de covarianza (las más comunes: identidad, diagonal, auto-regresiva de primer orden, simetría compuesta y no estructurada). La estructura de covarianza es importante porque se la utiliza como punto de partida para estimar los parámetros del modelo. Si especifica una estructura de covarianza que es demasiado simple, entonces es más probable que se cometa un error de Tipo I (encontrar que un parámetro es significativo cuando en realidad no lo es), por el contrario, si especifica una estructura de covarianza que es demasiado compleja, se corre el riesgo de cometer un error de Tipo II (encontrar parámetros que no son significativos cuando en realidad lo son) (Field, 2013).

El modelo inicial se ampliará en iteraciones sucesivas, de cada modelo se elegirá aquel que tenga el AIC más bajo entre todas las matrices de covarianza, siempre que exista normalidad en los residuos y no haya ocurrido ningún error con la matriz de covarianza. El modelo se ampliará en el siguiente orden:

1. *Incorporar la estructura jerárquica de los datos.* Los participantes se anidan dentro de experimentos, los experimentos se pueden anidar dentro de los sitios, etc. En este modelo se refleja la estructura definida en la actividad 4.1.
2. *Incorporar las variables moderadoras* como factores fijos asociadas a su nivel correspondiente. Las variables identificadas en todos los niveles y las posibles interacciones entre las variables del mismo nivel podrían afectar a la variable respuesta.
3. *Permitir la variación del comportamiento del factor principal experimental*, especificándolo como factor aleatorio (en lugar de fijo). El comportamiento relativo de los tratamientos podría variar entre los experimentos. Por ejemplo, debido a que el tamaño del efecto podría cambiar en dos experimentos, puede ocurrir que en un experimento dos tratamientos sean iguales, mientras que, en otro diferente, un tratamiento sea más efectivo que el otro.
4. *Incorporar posibles interacciones entre variables de diferentes niveles.* Por ejemplo, las variables de nivel de experimento podrían interactuar con los tratamientos de experimento.

Finalmente, para cada modelo debe verificarse la normalidad de los residuos, y el AIC más bajo entre las matrices de covarianza debe compararse con el AIC del modelo anterior. Cada vez que obtenemos un AIC más alto, significa que el modelo tiene un peor ajuste y, por lo tanto, se debe volver al modelo con el mejor AIC y proceder con el análisis. Por ejemplo, si el paso 4 (especificar el factor principal del experimento como factor aleatorio) genera un AIC peor que el paso 3, entonces el factor principal del experimento se eliminará de ser un factor aleatorio, y se continuará con el paso 5.

La Figura 4.12 muestra un diagrama que resume los pasos seguidos para conseguir el modelo de análisis final que responderá a las preguntas de investigación planteadas en la familia.

De ahora en adelante, no se puede continuar usando la familia [EP14] como ejemplo, porque no se dispone de los datos.

Figura 4.12 Iteraciones para obtener un modelo de análisis final

4.6.3 Interpretación de los Resultados

Una vez que se haya llegado a un modelo final, se pueden responder las preguntas de investigación identificadas en la primera actividad de este paso. Para este propósito, se tendrán en cuenta tanto las características del modelo como la significación estadística de las variables analizadas, y su tamaño de efecto e intervalos de confianza. Es decir, es igual que un experimento único.

A través de los resultados obtenidos de la ejecución del procedimiento MIXED se estudiarán aquellas variables en las que exista significancia estadística ($<0,05$). Para estas variables se realizará la misma tabla propuesta para los análisis individuales (Tabla 4.17) en la que se mira la media y la dirección y tamaño del efecto para cada una de las variables significativas.

Si las variables se han incorporado al modelo como factores, el tamaño del efecto es r , y está calculado como indica Field en el libro (Field, 2013). Por otro lado, si las variables se han incorporado como covariables, entonces el tamaño del efecto es b , este valor indica como es la

pendiente de la recta, es decir, si toma valores positivos la variable tiene dirección creciente, si por el contrario toma valores negativos entonces la dirección de la variable es decreciente, finalmente, si b es 0 entonces la variable es constante.

4.6.4 Interpretación Conjunta de los Resultados de los Distintos Grupos

Finalmente, en el caso de que en el primer paso del proceso se hayan creado diferentes grupos de experimentos, debido a que los experimentos tuvieran diferentes operacionalizaciones, se deben combinar los resultados de cada grupo para poder interpretarlos conjuntamente.

Para esta actividad se propone la utilización de Meta-análisis de AD.

4.7 Resumen del Proceso Propuesto

La Tabla 4.10 muestra un resumen del proceso propuesto, en ella se puede observar los cinco pasos del proceso con sus respectivas actividades, las entradas que se requieren en una de ellas y las técnicas y herramientas que se han propuesto para generar las salidas correspondientes.

Paso	Actividad	Entradas	Salidas	Técnica	Herramienta
Identificación de los Grupos de Agregación	Caracterización de los experimentos de la familia	Información sobre los experimentos	Experimentos caracterizados	Esquema de caracterización	Plantilla de caracterización (Tabla 4.3)
	Identificación de grupos de experimentos a agregar	Experimentos caracterizados	Grupos de agregación	Heurísticas para identificar experimentos con constructos similares	Elementos en gris claro en la Tabla 4.3
Identificación de variables de Diseño	Creación de un diseño agregado	Cada uno de los grupos de agregación	Diseño agregado	Heurísticas para agregación de experimentos	N/A
	Exploración de los resultados de los análisis individuales	Cada uno de los grupos de agregación	Variables de diseño	Significación estadística, dirección y tamaño del efecto e intervalos de confianza	Paquete estadístico
	Identificación del conjunto de variables de diseño	Diseño agregado y variables de diseño	Conjunto final de variables de diseño	Heurísticas para identificar el conjunto de variables de diseño	N/A
Identificación de las variables contextuales	Identificación de diferencias contextuales	Cada uno de los grupos de agregación	Diferencias contextuales	Diferencias contextuales con impacto en la variable respuesta	Elementos en gris intermedio en la Tabla 4.3
	Creación de un conjunto inicial de variables contextuales	Diferencias contextuales	Variables contextuales iniciales	Correspondencia entre diferencias y variables	N/A
	Refinamiento del conjunto inicial de variables contextuales	Variables contextuales iniciales	Variables contextuales definitivas	Variables confundidas entre sí o con la variable experimento	N/A
Identificación de las variables de participantes	Identificación de diferencias entre los participantes	Cada uno de los grupos de agregación	Diferencias entre los participantes	Características individuales de los participantes	Elementos en gris oscuro en la Tabla 4.3
	Creación de un conjunto inicial de variables de participantes	Diferencias medidas entre los participantes	Variables de participantes iniciales	Disponibilidad de las variables en todos los experimentos	N/A
	Refinamiento del conjunto inicial de variables de participantes	Variables de participantes iniciales y variables contextuales definitivas	Variables de participantes definitivas	Histogramas, gráficos de barras y dispersión	Paquete estadístico
Realización del Análisis	Identificación del modelo de análisis	Variables de diseño, de participantes y contextuales	Modelo de análisis	Estructura jerárquica de los datos	N/A
	Análisis del modelo	Modelo de análisis de datos	Resultados del análisis	Proceso iterativo de análisis	Paquete estadístico
	Interpretación de los resultados	Resultados del análisis	Interpretación de los resultados de cada grupo	Interpretación de resultados como en experimentos únicos	N/A
	Interpretación conjunta de los resultados de los distintos grupos	Resultados del análisis de cada grupo de agregación	Interpretación de los resultados de todos los grupos	Meta-análisis de AD	Paquete estadístico

Tabla 4.35 Resumen de proceso propuesto

Capítulo 5

ESTUDIO DE LA VIABILIDAD DE LA PROPUESTA: AGREGACIÓN DE UNA FAMILIA DE EXPERIMENTOS DE TÉCNICAS DE PRUEBA

En este capítulo se estudia la viabilidad del proceso propuesto en el capítulo anterior aplicándolo a una familia de experimentos de tamaño pequeño.

Los resultados de este capítulo se encuentran en la publicación:

- On (Mis)Perceptions of Testing Effectiveness: An Empirical Study. Empirical Software Engineering Journal, pp. 1-59.

5.1 Descripción de la Familia de Experimentos

La familia a analizar se compone de cuatro experimentos, los cuales fueron realizados durante los años 2004 y 2005. Dos de los experimentos se llevaron a cabo en la Universidad Politécnica de Madrid (UPM), los otros dos experimentos se realizaron: uno de ellos en la Universidad Politécnica de Valencia (UPV) y el otro en la Universidad ORT Uruguay (ORT).

El objetivo de esta familia de experimentos consiste en estudiar la efectividad de tres técnicas de evaluación de código: una estática (*lectura por abstracciones sucesivas*) y dos dinámicas (una estructural *-cobertura de decisión-* y otra funcional *-partición en clases de equivalencia-*). La hipótesis nula de esta familia de experimentos es:

H_0 : *No hay diferencia en la efectividad de las técnicas: lectura por abstracciones sucesivas, cobertura de decisión y partición en clases de equivalencia.*

De ahora en adelante se explica el experimento UPM, ya que es el experimento base, y posteriormente en la sección 5.1.7, se comentan las diferencias entre el resto de los experimentos.

5.1.1 Técnicas Utilizadas

La **técnica** es un factor del experimento con tres tratamientos (o niveles): lectura por abstracciones sucesivas (R), cobertura de decisión (S) y partición en clases de equivalencia (F).

- La técnica *lectura por abstracciones sucesivas* aparece descrita en (Linger, et al., 1979). Los participantes comienzan en el nivel más bajo (más detallado) y reemplazan las líneas consecutivas de código del programa por una abstracción o especificación. Los participantes repiten el proceso hasta que hayan abstraído todo el código fuente. Finalmente, los participantes comparan la especificación del programa con la abstracción que han generado. Las discrepancias que se encuentren reflejan las inconsistencias entre comportamiento real del programa específico y el esperado.
- La técnica *cobertura de decisión* aparece descrita en (Beizer, 1990). En esta técnica han de generarse casos de prueba, de tal forma que todas las decisiones del código tomen el valor verdadero o falso. Una decisión es una instrucción IF, una declaración de control de bucle (DO-WHILE o REPEAT-UNTIL), o una instrucción CASE, donde hay dos o más resultados a partir de la declaración.
- La técnica de *partición en clases de equivalencia* se describe en (Myers, et al., 2004). Esta técnica consta de dos pasos: el primero consiste en identificar un conjunto de clases de equivalencia para posteriormente definirse a partir ellas los casos de prueba. Las clases de equivalencia se identifican a través de las condiciones de entrada de un programa (normalmente una frase o declaración en la especificación). Se generan dos tipos de clases de equivalencia: las clases válidas, que representan los valores válidos para las entradas del programa, y las clases no válidas, que representan los valores erróneos para las entradas del programa. Las clases de equivalencia se identifican basándose en una serie de heurísticas predefinidas. A la hora de generar los casos de prueba, se debe garantizar que se prueben todas las clases al menos una vez. En un caso de prueba pueden aparecer varias clases válidas y una clase no válida como máximo.

5.1.2 Programas

El experimento utiliza tres **programas** diferentes sobre los que se aplican las técnicas.

Los programas utilizados están escritos en el lenguaje de programación C. Aunque el propósito de los programas es diferente, todos son similares en tamaño y complejidad. A continuación, se explica brevemente cada uno:

- Cmdline: Programa que recibe como entrada una línea de comandos, la analiza y devuelve un resumen de esta. Tiene 289 líneas de código y una complejidad ciclomática de 37.
- Nametbl: Programa que implementa la estructura de datos y las operaciones de una tabla de símbolos. Tiene 289 líneas de código y una complejidad ciclomática de 27.
- Ntree: Programa que implementa la estructura de datos y las operaciones de un árbol n-ario. Tiene 247 líneas de código y una complejidad ciclomática de 31.

El programa no es un factor, ya que el objetivo del experimento no es estudiar los programas, sino las técnicas de evaluación del código. Sin embargo, es una variable de bloque, porque los programas podrían estar influyendo en los resultados. El experimento ha sido diseñado para cancelar la influencia de los programas. Cada participante aplica cada técnica en un programa diferente, y cada técnica se aplica en diferentes programas (por diferentes participantes).

5.1.3 Faltas

A cada programa, se le han insertado artificialmente siete faltas. Las faltas se consideran por *omisión* si algo no se ha especificado o por *comisión* si algo es incorrecto (Basili, 1996). Se han usado los siguientes 4 tipos:

- Inicialización. Inicialización incorrecta de una estructura de datos:
 - Error por omisión (F3): la variable no se inicializa.
 - Error por comisión (F4): asignar un valor incorrecto a una variable.
- Control: El programa sigue un flujo de control incorrecto.
 - Error por omisión (F6): falta el bloque entero de sentencias en una estructura “if-then-else”.
 - Error por comisión (F5): el bloque de sentencias es incorrecto en una sentencia “if-then-else”.
- Cálculo: Estas faltas conducen a un cálculo incorrecto.
 - Error por comisión (F7): hay una expresión aritmética incorrecta en el lado derecho de una asignación.
- Estéticas: Estas faltas conducen a un error en la interfaz de usuario.
 - Error por omisión (F1): no aparece un mensaje de error cuando debe.
 - Error por comisión (F2): el mensaje de error contiene una falta de ortografía o de gramática.

Algunas fallas ocurrieron naturalmente cuando se codificaron los programas, mientras que otras son faltas típicas de programación. Se crearon 2 versiones de cada programa, los cuales incluían faltas diferentes, con el objetivo de replicar cada falta dos veces. Todas las faltas:

- Causan fallos visibles.
- Pueden ser detectadas por todas las técnicas.
- Se han elegido para que los programas fallen solo con determinadas entradas.
- Ninguna falta está enmascarada, excepto en una de las versiones de uno de los programas que contiene solo 6 faltas, ya que uno de los errores estaba enmascarado.
- Hay una relación unívoca entre falta y fallo.

Sin embargo, hay que tener en cuenta que es posible que un participante genere dos (o más) casos de prueba que ejerzan la misma falta y, por lo tanto, produzcan el mismo fallo. Se ha aconsejado a los participantes que informen estas faltas como una sola. Por ejemplo, hay un error en el programa ntree en la función encargada de imprimir el árbol. Esto provoca el fallo de que el árbol no se imprime correctamente. Cada vez que un participante genera un caso de prueba que

imprime el árbol (lo cual es bastante frecuente, ya que esta función es útil para verificar el contenido del árbol en cualquier momento), se mostrará el fallo.

5.1.4 Variable respuesta

La variable respuesta del experimento es la **efectividad** de las técnicas, esta variable se mide de dos formas distintas dependiendo de la técnica utilizada:

- Para la técnica de revisión de código, se mide como el porcentaje de faltas correctamente reportadas por cada participante (se descartan los falsos positivos).
- Para las técnicas funcional y estructural, se mide como el porcentaje de faltas ejercidas por el conjunto de casos de prueba generados por cada participante. Para medir la variable respuesta, los experimentadores ejecutan los casos de prueba generados por cada participante.

Tenga en cuenta que las técnicas dinámicas y la técnica de revisión de código no son directamente comparables, ya que las primeras encuentran fallos mientras que la última encuentra faltas.

5.1.5 Participantes

Dos de los experimentos se llevaron a cabo en la Universidad Politécnica de Madrid (UPM) durante los cursos académicos 2003/2004 y 2004/2005, con alumnos de quinto curso de Escuela Técnica Superior de Ingenieros Informáticos. Los experimentos formaban parte de la asignatura “Evaluación de Sistemas de Información”, en la cual se enseñaban conceptos y técnicas de verificación y validación de Software. Aunque los participantes tienen poca o ninguna experiencia profesional, tienen suficiente conocimiento de programación. No tienen capacitación formal en ninguna técnica de evaluación de código, ya que este es el curso en el que se les enseña.

Se realizaron otros dos experimentos. Uno de ellos en la Universidad Politécnica de Valencia en España (UPV) durante el curso académico 2004/2005, con alumnos de quinto curso de la Facultad de Informática.

El otro experimento se realizó en la Universidad ORT Uruguay (ORT) durante el curso académico 2004/2005, con alumnos de segundo y tercer curso (cuarto y quinto semestre) que cursaban las carreras “Ingeniería en Sistemas” y “Licenciatura en Sistemas”. El experimento formaba parte de la asignatura “Ingeniería del Software 1”, que es la primera asignatura en que se tratan temas de ingeniería de software, por lo tanto, los participantes no tienen conocimientos previos específicos sobre técnicas de prueba de software y tienen pocos conocimientos de programación.

5.1.6 Procedimiento experimental

El procedimiento experimental consta de siete semanas, tal y como se resume en la Tabla 5.1. En las primeras tres semanas se realizan las sesiones de entrenamiento en las cuales los participantes aprenden cómo aplicar las técnicas y practican con ellas. Se realiza dos sesiones por semana (martes y jueves) y cada una dura 2 horas, por lo tanto, el entrenamiento dura 12 horas (2 horas / sesión x 2

sesiones / semana x 3 semanas). Primero se les enseña a los participantes la técnica de revisión del código, luego la caja blanca y finalmente la caja negra. El entrenamiento no sigue ningún orden en particular, pero es el que mejor se adapta para cumplir con los objetivos de aprendizaje del curso.

Tipo	Entrenamiento			Libre	Operación		
Semana	S1	S2	S3	S4	S5	S6	S7
Actividad	E	R	F	Ejercicios	ntree	cmdline	nametbl
Número de sesiones	2	2	2	-	1	1	1
Duración de la sesión	2h	2h	2h	-	4h	4h	4h
Tiempo Total	4h	4h	4h	4h	4h	4h	4h

Tabla 5.1 Procedimiento experimental

La semana siguiente se les pide a los estudiantes que practiquen con las técnicas. Para este propósito, se les dan 3 pequeños programas en C (que contienen faltas), y se les pide que apliquen una técnica dada en cada programa (todos los estudiantes aplican la misma técnica en el mismo programa de entrenamiento). El rendimiento de estos ejercicios se utiliza para evaluarlos.

Las otras tres semanas se realiza la ejecución del experimento. Cada sesión de ejecución del experimento tiene lugar una vez a la semana (viernes) y dura cuatro horas. Esto es equivalente a que no hay límite de tiempo, ya que los participantes pueden completar la tarea en menos tiempo. Por lo tanto, la ejecución del experimento lleva 12 horas (4 horas / sesión x 1 sesión / semana x 3 semanas).

En cada sesión, los participantes aplican individualmente las técnicas y, para las técnicas dinámicas, también ejecutan los casos de prueba. Los participantes aplican todas las técnicas una vez y trabajan en cada programa una vez. Por lo tanto, un participante aplica cada técnica a un programa. En cada sesión los participantes aplican una técnica a un programa, pero siempre aplican técnicas diferentes. Durante cada sesión, se utilizan las tres técnicas y ellas se aplican al mismo programa, de esta manera se cubren todos los posibles órdenes de aplicación de la técnica. Los participantes realizan las siguientes tareas:

- Los participantes que aplican particiones en clase de equivalencia reciben la especificación del programa para diseñar los casos de prueba. Luego, reciben una versión ejecutable del programa y ejecutan los casos de prueba. Los participantes identifican faltas en términos de salidas incorrectas.
- Los participantes que aplican cobertura de decisiones reciben el código fuente sin su especificación para diseñar casos de prueba. Luego, se les entrega una versión ejecutable del programa para ejecutar los casos de prueba. A continuación, se les proporciona la especificación del programa para verificar las salidas del caso de prueba. Los participantes identifican faltas en términos de salidas inválidas.
- Los participantes que aplican la técnica lectura de código por abstracciones sucesivas reciben el código fuente del programa para identificar abstracciones y generar una especificación del programa. Luego se les entrega la especificación original para identificar las fallas en el programa. Los participantes identifican faltas por inconsistencias entre las especificaciones: abstracta y original.

Por lo tanto, los participantes generan casos de prueba y fallos (para las técnicas dinámicas) o faltas (para la técnica de revisión de código).

5.1.7 Cambios entre los experimentos

Cada sitio tiene algunas condiciones contextuales que afectan al diseño experimental y, por lo tanto, todos los diseños no son iguales. La Tabla 5.2 resume las principales diferencias entre estos experimentos y, a continuación, se explican detalladamente las condiciones contextuales.

Exp.	Cambio	Propósito
UPM-2	Orden de aplicación de los programas	Eliminar confusión día-programa
UPV	No se utiliza Revisión de código	Limitaciones de tiempo
	Sesión más corta (2h)	
	2 sesiones experimentales	
	Ejecutan casos de prueba para un programa	Participantes familiarizados con las técnicas
	Cambio de diseño experimental	
ORT	Entrenamiento más corto (2 sesiones de 3h)	Restricciones del contexto
	Entrenamiento intercalado	
	No se utiliza Revisión de código	Limitaciones de tiempo
	No se utiliza ntree	
	1 sesión experimental con 2 períodos	Cubrir todos los órdenes de aplicación de los tratamientos
	Cambio de diseño experimental	
	No se ejecutan los casos de prueba	Sin acceso a ordenadores

Tabla 5.2 Procedimiento experimental

5.1.7.1 UPM-2

Los experimentos examinan las tres técnicas y utilizan los tres programas. En un total de tres sesiones los participantes aplican las técnicas y en cada sesión los participantes aplican una técnica a un programa, pero siempre aplican técnicas diferentes. Durante cada sesión, se utilizan las tres técnicas y ellas se aplican al mismo programa, debido a que el día y el programa están confundidos los participantes aplican programas en diferentes órdenes en cada uno de los experimentos: en UPM-1 el orden de los programas es cmdline-ntree-nametbl, mientras que en UPM-2 es orden es ntree-cmdline-nametbl.

5.1.7.2 UPV

Por restricciones de tiempo, no se utiliza la técnica revisión de código, aunque si se utilizan los tres programas. El experimento se ejecuta en tres sesiones de dos horas cada una. Los participantes aplican individualmente una técnica a un programa durante las dos primeras sesiones y en la tercera sesión ejecutan los casos de prueba que obtuvieron para uno de los programas. En cada sesión, los participantes aplican la misma técnica a diferentes programas, por lo tanto, en este experimento la técnica está confundida con el día. Como los participantes ya están familiarizados con las técnicas, el entrenamiento consiste en dos breves tutoriales de dos horas, que se realizan de manera intercalada, es decir, cada tutorial se lleva a cabo antes de la aplicación de cada técnica.

5.1.7.3 ORT

Por restricciones de tiempo no se utiliza la técnica revisión de código ni tampoco el programa ntree. Aunque el experimento se ejecuta en una única sesión, existen dos períodos. Los participantes aplican

individualmente las dos técnicas a los dos programas, y no tienen restricción de tiempo, por tanto, la sesión es ilimitada. En cada período se aplican las dos técnicas y se utilizan los dos programas, por lo tanto, se pueden cubrir los cuatro posibles órdenes de aplicación de los tratamientos. Los participantes nunca realizan casos de prueba, ya que no tienen acceso a ordenadores. Los participantes reciben tres sesiones de entrenamiento de cuatro horas para aprender a aplicar las técnicas antes de ejecutar el experimento.

5.2 Aplicación del Proceso Propuesto a la Familia de Experimentos

5.2.1 Identificación de los Grupos de Agregación

5.2.1.1 Caracterización de los Experimentos de la Familia

Los cuatro experimentos de la familia se caracterizan según lo dispuesto en Tabla 4.11. Las tablas 5.3 y 5.4 muestran la caracterización de UPM. Los demás experimentos de la familia aparecen caracterizados en el Apéndice C.1.

Dimensión	Elemento	Valor
Operacionalización	Institución	UPM
	Sitio	Academia
	Publicación	(Juristo, et al., 2012)
	Factores	Técnica
	Causa	Revisión
		Estructural
		Funcional
		Lectura de código por abstracciones sucesivas (Linger, et al., 1979)
		Cobertura de sentencias (Beizer, 1990)
	Tratamiento como instrucciones	Particiones en clases de equivalencia (Myers, et al., 2004)
		Revisión de código
		Cobertura de decisión
	Proceso de aplicación del tratamiento	Partición en clases de equivalencia
		Procedimiento revisión de código
		Procedimiento cobertura de decisiones
	Efecto	Procedimiento particiones en clases de equivalencia
		Variable respuesta
		Efectividad
		Elaboración de la variable respuesta
	Métrica	- Revisión: $\frac{FaltasReportadas}{FaltasTotal} * 100$
		- Estr. y Func.: $\frac{Faltas(CasoPrueba)}{FaltasTotal} * 100$
Medición	Instrumentos de medición	Ordenador, Folios, Bolígrafos (manual)
	Procedimiento de medida	Criterio Relajado

Tabla 5.3 Caracterización del experimento UPM-1 (1/2)

Dimensión		Elemento	Valor
Ejecución experimental	Diseño	Diseño experimental	Crossover
		Número de sesiones	3
		Nº de factores por sesión	1
		Número de tratamientos por factor por sesión	3
		Variables de bloque	Programa
		Orden de los tratamientos (grupos)	R-F-E
			R-E-F
			E-R-F
			E-F-R
			F-R-E
			F-E-R
		Objetos experimentales por sesión	1
		Número de unidades experimentales	108
		Número de sujetos	32
		Duración de la tarea	Ilimitado
	Entrenamiento	Combinación entrenamiento y sesiones	Secuencial
		Aspectos de transmisión del tratamiento	Experiencia alta en enseñanza de técnicas de prueba
		Duración del entrenamiento	3 sesiones de 4 horas
	Instrumentación	Guías suplementarias	Guías
		Formularios	Formularios
		Transmisión de los tratamientos	Instrucciones
		Objetos experimentales	Programas
			Cmdline
			Nametbl
		Faltas	Ntree
			Instancia 1
			Instancia 2
		Infraestructura tecnológica	Ordenador
Población	Sujetos	Características comunes	Tipo de sujetos
			Estudiantes 5º año
			Experiencia con la técnica
			No familiarizado
			Conocimiento en programación
			Alta
		Características variables	Motivación
			Alta
	Objetos	Exp. Relativa	Exp. Absoluta
			Escala de Likert
		Complejidad	Número de años
			Media-baja 15-29
	Experimentadores	Dominio	Alta > 45
			Tareas de juguete
			Lenguaje de programación
	Experimentadores	Medidor	C
			Diseñador
			Juristo, Vegas
			Entrenador
	Experimentadores	Monitor	Juristo, Vegas
			Vegas, Moreno
	Experimentadores	Medidor	Vegas
			Vegas

Tabla 5.4 Caracterización del experimento UPM-1 (2/2)

5.2.1.2 Identificación de los Grupos de Experimentos a Agregar

Para establecer aquellos experimentos que pueden ser agregados, es necesario examinar su operacionalización. La tabla 5.5 muestra la operacionalización de los experimentos de la familia. En ella se puede observar que:

- *Los elementos de la dimensión efecto son siempre los mismos.* Todos los experimentos comparten la misma variable respuesta (efectividad). Asimismo, tanto la elaboración de la variable respuesta como la métrica son las mismas para los cuatro experimentos. Nótese que en el caso de UPV y ORT, al no trabajar con la técnica revisión de código, no se define para ella ni la elaboración de la variable respuesta, ni su métrica.
- *Los elementos de la dimensión medición son los mismos para todos los experimentos.* En todos los casos los instrumentos de medida utilizados son ordenador, folios y bolígrafos dado que la variable respuesta se mide manualmente y el procedimiento de medición sigue un proceso relajado.
- *Hay al menos un factor en común en la dimensión de causa.* Dos de los 3 tratamientos (técnicas estructural y funcional) son comunes a los 4 experimentos y el tercero (técnica revisión de código) aparece en 2 experimentos (UPM-1 y UPM-2). Asimismo, el proceso de aplicación de los tratamientos, sus instrucciones y la referencia bibliográfica seguida es el mismo para cada tratamiento en cada uno de los experimentos en que aparece.

Por tanto, **se establece un único grupo de agregación formado por los cuatro experimentos de la familia.**

OPERACIONALIZACIÓN			UPM-1	UPM-2	UPV	ORT
Causa	Factores	Técnica	x	x	x	x
	Tratamientos	Revisión	x	x	-	-
		Estructural	x	x	x	x
		Funcional	x	x	x	x
	Tratamiento acorde a la literatura	Lectura de código por abstracciones sucesivas	x	x	-	-
		Cobertura de sentencias	x	x	x	x
		Particiones en clases de equivalencia	x	x	x	x
	Tratamiento como instrucciones	Revisión de código	x	x	-	-
		Cobertura de decisiones	x	x	x	x
		Particiones en clases de equivalencia	x	x	x	x
	Proceso de aplicación del tratamiento	Procedimiento revisión de código	x	x	-	-
		Procedimiento cobertura de decisiones	x	x	x	x
		Procedimiento particiones en clases de equivalencia	x	x	x	x
Efecto	Variable Respuesta	Efectividad	x	x	x	x
	Elaboración de la variable respuesta	- Revisión: Porcentaje de faltas reportadas	x	x	x	x
		- Estructural y Funcional: Porcentaje de faltas para los cuales se genera un caso de prueba				
	Métrica	- Revisión: $\frac{FaltasReportadas}{FaltasTotal} * 100$	x	x	x	x
		- Estr. y Func.: $\frac{Faltas(CasoPrueba)}{FaltasTotal} * 100$				
Medición	Instrumentos de medida	Ordenador, Folios, Bolígrafos (manual)	x	x	x	x
	Procedimiento de medición	Criterio Relajado	x	x	x	x

Tabla 5.5 Elementos de la Operacionalización

5.2.2 Identificación de las Variables de Diseño

5.2.2.1 Creación de un Diseño Agregado

En la Tabla 5.6, la Tabla 5.7, la Tabla 5.8 y la Tabla 5.9 se muestra el diseño de los experimentos UPM-1, UPM-2, UPV y ORT, respectivamente.

Grupo	Sesión 1			Sesión 2			Sesión 3		
	R	E	F	R	E	F	R	E	F
G1.1 (6)	cmdline	-	-	-	-	ntree	-	nametbl	-
G1.2 (6)	cmdline	-	-	-	ntree	-	-	-	nametbl
G1.3 (6)	-	cmdline	-	ntree	-	-	-	-	nametbl
G1.4 (6)	-	cmdline	-	-	-	ntree	nametbl	-	-
G1.5 (6)	-	-	cmdline	-	ntree	-	nametbl	-	-
G1.6 (6)	-	-	cmdline	ntree	-	-	-	nametbl	-

Tabla 5.6 Diseño del experimento UPM-1

Grupo	Sesión 1			Sesión 2			Sesión 3		
	R	E	F	R	E	F	R	E	F
G2.1 (4)	ntree	-	-	-	-	cmdline	-	nametbl	-
G2.2 (7)	ntree	-	-	-	cmdline	-	-	-	nametbl
G2.3 (5)	-	ntree	-	cmdline	-	-	-	-	nametbl
G2.4 (7)	-	ntree	-	-	-	cmdline	nametbl	-	-
G2.5 (4)	-	-	ntree	-	cmdline	-	nametbl	-	-
G2.6 (5)	-	-	ntree	cmdline	-	-	-	nametbl	-

Tabla 5.7 Diseño del experimento UPM-2

Grupo	Sesión 1	Sesión 2
	E	F
G3.1 (6)	cmdline	ntree
G3.2 (5)	nametbl	cmdline
G3.3 (5)	ntree	nametbl
G3.4 (6)	cmdline	nametbl
G3.5 (5)	nametbl	ntree
G3.6 (6)	ntree	cmdline

Tabla 5.8 Diseño del experimento UPV

Grupo	Sesión 1			
	Período 1		Período 2	
	E	F	E	F
G4.1 (19)	cmdline	-	-	nametbl
G4.2 (20)	-	nametbl	cmdline	-
G4.3 (22)	nametbl	-	-	cmdline
G4.4 (20)	-	cmdline	nametbl	-

Tabla 5.9 Diseño del experimento ORT

La superposición se ha realizado de la siguiente manera:

- Los experimentos UPM-1 y UPM-2 tienen el mismo diseño (crossover). La única diferencia entre estos experimentos es el orden en que se utilizan los programas en

cada una de las sesiones. En la Tabla 5.10, se puede ver el resultado de la superposición de estos dos experimentos.

Grupo	Grupo Original	Sesión 1			Sesión 2			Sesión 3		
		R	E	F	R	E	F	R	E	F
G1	G1.1 (6)	Cm	-	-	-	-	Nt	-	Na	-
G2	G1.2 (6)	Cm	-	-	-	Nt	-	-	-	Na
G3	G1.3 (6)	-	Cm	-	Nt	-	-	-	-	Na
G4	G1.4 (6)	-	Cm	-	-	-	Nt	Na	-	-
G5	G1.5 (6)	-	-	Cm	-	Nt	-	Na	-	-
G6	G1.6 (6)	-	-	Cm	Nt	-	-	-	Na	-
G7	G2.1 (4)	Nt	-	-	-	-	Cm	-	Na	-
G8	G2.2 (7)	Nt	-	-	-	Cm	-	-	-	Na
G9	G2.3 (5)	-	Nt	-	Cm	-	-	-	-	Na
G10	G2.4 (7)	-	Nt	-	-	-	Cm	Na	-	-
G11	G2.5 (4)	-	-	Nt	-	Cm	-	Na	-	-
G12	G2.6 (5)	-	-	Nt	Cm	-	-	-	Na	-

Tabla 5.10 Superposición de los diseños UPM04-UPM05

- UPV es un experimento con un diseño de medidas repetidas, es decir, las técnicas se aplican siempre en el mismo orden. Este experimento se solapa parcialmente con el diseño agregado del paso anterior, por tanto, se crean nuevos grupos (G13 a G16) y se añaden participantes a los ya existentes grupos G4 y G10. En la Tabla 5.11, se puede ver el resultado de la superposición de los tres experimentos.

Grupo	Grupo Original	Sesión 1			Sesión 2			Sesión 3		
		R	E	F	R	E	F	R	E	F
G1	G1.1 (6)	Cm	-	-	-	-	Nt	-	Na	-
G2	G1.2 (6)	Cm	-	-	-	Nt	-	-	-	Na
G3	G1.3 (6)	-	Cm	-	Nt	-	-	-	-	Na
G4	G1.4 (6), G3.1 (6)	-	Cm	-	-	-	Nt	Na	-	-
G5	G1.5 (6)	-	-	Cm	-	Nt	-	Na	-	-
G6	G1.6 (6)	-	-	Cm	Nt	-	-	-	Na	-
G7	G2.1 (4)	Nt	-	-	-	-	Cm	-	Na	-
G8	G2.2 (7)	Nt	-	-	-	Cm	-	-	-	Na
G9	G2.3 (5)	-	Nt	-	Cm	-	-	-	-	Na
G10	G2.4 (7), G3.6 (6)	-	Nt	-	-	-	Cm	Na	-	-
G11	G2.5 (4)	-	-	Nt	-	Cm	-	Na	-	-
G12	G2.6 (5)	-	-	Nt	Cm	-	-	-	Na	-
G13	G3.2 (5)	-	Na	-	-	-	Cm	-	-	-
G14	G3.3 (5)	-	Nt	-	-	-	Na	-	-	-
G15	G3.4 (6)	-	Cm	-	-	-	Na	-	-	-
G16	G3.5 (5)	-	Na	-	-	-	Nt	-	-	-

Tabla 5.11 Superposición de los diseños UPM04-UPM05-UPV05

- Finalmente, en la Tabla 5.12, se puede observar que al superponer ORT05, su diseño se complementa con el diseño agregado del paso anterior. Se crea una nueva variable “Período”, distinta a sesión ya que en ORT los participantes aplican las técnicas en la misma sesión, una a continuación de la otra, por lo tanto, se crean nuevos grupos (G17 a G20). Es importante tener en cuenta que, en los otros tres experimentos, la sesión es lo mismo que el período.

Grupo	Grupo Original	Sesión 1						Sesión 2			Sesión 3		
		Período 1			Período 2			Período 2			Período 3		
		R	E	F	R	E	F	R	E	F	R	E	F
G1	G1.1 (6)	Cm	-	-	-	-	-	-	-	Nt	-	Na	-
G2	G1.2 (6)	Cm	-	-	-	-	-	-	Nt	-	-	-	Na
G3	G1.3 (6)	-	Cm	-	-	-	-	Nt	-	-	-	-	Na
G4	G1.4 (6), G3.1 (6)	-	Cm	-	-	-	-	-	-	Nt	Na	-	-
G5	G1.5 (6)	-	-	Cm	-	-	-	-	Nt	-	Na	-	-
G6	G1.6 (6)	-	-	Cm	-	-	-	Nt	-	-	-	Na	-
G7	G2.1 (4)	Nt	-	-	-	-	-	-	-	Cm	-	Na	-
G8	G2.2 (7)	Nt	-	-	-	-	-	-	Cm	-	-	-	Na
G9	G2.3 (5)	-	Nt	-	-	-	-	Cm	-	-	-	-	Na
G10	G2.4 (7), G3.6 (6)	-	Nt	-	-	-	-	-	-	Cm	Na	-	-
G11	G2.5 (4)	-	-	Nt	-	-	-	-	Cm	-	Na	-	-
G12	G2.6 (5)	-	-	Nt	-	-	-	Cm	-	-	-	Na	-
G13	G3.2 (5)	-	Na	-	-	-	-	-	-	Cm	-	-	-
G14	G3.3 (5)	-	Nt	-	-	-	-	-	-	Na	-	-	-
G15	G3.4 (6)	-	Cm	-	-	-	-	-	-	Na	-	-	-
G16	G3.5 (5)	-	Na	-	-	-	-	-	-	Nt	-	-	-
G17	G4.1 (19)	-	Cm	-	-	-	Na	-	-	-	-	-	-
G18	G4.2 (20)	-	-	Na	-	Cm	-	-	-	-	-	-	-
G19	G4.3 (22)	-	Na	-	-	-	Cm	-	-	-	-	-	-
G20	G4.4 (20)	-	-	Cm	-	Na	-	-	-	-	-	-	-

Tabla 5.12 Superposición UPM04-UPM05-UPV05-ORT05

5.2.2.2 Exploración de los Resultados de los Análisis Individuales

La Tabla 5.13 muestra los modelos estadísticos utilizados en el análisis individual de cada uno de los experimentos. Se han estudiado los factores principales y la interacción entre ellos. Debido a que no todos los diseños son iguales, se tienen que analizar de manera distinta:

Variables	UPM-1	UPM-2	UPV	ORT
Técnica	x	x	x ¹	x
Programa	x ²	x ²	x	x
Grupo	x	x	-	x
Período	-	-	-	x
Técnica * Programa	x ²	x ²	x ¹	x
Técnica * Período	-	-	-	x

¹ La Técnica está confundida con la Sesión y el Período

² El Programa está confundido con la Sesión y el Período

Tabla 5.13 Variables del análisis

- En UPM-1 y UPM-2 se analiza la Técnica, el Programa, el Grupo y la interacción entre la Técnica y el Programa, ya que es un diseño crossover. Para analizarlo se han seguido las guías descritas en (Vegas, et al., 2016).
- En UPV05 se analiza la Técnica, el Programa, y la interacción entre la Técnica y el Programa.
- En ORT05 se analiza la Técnica, el Programa, el Grupo, el Período, la interacción entre la Técnica y el Programa, la interacción entre la Técnica y el Período y la interacción entre la Técnica, el Programa y el Período.

Nótese que algunas variables están confundidas con otras. En UPM-1 y UPM-2 el Programa está confundido con la Sesión y el Período, mientras que en UPV es la Técnica la que está confundida con la Sesión y el Período. La variable Grupo representa la secuencia de aplicación de las técnicas en UPM-1 y UPM-2, mientras que en ORT este representa la secuencia de aplicación de los programas. Esto deberá tenerse en cuenta a la hora de interpretar los resultados.

Los resultados completos de los análisis individuales de cada experimento se muestran en el Apéndice D.1. La tabla 5.14 muestra los casos en los que se ha obtenido significación estadística (p -valor $< 0,05$) y/o significación práctica ($r > 0,23$).

Variable	UPM-1	UPM-2	UPV	ORT
Técnica	Estadística Práctica	Estadística Práctica	-	Estadística
Programa	-	Estadística Práctica	Práctica	Estadística Práctica
Grupo	Práctica	-	N/A	-
Período	N/A	N/A	N/A	Estadística
Técnica * Programa	Estadística Práctica	Práctica	Estadística Práctica	Práctica
Técnica * Período	N/A	N/A	N/A	Práctica

Tabla 5.14 Resultados de los análisis individuales

Los casos en los que no existe ni significación estadística ni práctica aparecen representados con un guión, mientras que las variables que no se han considerado en ese análisis, aparecen representados con un N/A. Los análisis de todos los experimentos de esta familia se realizaron mediante el procedimiento MIXED y fueron hechos con SPSS v25 (IBM SPSS Statistics for Macintosh, Version 25, 2017). Por otro lado, el tamaño del efecto es r , y está calculado como indica Field en el libro (Field, 2013).

5.2.2.3 Identificación del Conjunto de Variables de Diseño

Considerando el diseño agregado mostrado en la Tabla 5.12 las variables que se deben incluir en el análisis que vienen impuestas por el diseño son:

- Técnica, es el factor principal de la familia.
- Programa.
- Período.
- La interacción Técnica*Programa.

Por otro lado, de los análisis individuales se ha decidido:

- No incluir la variable Grupo, dado que esta sale significativa una única vez y no será relevante para el modelo.
- No incluir la variable Sesión, ya que en tres de cuatro experimentos está confundida con el período, que ya se tiene en cuenta.

5.2.3 Identificación de las Variables Contextuales

En este paso se identifican aquellos elementos cuyos valores podrían provocar diferencias en los resultados experimentales.

5.2.3.1 Identificación de Diferencias Contextuales

Los elementos de la ejecución experimental que se deben tener en cuenta a la hora de la identificación de diferencias contextuales aparecen en la Tabla 5.15 junto con los valores que toman para cada experimento de la familia. En gris se muestran aquellos que se considera podrían provocar diferencias en los resultados.

EJECUCIÓN EXPERIMENTAL		UPM-1	UPM-2	UPV	ORT
Diseño	Duración de la tarea Limitado	-	-	X	-
	Ilimitado	X	X	-	X
Entrenamiento	Combinación entre entrenamiento y sesión Secuencial	X	X	-	X
	Intercalado	-	-	X	-
	Aspectos de transmisión de los tratamientos Experiencia en enseñanza de técnicas de prueba alta	X	X	X	X
	Duración del entrenamiento 3 sesiones de 4 horas	X	X	-	X
	2 sesiones de 2 horas	-	-	X	-

Tabla 5.15 Elementos del diseño

- La *duración de la tarea* podría tener implicaciones en los resultados, porque los participantes que trabajan con tiempo limitado podrían tener peores resultados, ya que estarían trabajando bajo presión para conseguir terminar la tarea a tiempo, y por tanto podrían aplicar peor las técnicas.
- La *combinación entre el entrenamiento y las sesiones* podría tener implicaciones en los resultados, porque, los participantes que lo recibieron de manera secuencial podrían no haber tenido suficiente tiempo para estudiar el último tratamiento enseñado o viceversa, podrían haber olvidado el primero, provocando un conocimiento desigual de los mismos.
- Los *Aspectos de transmisión de los tratamientos*, no podría tener implicaciones en los resultados en esta familia ya que siempre toma el mismo valor.
- La *duración del entrenamiento* podría tener implicación en los resultados porque, aunque se intenta adaptarlo a las necesidades de los participantes, no se puede garantizar cuales son exactamente los conocimientos que posee cada participante. Ellos pueden tener conocimientos desiguales y, por tanto, aquellos que recibieron un entrenamiento más corto podrían aplicar peor las técnicas.

Los elementos de la instrumentación a tener en cuenta a la hora de identificación de diferencias contextuales aparecen en la Tabla 5.16 junto con los valores que toman para cada experimento de la familia. *Ninguno de estos elementos provocará diferencias en los resultados.*

- Las *guías suplementarias* no deberían tener implicaciones en los resultados, ya que son una combinación entre la especificación del programa y un breve resumen sobre cómo

aplicar las técnicas. Por lo tanto, estas dependen de la técnica y los objetos experimentales, que son iguales en todos los experimentos.

INSTRUMENTACIÓN		UPM-1	UPM-2	UPV	ORT
Guías suplementarias	Guías	x	x	x	x
Formularios	Formularios	x	x	x	x
Transmisión de los tratamientos	Instrucciones	x	x	x	x
Infraestructura tecnológica	Ordenador	x	x	x	-

Tabla 5.16 Elementos de la instrumentación

- Los *formularios* no deberían tener implicaciones en los resultados porque se han usado los mismos en todos los experimentos. Los formularios se necesitan para recoger datos sobre los participantes y obtener los resultados después de aplicar las técnicas.
- La *transmisión de los tratamientos* son un conjunto completo de instrucciones sobre cómo aplicar las técnicas. No debería tener implicaciones en los resultados porque éstas son las mismas en todos los casos.
- La *infraestructura tecnológica* no debería tener implicaciones en los resultados porque es la misma en todos los casos.

Los elementos de la población a tener en cuenta a la hora de identificación de diferencias contextuales aparecen en la Tabla 5.17 junto con los valores que toman para cada experimento. En gris se muestran aquellos que se considera podrían provocar diferencias en los resultados.

POBLACIÓN				UPM-1	UPM-2	UPV	ORT
Prop. de los Sujetos	Características comunes	Tipo de sujeto	Estudiantes de grado	x	x	x	x
		Experiencia con la técnica	Familiarizado	-	-	x	-
			No familiarizado	x	x	-	x
		Conocimientos de programación	Alto	x	x	x	-
			Bajo	-	-	-	x
		Motivación	Alto	x	x	-	x
Bajo	-		-	x	-		
Prop. de los Objetos	Complejidad	Media-baja 15-29	x	x	x	x	
		Alta > 30	x	x	x	x	
	Dominio	Tareas juguete	x	x	x	x	
	Lenguaje	C	x	x	x	x	

Tabla 5.17 Elementos de la población

- El *tipo de sujetos* durante el experimento no debería tener implicaciones en los resultados, ya que todos los experimentos se realizaron con estudiantes de grado.
- La *experiencia con la técnica* podría tener implicaciones en los resultados, porque los participantes que están familiarizados con las técnicas podrían aplicarlas mejor que aquellos que no lo están. A pesar de que el entrenamiento debería compensarlo, se tendrá en cuenta.
- El *conocimiento en programación* podría tener implicaciones en los resultados porque los participantes que tienen menos conocimientos podrían tener más dificultades a la hora de entender el código fuente, en el caso de la técnica revisión de código o estructural. Por lo tanto, estas podrían ser menos efectivas.

- La *motivación* que tengan los participantes al realizar el experimento podría tener implicaciones en los resultados. En el caso de UPM y ORT, la realización del experimento era la calificación de la asignatura, mientras que en UPV apenas tenía impacto en la calificación. Por lo tanto, los participantes de UPV podrían estar menos motivados y las técnicas podrían ser menos efectivas
- La *complejidad* de los objetos no debería tener implicaciones en los resultados, porque dependen de los programas que fueron los mismos en los 4 experimentos.
- El *dominio* de los objetos no debería tener implicaciones en los resultados ya que está confundido con los objetos experimentales (adaptados a los experimentos) y con las características de los participantes (si ellos resuelven el mismo tipo de programas) y en todos los casos toman los mismos valores.
- El *lenguaje de programación* no debería tener implicaciones en los resultados, ya que, en los 4 experimentos tiene el mismo valor.

Los elementos de los experimentadores a tener en cuenta a la hora de identificación de diferencias contextuales aparecen en la Tabla 5.18 junto con los valores que toman para cada experimento de la familia. En gris se muestran aquellos que se considera podrían provocar diferencias en los resultados.

EXPERIMENTADORES		UPM-1	UPM-2	UPV	ORT
Diseñador	Vegas	x	x	x	-
	Juristo	x	x	x	-
	Solari	-	-	-	x
Entrenador	Juristo	x	x	-	-
	Vegas	x	-	-	-
	Moreno	-	x	-	-
	Letelier	-	-	x	-
	Solari	-	-	-	x
Monitor	Vegas	x	x	-	-
	Moreno	x	x	-	-
	Letelier	-	-	x	-
	Solari	-	-	-	x
Medidor	Vegas	x	-	-	-
	Solari	-	x	x	x

Tabla 5.18 Elementos de los experimentadores

- El *diseñador* no debería tener implicaciones en los resultados, ya que, por lo general, suele ser la misma persona o hay al menos un diseñador en común en todos los experimentos y si no, éste se mantiene en contacto con los nuevos diseñadores. Además, el posible impacto ya estaría contemplado en el diseño en sí mismo.
- El *entrenador* podría tener implicaciones en los resultados, porque al no tener la misma experiencia podría influir en la enseñanza de los tratamientos.
- El *monitor* no debería tener implicaciones en los resultados, ya que, siguieron el mismo procedimiento durante el experimento.
- El *medidor* no debería tener implicaciones en los resultados, ya que, por lo general, suele ser la misma persona, si no, éste transmite la información a los nuevos medidores.

Concretamente, en este caso Vegas transmite el procedimiento a seguir a Solari y él lo realiza de la misma manera.

5.2.3.2 Creación de un Conjunto Inicial de Variables Contextuales

Una vez se han obtenido los elementos que pueden tener implicaciones en los resultados, es necesario transformarlos a las correspondientes variables que serán las que se incluirán en el paso 5 del proceso. La tabla 5.19 muestra el conjunto inicial de elementos que se podrían considerar como variables junto con sus respectivos valores. Se puede observar que en este momento, hay una correspondencia 1 a 1 entre elemento y variable.

Elemento		Valores	UPM-1	UPM-2	UPV	ORT
Duración de la tarea		Limitada	-	-	X	-
		Ilimitada	X	X	-	X
Duración del entrenamiento		3 sesiones de 4 horas	X	X	-	X
		2 sesiones de 2 horas	-	-	X	-
Combinación entre el entrenamiento con las sesiones		Secuencial	X	X	-	X
		Intercalado	-	-	X	-
Características comunes de los sujetos	Experiencia con la técnica	Familiarizado	-	-	X	-
		No familiarizado	X	X	-	X
	Conocimientos de programación	Alto	X	X	X	-
		Bajo	-	-	-	X
	Motivación	Alto	X	X	-	X
		Bajo	-	-	X	-
Entrenador		Juristo-Vegas	X	-	-	-
		Juristo-Moreno	-	X	-	-
		Letelier	-	-	X	-
		Solari	-	-	-	X

Tabla 5.19 Listado de variables

5.2.3.3 Refinamiento del Conjunto Inicial de Variables Contextuales

En la Tabla 5.19 se puede observar que hay variables que están confundidas con otras. Concretamente, se puede observar los siguientes grupos de variables:

- Variable a la que se ha denominado DOTSEM, por las iniciales de las variables que agrupa (en inglés: **D**uration, **O**rdering **T**raining and **S**ession, **E**xperience, **M**otivation). Agrupa las siguientes 5 variables:
 - Duración de la tarea.
 - Duración del entrenamiento.
 - Combinación entre el entrenamiento y las sesiones.
 - Experiencia con la técnica.
 - Motivación.
- Entrenador. Esta variable está confundida con la variable experimento y no se podrá incluir en el análisis. Nótese que en caso de que el experimento sea significativo no se podrá saber si es el experimento o es el entrenador lo que influye en los resultados.

Finalmente, la Tabla 5.20 muestra el conjunto final de variables contextuales junto con su distribución, en ella se puede observar que solo se tendrán en cuenta dos variables: Conocimientos de Programación y DOTSEM, además se puede observar el número de participantes y observaciones que hay para cada variable y también para cada experimento.

Variable	Valores	Experimento	Sujetos	Observaciones
Conocimientos de programación	Alto	UPM-1	36	108
		UPM-2	32	96
		UPV	33	62
		Total	101	266
	Bajo	ORT	78	152
DOTSEM	DOTSEM 1	UPM-1	36	108
		UPM-2	32	96
		ORT	78	152
		Total	146	356
	DOTSEM 2	UPV	33	62

Tabla 5.20 Distribución del conjunto final de variables

En la Tabla 5.20 se puede observar que el grupo de variables DOTSEM está bastante desbalanceado, ya que, para el primer valor hay un total de 356 observaciones, mientras que para el segundo valor solo hay 62 observaciones. Esto podría afectar al análisis, teniendo avisos de error con la matriz de convergencia utilizada.

5.2.4 Identificación de las Variables de Participantes

En este paso, se usa la información recogida de los participantes sobre su experiencia o conocimientos.

5.2.4.1 Identificación de las Diferencias entre los Participantes

La Tabla 5.21 muestra los elementos que definen las diferencias entre los participantes. Debido a que estas variables son diferentes para cada participante solo se muestran el tipo de valores que pueden tomar. A continuación, se detalla cada una de ellas:

- **Experiencia Relativa.** Recoge información sobre la experiencia relativa con el lenguaje de programación C de los participantes. Aunque, esta variable tenía niveles del 0 al 5, algunos participantes contestaron valores intermedios a los mismos. Los niveles eran:
 - 0. Ninguna
 - 1. Conozco la teoría
 - 2. Pequeños ejercicios
 - 3. Usado en prácticas
 - 4. Usado en un desarrollo
 - 5. Experto
- **Experiencia Absoluta.** Recoge información sobre la experiencia absoluta con el lenguaje de programación C de los participantes. Es una variable numérica que indica el número

de años que los participantes han estado en contacto con la programación. Sirve como corroboración de la experiencia absoluta.

- Asistencia. Indica si los participantes asistieron a las clases de la asignatura.

Elemento	Escala	UPM-1	UPM-2	UPV	ORT
Experiencia Relativa	Escala de Likert (con 6 respuestas)	x	x	x	x
Experiencia Absoluta	Años	x	x	x	x
Asistencia	Binario	-	x	-	-

Tabla 5.21 Distribución de las variables de los participantes

5.2.4.2 Creación de un Conjunto Inicial de las Variables de Participantes

Del conjunto de variables obtenido, no se tendrá en cuenta la variable asistencia, debido a que esta solo se midió en un experimento y por lo tanto no tiene sentido para la agregación.

Una vez se han obtenido las variables de los participantes, es necesario transformarlas a variables para que se puedan incluir en el paso 5 del proceso. La tabla 5.22 muestra el conjunto final de variables con sus valores y la escala usada para cada variable.

Elemento	Valor	Escala
Experiencia Absoluta	Escala de Likert (con 6 respuestas)	Ordinal
Experiencia Relativa	Años	Ratio

Tabla 5.22 Conjunto final de las variables de participantes

5.2.4.3 Refinamiento del Conjunto Inicial de las Variables de Participantes

La última actividad de este paso es conocer cómo se distribuyen las variables del conjunto determinado en el paso anterior.

La Figura 5.1 muestra la experiencia relativa con el lenguaje de programación C de los participantes. Se puede observar que los participantes de ORT son los que menor experiencia tienen, con valores comprendidos entre 0 y 3,5. Por otro lado, se puede observar que los participantes con mayor experiencia son los de UPM-2, con valores comprendidos entre 2 y 5. Además, se puede observar que muy pocos participantes de UPM-1 y UPV superan el valor 3,5. En términos generales, los participantes de UPM-1 y UPM-2 tienen mayor experiencia en C que los participantes de UPV y a su vez estos últimos tienen más experiencia que los participantes de ORT. Finalmente, en todos los experimentos la mayoría de los participantes tienen una experiencia relativa de 3.

La figura 5.2 muestra la experiencia absoluta con el lenguaje de programación C de los participantes. Al igual que en la experiencia relativa, se puede observar que los participantes de ORT son los que menor experiencia tienen con valores comprendidos entre 0 y 4 años. Por otro lado, se puede observar que los participantes con mayor experiencia son los de UPM-2, con valores comprendidos entre 1 y 7. Además, se puede observar que en UPV hay participantes con 0 años de experiencia. En términos generales, los participantes de UPM-1, UPM-2 y UPV tienen más experiencia absoluta en C que los participantes de ORT.

Figura 5.1 Experiencia relativa de los participantes

Figura 5.2 Experiencia absoluta de los participantes

Asimismo, teniendo en cuenta la distribución de las variables se puede observar que en la variable experiencia relativa hay desbalanceo y esto podría dar problemas con la técnica de análisis.

Finalmente, ninguna de estas dos variables está confundida con ninguna de las variables contextuales identificadas en el paso anterior del proceso. Por ejemplo, aunque en términos generales, los participantes de UPM-1, UPM-2 y UPV tienen más experiencia absoluta en C que los participantes de ORT, esta variable no se confunde con la variable contextual Conocimientos de Programación, ya que los participantes de estos experimentos tienen un amplio rango de experiencia. Por lo tanto, se incluirán ambas en el modelo de análisis.

5.2.5 Realización del Análisis

La ejecución de los análisis se ha realizado con IBM SPSS V25. El procedimiento MIXED es muy flexible y permite analizar una amplia variedad de diseños multinivel, multinivel cruzado y multinivel de medidas repetidas con variables continuas y categóricas.

5.2.5.1 Identificación del Modelo de Análisis

Hay 4 experimentos intra-sujetos, y 2 de ellos se realizaron en años académicos sucesivos en la misma universidad. Una posible estructura jerárquica podría ser de 4 niveles. El nivel más bajo corresponde a las medidas repetidas tomadas de cada participante, el segundo nivel a los participantes, el tercer nivel a los experimentos y el cuarto nivel a los sitios. Sin embargo, esta estructura no está balanceada, ya que los sitios UPV y ORT tienen un experimento cada uno. Por lo tanto, se ha optado por una jerarquía de 3 niveles como la que se muestra en la Figura 5.3. El nivel 1 corresponde a medidas repetidas, el nivel 2 a los participantes y el nivel 3 a los experimentos.

Figura 5.3 Jerarquía de los experimentos

La definición de las variables que se analizarán se muestra en la Tabla 5.23.

Tipo	Variable	Nivel	Valores	Escala
Diseño	Efectividad	1	0-100	Intervalo
	Técnica	1	Revisión Estructural Funcional	Nominal
	Programa	1	cmdline nametbl ntree	Nominal
	Período	1	1-2	Ordinal
Participante	Experiencia Relativa	2	0-5	Ordinal
	Experiencia Absoluta	2	0-n	Escala
Contextual	Conocimientos de programación	3	Alto Bajo	Nominal
	DOTSEM	3	DOTSEM1 DOTSEM2	Nominal

Tabla 5.23 Definición de las variables de análisis

Las preguntas de investigación a responder con el análisis son:

- RQ_1 : ¿Depende la efectividad de la técnica del programa, el período y/o la interacción de técnica*programa?
- RQ_2 : ¿Está relacionada la efectividad con la experiencia relativa y absoluta con el lenguaje de programación C de los participantes?
- RQ_3 : ¿Está relacionada la efectividad con las diferencias entre los experimentos representadas por las variables Conocimientos de programación y DOTSEM?
- RQ_4 : ¿Se comporta de forma distinta la técnica en los diferentes experimentos?

- RQ_5 : ¿Interactúa la experiencia relativa o absoluta de los participantes en el lenguaje de programación C con las técnicas?
- RQ_6 : ¿Interactúan las características de los experimentos, representadas por las variables DOTSEM y conocimientos de programación con las técnicas?

5.2.5.2 Análisis del Modelo

La Tabla 5.24 muestra los resultados del AIC para el modelo nulo, incluyendo solo las variables de diseño (sin distinguir entre experimentos): período, técnica, programa y la interacción técnica*programa. Como con la variable efectividad original no se obtiene normalidad, se ha tenido que transformar, utilizando la fórmula $\sqrt{\max(Efectividad) - Efectividad_{sujeto}}$, es decir, a cada una de las medidas de los participantes se les restó 100, que es el porcentaje máximo existente en el conjunto de datos, y a este valor se le aplicó la raíz cuadrada.

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Técnica	Identidad	1989,951	Si	
	Diagonal	1973,244	Si	
	AR(1)	1991,707	Si	
	CS	1991,951	Si	No converge
	No estructurada	1978,534	Si	No converge
Período	Identidad	1989,951	Si	
	Diagonal	1985,822	Si	
	AR(1)	1990,429	Si	
	CS	1991,951	Si	No converge
	No estructurada	1989,794	Si	No converge
Programa	Identidad	1989,951	Si	
	Diagonal	1990,516	Si	
	AR(1)	1991,440	Si	
	CS	1991,951	Si	No converge
	No estructurada	1997,444	Si	No converge

Tabla 5.24 AICs para el modelo básico familia de técnicas de prueba

En todos los casos, se prueban las 5 matrices de covarianza: identidad, diagonal, auto-regresiva de primer orden, simetría compuesta y no estructurada. Dado que hay 3 variables intra-participantes (técnica, período y programa), se prueba como efectos repetidos cada una de ellas. El modelo con mejor ajuste (AIC más bajo) aparece en gris claro y corresponde a los efectos repetidos de la técnica y la estructura de covarianza diagonal.

Este modelo se refina aún más como se ha explicado en la Sección 4.6.2. Los resultados detallados del cálculo del AIC se muestran en el Apéndice E.1. La Tabla 5.25 muestra un resumen de los resultados de los análisis. Muestra el AIC para el mejor modelo obtenido cada vez que se agregan nuevas variables al análisis.

En la Tabla 5.25 se puede observar que el AIC para el modelo básico cuando se incorpora la estructura jerárquica de los datos es 1973,825 que es ligeramente más alto que el AIC anterior (1973,244). Esto podría indicar que no es necesario reconocer el hecho de que existan 4 conjuntos de datos, como si cada uno proveniente de un experimento diferente, y los datos podrían analizarse

como si provinieran de un solo experimento. Como el AIC varía en décimas (0,581), lo que indica que es un caso límite, se mantendrá la estructura jerárquica de los datos.

Modelo	Efecto Repetido	Matriz de covarianza	AIC mejor
Nulo (sin jerarquía)	Técnica	Diagonal	1973,244
Añadiendo jerarquía	Técnica	Diagonal	1973,825
Añadiendo variables de participantes y experimento	Período	Diagonal	1413,516
Especificando técnica como efecto aleatorio	Período	Diagonal	1414,044
Técnica*Experiencia Relativa	Período	Diagonal	1412,450
Técnica*Experiencia Relativa, Técnica*Experiencia Absoluta	Período	Diagonal	1414,316
Técnica*Experiencia Relativa, Técnica*Conocimientos de Programación	Período	Diagonal	1413,619

Tabla 5.25 Resultados del análisis para el modelo refinado

Cuando se incorporan las variables de participantes y experimento, el análisis muestra un error y no se pueden obtener resultados confiables. Como una de las variables del experimento (DOTSEM) estaba muy desequilibrada y se preveían posibles problemas de análisis, se elimina del modelo y se vuelve a ejecutar el análisis. Después de hacer esto, el AIC disminuye a 1413,516. Sin embargo, cuando la técnica se especifica como factor aleatorio (para expresar la posibilidad de que las técnicas se comporten de manera diferente para cada experimento), el AIC aumenta ligeramente nuevamente a 1414,044. Esto significa que es muy probable que las técnicas se comporten igualmente en todos los experimentos, ya que el AIC varía en décimas (0,528), lo que indica que es un caso límite.

Cuando se agrega la interacción tratamiento*experiencia relativa, el AIC disminuye nuevamente, por lo tanto, esta interacción se mantiene en el modelo. Seguidamente a este último modelo se le incorpora la interacción tratamiento*experiencia absoluta, sin embargo, el AIC aumenta a 1414,316 y, por lo tanto, esta interacción se eliminará del modelo y se vuelve al modelo anterior que contenía una única interacción. Finalmente, a dicho modelo se le añade la interacción tratamiento*conocimientos de programación y el AIC disminuye 1413,619, por lo tanto, esta interacción también se mantiene en el modelo.

El modelo final obtenido para esta familia de experimentos incluye:

- Técnica, Programa, Periodo y la interacción Técnica*Programa como factores fijos.
- El participante como una variable aleatoria.
- El experimento como una variable aleatoria.
- Las variables a los dos niveles existentes son:
 - De participantes: experiencia absoluta y relativa son factores fijos.
 - De experimento: conocimientos de programación.
- La técnica como un factor aleatorio.
- Las interacciones Técnica*Experiencia relativa y Técnica*Conocimientos de programación como factores fijos.

Una vez se ha elegido el mejor modelo, se realiza el análisis de este. Nótese que antes de llegar al modelo elegido, sobre el que finalmente se realiza el análisis se ha calculado el AIC de un total de 120 modelos: 5 matrices de covarianza x 3 variables de medidas repetidas x 8 modelos:

1. Modelo nulo
2. Modelo básico con jerarquía
3. Incorporando variables (error por desbalanceo en la variable DOTSEM)
4. Incorporando variables de análisis
5. Técnica como efecto aleatorio
6. Interacción Técnica*Experiencia Relativa
7. Interacciones: Técnica*Experiencia Relativa y Técnica*Experiencia Absoluta
8. Interacciones: Técnica*Experiencia Relativa y Técnica*Conocimientos de Programación

5.2.5.3 Interpretación de los Resultados

La Tabla 5.26, la Tabla 5.27 y la Figura 5.4 muestran los resultados del análisis para el modelo final, con la que se podrán responder las preguntas de investigación:

Origen	gl de numerador	gl de denominador	F	Sig.
Intersección	1	144,364	45,597	,000
Técnica	2	161,305	4,562	,012
Programa	2	219,060	6,176	,002
Período	2	118,153	2,344	,100
Experiencia Relativa	1	148,516	,950	,331
Experiencia Absoluta	1	94,394	,010	,919
Conocimientos de Programación	1	11,834	,207	,657
Técnica*Programa	4	243,371	1,057	,378
Técnica*Experiencia Relativa	2	167,094	3,301	,039
Técnica*Conocimientos de Programación	1	23,472	,869	,361

Tabla 5.26 Resultados del análisis para el modelo básico

- **RQ₁:** ¿Depende la efectividad de la técnica, el programa, el período y/o la interacción de técnica*programa?
 - La técnica influye en la efectividad. Las técnicas Estructural y Funcional se comportan igual y mejor que la técnica Revisión de código.
 - El programa influye en la efectividad. Aunque cmdline se comporta peor que nametbl y ntree, que se comportan igual el tamaño del efecto es despreciable.
 - El período no influye en la efectividad.
 - La interacción Técnica*Programa no influye en la efectividad
- **RQ₂:** ¿Depende la efectividad de la experiencia relativa y absoluta con el lenguaje de programación C de los participantes?
 - La experiencia relativa y absoluta no influyen en la efectividad.
- **RQ₃:** ¿Depende la efectividad de las diferencias entre los experimentos representadas por las variables Conocimientos de programación y DOTSEM?
 - No, en el caso de DOTSEM ocurren problemas con la matriz de covarianza y es necesario eliminarla del modelo. Por otro lado, los conocimientos de programación no influyen en la efectividad.
- **RQ₄:** ¿Se comporta de forma distinta la técnica en los diferentes experimentos?
 - Están al límite, pero como empeora ligeramente, no.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Técnica	0,012	Revisión	54,545	Revisión<Estructural	0,001	-0,40
		Estructural	79,170	Revisión<Funcional	0,000	-0,50
		Funcional	83,600	Estructural<Funcional	0,562	-0,10
Programa	0,002	cmdline	68,179	cmdline<ntree	0,074	-0,10
		ntree	77,226	cmdline<nametbl	0,001	-0,18
		nametbl	84,223	ntree<nametbl	0,407	-0,12
Técnica* Experiencia Relativa	0,039	R=>2,5	42,857	creciente	N/A	0,79
		R=>3	54,464			
		R=>3,5	65,079			
		R=>4	52,381			
		E=>0	26,191	creciente	N/A	0,49
		E=>1	69,048			
		E=>1,5	58,333			
		E=>2	61,404			
		E=>2,5	72,024			
		E=>3	73,309			
		E=>3,5	70,996			
		E=>4	82,993			
		F=>0	70,635	constante	N/A	0
		F=>1	85,714			
		F=>1,5	91,667			
		F=>2	74,376			
		F=>2,5	80,952			
		F=>3	76,029			
		F=>3,5	92,208			
		F=>4	78,571			

Tabla 5.27 Resultados del análisis

- **RQ₅:** ¿Interactúa la experiencia relativa o absoluta de los participantes en el lenguaje de programación C con las técnicas?
 - La interacción técnica por experiencia relativa está afectando a la efectividad. A mayor experiencia de los participantes las técnicas revisión y estructural consiguen mejor efectividad. Por el contrario, la técnica funcional consigue la misma efectividad independientemente de la experiencia relativa con el lenguaje de programación C, lo cual tiene sentido, ya que en la técnica funcional los participantes no ven el código. Tal y como se puede observar en la Figura 5.4.
- **RQ₆:** ¿Interactúan las características de los experimentos con las técnicas?
 - La interacción técnica por conocimientos de programación no afecta a la efectividad.

De todo esto se puede concluir que:

- La técnica influye en la efectividad $(F=E)>R$.
- Las técnicas se están comportando igual para diferentes experimentos.
- Los diferentes experimentos se están comportando de la misma manera.
- La única variable moderadora que influye en los resultados es la experiencia relativa. A mayor experiencia se consiguen mejor efectividad para revisión y estructural. Funcional consigue la misma efectividad independientemente de la experiencia.

Figura 5.4 Resultados de la interacción Técnica*Experiencia Relativa

5.3 Discusión sobre el proceso

El proceso de agregación resulta especialmente útil para distinguir la influencia de los distintos niveles en la variable respuesta.

Una vez caracterizados los experimentos (Sección 4.2.1), es sencillo, distinguir las diferencias entre los experimentos. Eso sí, como se ha podido comprobar en este capítulo, para la elección de las posibles variables moderadoras, se debe conocer muy bien cada uno de los experimentos puesto que es la única manera de determinar el “posible” impacto de un elemento, ya que pueden ocurrir casos en los que precisamente no realizar un cambio influya en los resultados. Considerar incluir todas las variables, sin distinguir su posible influencia, generaría un modelo demasiado complicado y, por lo tanto, el test estadístico mostraría errores.

Las reglas de diseño permiten acoplar los experimentos que tienen diferentes diseños, eso si gracias al análisis individual de cada uno de los experimentos, se puede nuevamente *simplificar* el número de variables a incluir en el modelo.

Finalmente, el análisis conjunto de los experimentos, es la parte más extensa, dado que puede ocurrir que se tenga que transformar varias veces la variable respuesta para conseguir cumplir las restricciones de la prueba estadística. Además, que dependiendo del número de variables intra-participantes se puede extender mucho más cada uno de los cinco modelos.

Por otro lado, es importantísimo comprobar el balanceo de las variables, ya que cuando ocurren errores en la matriz y no se puede avanzar a un siguiente modelo si existen muchas variables ir descartando una a una alargaría mucho más el análisis.

Capítulo 6

Estudio de la Viabilidad de la Propuesta: Agregación de una Familia de Experimentos sobre Desarrollo Guiado por Pruebas (TDD)

En este capítulo se estudia la viabilidad del proceso propuesto en el capítulo 4 aplicado a una familia de tamaño grande (16 experimentos). Téngase en cuenta que, de acuerdo con la revisión sistemática en la literatura de familias de experimentos mostrada en el capítulo 2, el máximo número de experimentos reportados en una familia ha sido 12. Esto quiere decir que esta familia tiene un tamaño grande.

6.1 Descripción de la Familia de Experimentos

La familia se compone de un total de 16 experimentos, los cuales fueron ejecutados entre el año 2011 y 2017, en Finlandia, Malasia, Estonia, España, Italia, Ecuador, Dinamarca y Serbia. Un amplio número de estos experimentos se realizaron en un entorno industrial (con profesionales).

La Tabla 6.1 sintetiza los detalles de esta familia de experimentos. Concretamente, la mitad de estos experimentos fueron realizados en empresas. Por otro lado, se puede observar que los años 2011 y 2017 se realizó un experimento en cada año; en el año 2012 no se llevó a cabo ningún experimento; desde el 2014 al 2016 se realizaron cuatro experimentos en cada año y finalmente en el 2013, se realizaron tres.

Además, se puede observar que siete (43,75%) experimentos fueron realizados en Finlandia, en España, Italia y Dinamarca se realizaron dos experimentos en cada país, mientras que Malasia, Ecuador, Serbia y Estonia se realizó un único experimento en cada país.

En los experimentos que se realizaron en la academia, participaron estudiantes de diferentes niveles de educación superior: estudiantes de grado, estudiantes de máster y estudiantes de posgrado (máster y doctorado).

Esta familia de experimentos se realizó con el objetivo de estudiar distintos métodos de desarrollo del software: desarrollo guiado por pruebas (de ahora en adelante TDD, del inglés Test-Driven

Development) frente a otros. En concreto se estudia si el método de desarrollo de software influye en la calidad del software desarrollado. De este objetivo se ha derivado la siguiente pregunta de investigación:

Exp	Institución	Entorno	Localidad	Fecha
OY1	Oulun yliopisto	Academia	Oulu, Finlandia	Oct/2011
FSH	FSecure	Industria	Helsinki, Finlandia	Oct/2013
FSK	FSecure	Industria	Kuala Lumpur, Malasia	Nov/2013
FSO	FSecure	Industria	Oulu, Finlandia	Dic/2013
EB	Elektrobit	Industria	Kajaani, Finlandia	Feb/2014
UPM	Universidad Politécnica de Madrid	Academia	Madrid, España	Mar/2014
PT	Playtech	Industria	Tartu, Estonia	Abr/2014
UPV	Universitat Politècnica de València	Academia	Valencia, España	May/2014
ER	Ericsson	Industria	Jorvas, Finlandia	Abr/2015
UB	Università degli Studi della Basilicata	Academia	Potenza, Italia	Abr/2015
USD	University of Southern Denmark	Academia	Odense, Dinamarca	Nov/2015
PAF	PAF	Industria	Helsinki, Finlandia	Mar/2016
ETAPA	ETAPA	Industria	Cuenca, Ecuador	Jun/2016
USD2	University of Southern Denmark	Academia	Odense, Dinamarca	Oct/2016
OY2	Oulun yliopisto	Academia	Oulu, Finlandia	Oct/2016
NS	Universidad de Novi Sad	Academia	Novi Sad, Serbia	May/2017

Tabla 6.1 Detalles de la familia FiDiPro

- RQ₁: ¿Depende la calidad del software desarrollado del método de desarrollo utilizado?

6.1.1 Factores y Tratamientos

Para poder contestar a las preguntas de investigación, en esta familia se han utilizado los siguientes factores.

6.1.1.1 Método de Desarrollo

Este es el factor **principal** utilizado en la familia y por lo tanto está presente en todos los experimentos de ésta. Se refiere al método seguido por los participantes a la hora de desarrollar el código fuente de un programa. Este factor tiene cuatro tratamientos diferentes:

- **TDD** (Beck, 2003). Este es el tratamiento utilizado como grupo experimental. TDD es una practica ágil donde las pruebas unitarias se escriben antes que el código de producción. Por lo tanto, el desarrollador debe centrarse en el comportamiento correcto del programa. Las tareas que se realizan son las siguientes: primero se descompone la especificación del programa en sub-tareas, seguidamente se escribe una prueba funcional asociado a cada sub-tarea, y para finalizar se escribe el código fuente necesario para hacer pasar la prueba funcional.
- Como grupo de control se han utilizado distintos tratamientos. En concreto:
 - **Incremental Test Last (ITL)**. Para este tratamiento, también se descomponen las especificaciones en pequeñas sub-tareas, pero a diferencia de TDD, primero se escribe el código fuente e inmediatamente se escribe la prueba funcional antes de pasar a una nueva funcionalidad. Se trata de un proceso incremental.

- **Your Way (YW).** Los participantes realizan la codificación sin seguir ningún proceso determinado. Es decir, ellos emplean el método con el que habitualmente desarrollan software en su entorno de trabajo.

6.1.1.2 Granularidad

Se refiere a la forma en la que se ha escrito la especificación de los objetos experimentales. Este factor tiene los siguientes dos tratamientos. Aparecen reflejados en las tablas correspondientes al diseño como S (del inglés Slicing) y NS (del inglés No Slicing):

- **Grano fino,** la especificación está dividida en requisitos individuales.
- **Grano grueso,** la especificación no está dividida en requisitos individuales.

Es importante remarcar que, este no ha sido un factor en todos los experimentos.

6.1.1.3 Efecto de Aprendizaje

Los participantes aplican los tratamientos dos veces: el primer tratamiento TDD se aplica en la primera y tercera sesión y el segundo tratamiento ITL se aplica en la segunda y cuarta sesión. El efecto de aprendizaje se refiere a la posible mejora al aplicar por segunda vez los tratamientos. Este factor tiene dos tratamientos:

- **Primera,** se refiere a la primera vez que los participantes aplicaron cada uno de los tratamientos: ITL (primera sesión experimental) y TDD (segunda sesión experimental).
- **Segunda,** se refiere a la segunda vez que los participantes aplicaron cada uno de los tratamientos: ITL (tercera sesión experimental) y TDD (cuarta sesión experimental).

Es importante remarcar que, este no ha sido un factor en todos los experimentos.

6.1.1.4 Realización del Trabajo

Se refiere a la agrupación de los participantes al momento de realizar el experimento. Este tiene los siguientes dos niveles:

- **Individual,** si los participantes realizaban la tarea de forma individual.
- **Pareja,** si los participantes realizaban la tarea en parejas.

Es importante remarcar que, este no ha sido un factor en todos los experimentos.

6.1.1.5 Objetos experimentales

Se refiere a los programas sobre los cuales fueron aplicados los tratamientos. En esta familia de experimentos se utilizaron los siguientes objetos experimentales:

- **Mars Rover API (MR).** En este programa los sujetos controlan los movimientos de un vehículo que se mueve en un planeta (cuadrícula) de tamaño arbitrario (x, y). El rover

- empieza en la posición (0, 0) y de cara al Norte, la dirección que puede tener el rover es: norte, sur, este y oeste. Para poder moverse el rover realiza cuatro movimientos derecha, izquierda, delante y atrás. Los movimientos derecha e izquierda cambian la dirección del rover en la misma celda, sin embargo, los movimientos delante y atrás cambian su posición (una única celda). Además, el grid puede contener obstáculos, los cuales deben ser reportados una única vez y en el orden en el que fueron encontrados.
- **Bowling Score Keeper (BSK).** El objetivo de este programa es desarrollar una aplicación que calcule la puntuación del juego bowling. El juego está compuesto de 10 marcos, en el que cada jugador tiene dos oportunidades para derribar 10 bolos y la puntuación de cada marco es el total de bolos derribados. La puntuación final del juego es la suma total de los 10 marcos. Sin embargo, hay cinco casos excepcionales que se puntúan distinto:
 - **Strike:** es cuando se derriban los 10 bolos en la primera oportunidad, en este caso el valor total del marco es igual a 10 (strike) más la suma de las siguientes dos oportunidades del marco siguiente.
 - **Spare:** es cuando se derriban los 10 bolos en las dos oportunidades, en este caso el valor total del marco es igual a 10 (spare) más la suma de la primera oportunidad del marco siguiente.
 - **Dos strikes seguidos:** en este caso el valor del primer strike es la suma de los dos marcos siguientes
 - **Strike en el último marco:** en este caso el valor total del marco es igual a 10 (strike) más la suma de dos oportunidades que no pertenecen a ningún marco en concreto.
 - **Spare en el último marco:** en este caso el valor total del marco es igual a 10 (spare) más la suma de una oportunidad que no pertenece a ningún marco en concreto.
 - **MusicFone (MF).** Es una aplicación que se ejecuta en un teléfono móvil habilitado de GPS y compatible con el formato MP3. La aplicación recomendará a los usuarios, artistas que le gusten y encontrará los próximos eventos para conciertos de dichos artistas consultando los datos recopilados del sitio web de LastFM. Los sujetos tienen que implementar la lógica necesaria para permitir a MF: recomendar artistas, encontrar conciertos de los artistas, calcular la distancia a los conciertos y calcular un itinerario para los conciertos.
 - **Spreadsheet (SS).** El objetivo es crear una hoja de cálculo básica. Los sujetos tienen que implementar el código de las estructuras de datos y de las operaciones, pero no tienen que desarrollar la interfaz de usuario. La hoja de cálculo implementa tres operaciones: `set()` establece valores a las celdas, `get()` obtiene valores de la celdas y `evaluate()` toma el contenido de la celda y devuelve el resultado (por ejemplo, una celda que contiene "`= 1 + 1`" se evalúa como 2). Las celdas pueden almacenar enteros (positivos o negativos), cadenas (entrecomillado simple 'ejemplo') o fórmulas (empiezan por el signo igual "`=`"). Las operaciones permitidas son: suma (+), resta (-), multiplicación (*), división (/), modulo (%) y concatenación (&).
 - **Sudoku (S).** Es un juego en el que el objetivo es colocar 9 dígitos positivos (1...9) en las celdas de una cuadrícula fija de 9x9. Además, la cuadrícula está dividida en sub-cuadrículas de 3x3 las cuales tienen que rellenarse con 9 dígitos positivos (1...9) sin repetirse. Finalmente, un dígito puede aparecer una única vez en las filas y una única vez las columnas globales de la cuadrícula.

Existen dos especificaciones diferentes para los programas MR y BSK, de esta manera se incluye el factor Granularidad mencionado anteriormente. Para el resto de los programas solo existe la versión en grano grueso (NS).

6.1.2 Variable respuesta y Medición

La variable respuesta utilizada en esta familia de experimentos es la calidad del software desarrollado (**QLTY**, del inglés Quality).

Esta variable se refiere a la cantidad de trabajo correctamente desarrollado por el usuario. Y se mide como el porcentaje del total de aserciones correctamente abordadas dividido entre el total del conjunto de aserciones creadas para un objeto experimental, como se muestra en la siguiente ecuación:

$$QLTY = \frac{\#Assert_i(Pass)}{\#Assert_i(All)} \times 100$$

Para poder calcular las variables respuesta, se ha diseñado un conjunto de casos de prueba para cada objeto experimental. En la tabla 6.2, se puede observar los casos de prueba y aserciones que utiliza cada objeto experimental, nótese que un caso de prueba puede tener una o varias aserciones.

Programa	Casos de Prueba	Aserciones
MR	48	56
BSK	52	89
MP	45	123
SS	43	43
S	11	13

Tabla 6.2. Número de casos de prueba y aserciones

Además de los instrumentos utilizados es importante detallar el proceso utilizado para la medición de la variable respuesta. El proceso de medición tiene dos características importantes: la intervención del medidor para poder obtener los resultados del código fuente y posteriormente la manera en la que se obtienen los valores de la variable respuesta (Uyaguari, 2017).

Debido a que, en algunas ocasiones, el código fuente de los participantes puede tener errores que impiden la correcta ejecución de los casos de prueba, es necesario manipular el código, para poder corregirlos y ejecutarlos. Es importante tener en cuenta que estas correcciones permiten obtener mejores resultados en las variables respuesta, porque de lo contrario el código correctamente implementado no podría medirse y se obtendrían valores muy bajos no atribuibles a la implementación de cada participante. Esto generaría resultados de baja fiabilidad. Los valores para este elemento son:

- **Sin Intervención:** los casos de prueba son ejecutados sobre el código fuente, tal y como este fue entregado, aunque contenga errores, estos no son corregidos. Las únicas modificaciones que se realizan son las siguientes:
 - Carga de librerías necesarias
 - Inicialización de variables
 - Renombrado de carpetas o métodos

- **Sintáctica:** se realizan intervenciones básicas en el código fuente, con el fin de evitar errores al ejecutar los casos de prueba. Las posibles modificaciones son:
 - Renombrado de clases y métodos
 - Modificación del formato de los valores devueltos por los métodos
 - Inicialización de variables
 - Creación de constructores con los parámetros sugeridos
 - Creación de los métodos sugeridos faltantes
 - Eliminación de valores fijos en las variables
- **Semántica:** se realizan intervenciones más profundas, dando por válidas soluciones diferentes a las esperadas, o corrigiendo los métodos para que se ajusten a los casos de prueba. Es importante aclarar, que en ningún caso se modifica la lógica del programa implementada por los participantes.
- **Semántica-Sintáctica:** es una combinación de la medición Sintáctica y la Semántica, esta medición fue realizada con la ayuda de un protocolo de medición. El protocolo unifica los pasos que se deben realizar y las consideraciones para tener en cuenta, para medir el código fuente.

Por otro lado, existen dos maneras de transcribir los resultados obtenidos:

- **Manual:** una vez se han ejecutado los casos de prueba, se copian manualmente los resultados obtenidos en una hoja de cálculo Excel en la cual se calculan las métricas de la variable respuesta.
- **Automática:** con la ayuda de un plugin, desarrollado específicamente en el ámbito del proyecto FiDiPro, se obtienen los resultados de la ejecución de los casos de prueba sobre el código fuente de los participantes y a su vez el plugin automáticamente calcula las métricas de la variable respuesta. El plugin está escrito en el lenguaje de programación Java y únicamente se ha implementado para medir los objetos experimentales MR y BSK.

6.2 Analizando la Familia de Experimentos

6.2.1 Identificación de los Grupos de Agregación

6.2.1.1 Caracterización de los Experimentos de la Familia

Los 16 experimentos de esta familia están caracterizados de acuerdo con la Tabla 4.11. La Tabla 6.3 muestra la caracterización de OY1. La caracterización de los demás experimentos se muestra en el Apéndice C.2.

Dimensión		Elemento	Valor
Identificación		Institución	OY1
		Sitio	Academia
		Publicación	(Fucci & Turhan, 2013)
Operacionalización	Causa	Factor	Enfoque de desarrollo Sujeto
		Tratamientos	Enfoque de desarrollo
			ITL
			TDD
			Realización del trabajo
			Pareja Individual
		Tratamiento acorde a la literatura	Sección 6.1.1
		Tratamiento como instrucciones	
		Proceso de aplicación del tratamiento	
	Efecto	Variable Respuesta	Calidad
		Elaboración de la variable respuesta	Sección 6.1.2
	Medición	Métrica	
		Instrumentos de medida	Ad-hoc
		Procedimiento de medición	Sintáctica (Manual)
Ejecución experimental	Diseño	Diseño Experimental	Entre-sujetos
		Sesiones	1
		Factores por sesión	1
		Tratamientos por factor por sesión	2
		Variables de Bloque	Sin variables de bloque
		Orden de los tratamientos (grupos)	Ninguno (2)
		Objetos experimentales por sesión	1
		Duración de la tarea	Ilimitada
	Entrenamiento	Combinación entrenamiento y sesión	Secuencial
		Aspectos de transmisión del tratamiento	Experiencia alta en la enseñanza del tratamiento
		Duración del entrenamiento	Adaptado
Instrumentación		Granularidad	BSK_S
		Guías suplementarias	BSK-Java
		Formularios	Ninguno
		Transmisión de los tratamientos	TLD
		Objetos experimentales	BSK
Población	Sujetos	Características Comunes	Tipo de sujetos
		Características variables	Pre y Posgrado
			Exp Programación
			Escala de Likert
			Exp con Lenguaje
			Escala de Likert
	Objetos	Complejidad	Exp pruebas unitarias
			Escala de Likert
			Exp con framework de pruebas
Experimentadores		Complejidad	Igual a las tareas que los participantes realizan habitualmente
		Dominio	No real
		Lenguaje	Java
		Diseñador	Juristo, Turhan, Fucci
		Entrenador	Fucci
		Monitor	Fucci
		Medidor	Fucci

Tabla 6.3 Caracterización Universidad de Oulu 1

6.2.1.2 Identificación de los Grupos de Experimentos a Agregar

Para establecer aquellos experimentos que pueden ser agregados, es necesario examinar su operacionalización. La Tabla 6.4 muestra la operacionalización de los experimentos de la familia. En ella se observa que:

OPERACIONALIZACIÓN				FS	OY1	EB	UPM	PT	UPV	ER	UB	USD	PAF	ETAPA	USD2	OY2	NoviSad	
Causa	Factores	Enfoque de desarrollo		x	x	x	x	x	x	x	x	x	x	x	x	x	x	
		Efecto de aprendizaje		-	-	-	x	-	-	-	-	-	-	-	-	-	-	
		Granularidad		-	-	-	-	-	-	x	-	x	-	-	-	x	x	-
		Realización del trabajo		-	x	-	-	-	-	-	-	-	-	-	-	-	-	-
	Tratamientos Tratamiento acorde a la literatura Procedimiento de aplicación del tratamiento	Enfoque de desarrollo	ITL	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
			TDD	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
		YW	-	-	-	-	-	-	-	-	-	-	x	x	-	-	-	-
	Efecto de aprendizaje	Primera	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-
Segunda		-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	
Tratamiento como instrucciones	Granularidad	Gruesa	-	-	-	-	-	-	x	-	x	-	-	-	x	x	-	
		Fina	-	-	-	-	-	-	x	-	x	-	-	-	x	x	-	
	Realización del trabajo	Pareja	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	
		Individual	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	
Efecto	Variable respuesta																	
	Elaboración de la VR	Calidad	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Medición	Métrica/Fórmula																	
	Instrumentos de medida	Ad-hoc	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Procedimiento de medición	Sintáctica (M)	x	x	-	x	x	x	x	-	-	-	-	-	-	-	-	-
		Sintáctica (A)	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
		Sin Intervención (M)	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-
		Sintáctica-Semántica (M)	-	-	-	-	-	-	-	-	-	x	-	-	x	x	x	x
		Semántica (A)	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-
Semántica (M)		-	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	

Tabla 6.4 Elementos de la Operacionalización

- *Los elementos de la dimensión efecto son siempre los mismos.* Todos los experimentos comparten la misma variable respuesta (calidad). Asimismo, tanto la elaboración de la variable respuesta como la métrica son las mismas para todos experimentos.
- *Los elementos de la dimensión medición son equivalentes para todos los experimentos.* La familia tiene un único instrumento de medidas, sin embargo, existen distintos procedimientos de medición. El procedimiento de medición es la combinación entre la intervención del medidor y la forma de transcribir los resultados:
 - El tipo de *intervención* (sin intervención, sintáctica, semántica o sintáctica-semántica) permite conectar los casos de prueba para poder calcular las métricas correctamente, debido a que la existencia de errores no atribuibles a la implementación del participante suele evitar la correcta ejecución de los casos de prueba.
 - La *transcripción* de resultados (manual o automática) es una mejora que evita cometer errores humanos. Aparecen reflejados en la Tabla 6.4 como A y M respectivamente.

Los procedimientos de medición son equivalentes, ya que no producen ninguna alteración en los resultados, sino más bien permiten calcular correctamente el código desarrollado por los participantes.

- *Hay al menos un factor en común en la dimensión de causa.* El factor principal de la familia y común a todos los experimentos es el “Enfoque de Desarrollo”, este tiene 3 tratamientos: ITL y TDD se estudian en todos los experimentos mientras que YW, solamente se estudia en 2 experimentos (PAF y ETAPA). Asimismo, el proceso de aplicación de los tratamientos, sus instrucciones y la referencia bibliográfica seguida es el mismo para cada tratamiento en cada uno de los experimentos en que aparece. Con respecto a los otros factores, no se considerará ni el “Efecto de aprendizaje” (solo se estudia en UPM) ni la “Realización del Trabajo” (solo se estudia en OY1) ya que aparecen estudiados en un único experimento, y por tanto su agregación no tiene sentido. Finalmente, la “Granularidad” se incorpora al programa ya que es factor únicamente en algunos experimentos. En conclusión, el único factor a considerar es el Enfoque de desarrollo y la granularidad se estudiará asociada en el programa.

Por tanto, **se establece un único grupo de agregación formado por todos experimentos de la familia.**

6.2.2 Identificación de las Variables de Diseño

6.2.2.1 Creación de un Diseño Agregado

Las Tablas 6.5 a 6.16 muestran el diseño de cada uno de los experimentos de la familia. Los experimentos que comparten exactamente el mismo diseño se representan en una única tabla.

Exp	Gr	Sesión 1	Sesión 2	Sesión 3
		ITL	TDD	TDD
FSH	G1.1(6)	MR_NS	BSK_S	MF_NS
FSK	G2.1(7)	MR_NS	BSK_S	MF_NS
FSO	G3.1(11)	MR_NS	BSK_S	MF_NS
EB	G4.1(9)	MR_NS	BSK_S	MF_NS

Tabla 6.5 Diseño FSH, FSK, FSO y EB

	Gr	Sesión 1			
		ITL		TDD	
		Pareja	Individual	Pareja	Individual
OY1	G5.1(18)	BSK_S	BSK_S	-	-
	G5.2(23)	-	-	BSK_S	BSK_S

Tabla 6.6 Diseño OY1

	Gr	Sesión 1	Sesión 2	Sesión 3
		ITL	TDD	TDD
PT	G7.1(8)	MR_NS	BSK_S	MF_NS
	G7.2(9)	BSK_S	MR_NS	MF_NS

Tabla 6.7 Diseño PT

	Gr	Sesión 1	Sesión 2
		ITL	TDD
UPV	G8.1(31)	MR_NS	BSK_S

Tabla 6.8 Diseño UPV

	Gr	Sesión 1	Sesión 2	Sesión 3	Sesión 4
		ITL	TDD	ITL	TDD
UPM	G6.1(1)	MR_NS	MF_NS	BSK_S	S_NS
	G6.2(1)	BSK_S	MR_NS	MF_NS	S_NS
	G6.3(1)	BSK_S	MF_NS	S_NS	MR_NS
	G6.4(1)	MF_NS	S_NS	BSK_S	MR_NS
	G6.5(1)	S_NS	MF_NS	MR_NS	BSK_S
	G6.6(1)	S_NS	BSK_S	MF_NS	MR_NS
	G6.7(1)	MR_NS	BSK_S	MF_NS	S_NS
	G6.8(1)	BSK_S	S_NS	MR_NS	MF_NS
	G6.9(1)	BSK_S	MR_NS	S_NS	MF_NS
	G6.10(1)	MF_NS	BSK_S	MR_NS	S_NS
	G6.11(1)	MR_NS	S_NS	MF_NS	BSK_S
	G6.12(1)	MF_NS	MR_NS	S_NS	BSK_S
	G6.13(1)	S_NS	MR_NS	BSK_S	MF_NS
	G6.14(1)	MF_NS	S_NS	MR_NS	BSK_S
	G6.15(1)	BSK_S	S_NS	MF_NS	MR_NS
	G6.16(1)	MR_NS	MF_NS	S_NS	BSK_S

Tabla 6.9 Diseño UPM

	Gr	Sesión 1	Sesión 2
		ITL	TDD
ER	G9.1(3)	MR_NS	BSK_NS
	G9.2(2)	MR_NS	BSK_S
	G9.3(3)	MR_S	BSK_NS
	G9.4(2)	MR_S	BSK_S
	G9.5(3)	BSK_NS	MR_NS
	G9.6(3)	BSK_NS	MR_S
	G9.7(2)	BSK_S	MR_NS
	G9.8(2)	BSK_S	MR_S

Tabla 6.10 Diseño ER

	Gr	Sesión 1		Sesión 2	
		ITL	TDD	ITL	TDD
UB	G10.1(11)	-	MR_S	BSK_S	-
	G10.2(9)	MR_S	-	-	BSK_S

Tabla 6.11 Diseño UB

	Gr	Sesión 1	
		ITL	TDD
USD2	G12.1(15)	-	BSK_NS
	G12.2(13)	-	BSK_S
	G12.3(17)	BSK_NS	-
	G12.4(19)	BSK_S	-

Tabla 6.12 Diseño USD2

	Gr	Sesión 1	Sesión 2	Sesión 3
		YW	ITL	TDD
PAF	G14.1(4)	BSK_NS	SS_NS	MR_NS
	G14.2(4)	SS_NS	MR_NS	BSK_NS
	G14.3(1)	MR_NS	BSK_NS	SS_NS
	G14.4(0)	SS_NS	BSK_NS	MR_NS
ETAPA	G15.1(4)	BSK_NS	SS_NS	MR_NS
	G15.2(3)	SS_NS	MR_NS	BSK_NS
	G15.3(4)	MR_NS	BSK_NS	SS_NS
	G15.4(1)	SS_NS	BSK_NS	MR_NS

Tabla 6.13 Diseño PAF y ETAPA

	Gr	Período 1			
		Sesión 1		Sesión 2	
		ITL	TDD	ITL	TDD
USD-UG	G11.1(15)	-	BSK_NS	-	-
	G11.2(17)	-	BSK_S	-	-
	G11.3(16)	BSK_NS	-	-	-
	G11.4(15)	BSK_S	-	-	-
USD-G	G11.5(3)	-	-	-	BSK_NS
	G11.6(4)	-	-	-	BSK_S
	G11.7(2)	-	-	BSK_NS	-
	G11.8(3)	-	-	BSK_S	-

Tabla 6.14 Diseño USD

	Gr	Sesión 1	Sesión 2
		ITL	TDD
OY2	G16.1(23)	BSK_S	MR_NS
	G16.2(23)	BSK_NS	MR_S

Tabla 6.15 Diseño OY2

	Gr	Sesión 1	Sesión 2
		ITL	TDD
NoviSad	G17.1(30)	BSK_NS	MR_NS
	G17.2(23)	MR_NS	BSK_NS

Tabla 6.16 Diseño NoviSad

Debido a que YW solo se estudia en 2 de los 16 experimentos, no contribuye mucho a la agregación y no se estudiará en la misma. Lo mismo ocurre con los objetos experimentales MF (se utiliza en 6 de los 16 experimentos), Sudoku (se utiliza en tan solo un experimento) y SS (se utiliza en 2 experimentos); no contribuyen mucho en la agregación y por tanto se prescinde de ellos. Por lo tanto, para el enfoque de desarrollo se estudiará ITL y TDD y para los objetos y su granularidad BSK_NS, BSK_S, MR_NS y MR_S.

A continuación, se describe como se han ido agregando cada uno de los diseños:

1. La Tabla 6.17 muestra que al agrupar FSH, FSK, FSO, EB y UPV solo se añaden participantes al único grupo existente G1.

Gr	Grupo Original	Sesión 1	Sesión 2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G8.1(31)	MR_NS	BSK_S

Tabla 6.17 Diseño agregado (Paso 1)

2. Al agrupar PT, ambos grupos de experimentos tienen un diseño intra-sujetos. Estos se complementan y se añaden el factor *Programa*, se crea un nuevo grupo (G2) y se añaden participantes al grupo ya existente G1. La Tabla 6.18, muestra el resultado de la superposición de estos experimentos.

Gr	Grupo Original	Sesión 1	Sesión 2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G8.1(31), G7.1(8)	MR_NS	BSK_S
G2	G7.2(9)	BSK_S	MR_NS

Tabla 6.18 Diseño agregado (Paso 2, añadir PT)

3. La Tabla 6.19 muestra que al superponer OY2, los diseños se complementan, se añaden nuevos niveles al factor programa, se crea un nuevo grupo (G3) y se añaden participantes al grupo ya existente G2.

Gr	Grupo Original	Sesión 1	Sesión 2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G7.1(8), G8.1(31)	MR_NS	BSK_S
G2	G7.2(9), G16.1(23)	BSK_S	MR_NS
G3	G16.2(23)	BSK_NS	MR_S

Tabla 6.19 Diseño agregado (Paso 3, añadir OY2)

4. Para agrupar UPM, de las sesiones 1 y 2 se consideran los grupos G6.2, G6.7 y G6.9 que es donde se utilizan los programas comunes BSK_S y MR_NS. Por otro lado, debido a que los sujetos aplican los tratamientos siempre sobre un programa distinto, de las sesiones 3 y 4 se consideran los grupos G6.4, G6.5 y G6.14, que es donde se utilizan los programas BSK_S y MR_NS, y se tiene en cuenta la vez en la que se aplica un tratamiento sobre un programa, por lo tanto, estos grupos se incluyen en la sesión 1 (primera vez) y 2 (segunda vez). La Tabla 6.20 muestra que este diseño está contenido en el anterior y se añaden más participantes a los grupos ya existentes G1 y G2.

Gr	Grupo Original	Sesión 1	Sesión 2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G7.1(8), G8.1(31), G6.7(1), G6.5(1), G6.14(1)	MR_NS	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.9 (1), G6.4(1)	BSK_S	MR_NS
G3	G16.2(22)	BSK_NS	MR_S

Tabla 6.20 Diseño agregado (Paso 4, añadir UPM)

5. La Tabla 6.21 muestra que al superponer NoviSad, se añaden dos nuevos grupos (G4 y G5) al diseño agregado.

Gr	Grupo Original	Sesión 1	Sesión 2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G6.5(1), G6.7(1), G6.14(1), G7.1(8), G8.1 (31), G9.2(2)	MR_NS	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.4(1), G6.9(1), G9.7(2)	BSK_S	MR_NS
G3	G16.2(22)	BSK_NS	MR_S
G4	G17.2(23)	MR_NS	BSK_NS
G5	G17.1(30)	BSK_NS	MR_NS

Tabla 6.21 Diseño agregado (Paso 7, añadir NoviSad)

6. En el siguiente paso se agregan PAF y ETAPA, dado que estos dos experimentos tienen el mismo diseño. Como ya se había mencionado anteriormente, se elimina la primera sesión, que es en la que se realiza YW y ahora las sesiones 2 y 3, se van a considerar como si fueran la primera y segunda sesión. La Tabla 6.22 muestra que este diseño está contenido en el anterior y se añaden más participantes a los grupos ya existentes G4 y G5.

Gr	Grupo Original	Sesión 1	Sesión 2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G6.5(1), G6.7(1), G6.14(1), G7.1(8), G8.1 (31), G9.2(2)	MR_NS	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.4(1), G6.9(1), G9.7(2)	BSK_S	MR_NS
G3	G16.2(22), G9.6(3)	BSK_NS	MR_S
G4	G17.2(23), G14.2(4), G15.2(3)	MR_NS	BSK_NS
G5	G17.1(30), G14.4(0), G15.4(1)	BSK_NS	MR_NS

Tabla 6.22 Diseño agregado (Paso 6, añadir PAF y ETAPA)

7. La Tabla 6.23 muestra que al agregar ER que se añaden nuevos niveles al factor programa, se crean dos nuevos grupos (G6 a G8) y se añaden participantes a cinco grupos ya existentes.

Gr	Grupo Original	Sesión 1	Sesión 2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G6.5(1), G6.7(1), G6.14(1), G7.1(8), G8.1 (31), G9.2(2)	MR_NS	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.4(1), G6.9(1), G9.7(2)	BSK_S	MR_NS
G3	G16.2(22), G9.6(3)	BSK_NS	MR_S
G4	G14.2(4), G15.2(3), G17.2(23), G9.1(3)	MR_NS	BSK_NS
G5	G14.4(0), G15.4(1), G17.1(30), G9.5(3)	BSK_NS	MR_NS
G6	G9.3(3)	MR_S	BSK_NS
G7	G9.4(2)	MR_S	BSK_S
G8	G9.8(2)	BSK_S	MR_S

Tabla 6.23 Diseño agregado (Paso 5, añadir ER)

8. Para agregar OY1, no se considera el factor *Realización del Trabajo*, es decir, se toman los datos como si todos los participantes hubiesen trabajado individualmente. En la Tabla 6.24 se puede observar que el diseño interseca parcialmente y se crean dos nuevos grupos (G9 y G10).

Gr	Grupo Original	S1	S2
		ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G6.5(1), G6.7(1), G6.14(1), G7.1(8), G8.1 (31), G9.2(2)	MR_NS	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.4(1), G6.9(1), G9.7(2)	BSK_S	MR_NS
G3	G16.2(22), G9.6(3)	BSK_NS	MR_S
G4	G9.1(3), G14.2(4), G15.2(3), G17.2(23)	MR_NS	BSK_NS
G5	G9.5(3), G14.4(0), G15.4(1), G17.1(30)	BSK_NS	MR_NS
G6	G9.3(3)	MR_S	BSK_NS
G7	G9.4(2)	MR_S	BSK_S
G8	G9.8(2)	BSK_S	MR_S
G9	G5.1(18)	BSK_S	-
G10	G5.2(23)	-	BSK_S

Tabla 6.24 Diseño agregado (Paso 8, añadir OY1)

9. En la Tabla 6.25 se puede observar que al agregar UB, éste se complementa con el diseño anterior y se crea un nuevo factor *Período*, el cual tiene dos niveles. Este coincide con la sesión, sin embargo, hay que distinguir que en una misma sesión se aplican los dos niveles del factor principal *Enfoque de desarrollo*, se crea un nuevo grupo (G11) y se añaden participantes al grupo ya existente G7.

Gr	Grupo Original	S1		S2	
		Período 1		Período 2	
		ITL	TDD	ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G6.5(1), G6.7(1), G6.14(1), G7.1(8), G8.1 (31), G9.2(2)	MR_NS	-	-	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.4(1), G6.9(1), G9.7(2)	BSK_S	-	-	MR_NS
G3	G16.2(22), G9.6(3)	BSK_NS	-	-	MR_S
G4	G9.1(3), G14.2(4), G15.2(3), G17.2(23)	MR_NS	-	-	BSK_NS
G5	G9.5(3), G14.4(0), G15.4(1), G17.1(30)	BSK_NS	-	-	MR_NS
G6	G9.3(3)	MR_S	-	-	BSK_NS
G7	G9.4(2), G10.2 (9)	MR_S	-	-	BSK_S
G8	G9.8(2)	BSK_S	-	-	MR_S
G9	G5.1(18)	BSK_S	-	-	-
G10	G5.2(23)	-	-	-	BSK_S
G11	G10.1 (11)	-	MR_S	BSK_S	-

Tabla 6.25 Diseño agregado (Paso 9, añadir UB)

10. Al agregar USD, se puede observar que este experimento, en realidad son dos: uno realizado con estudiantes de grado y el otro con estudiantes de posgrado. Ya que, aunque en el experimento con estudiantes de postgrado (USD-G) se aplican los tratamientos en la sesión 2, existe un único período (no hay orden de aplicación). Por lo tanto, se añaden estos grupos al primer período, sin tener en cuenta la sesión. La Tabla 6.26 muestra que el diseño interseca parcialmente, se crean nuevos grupos (G12 a G14) y se añaden participantes al grupo G9 ya existente.
11. Finalmente, en la Tabla 6.27 se puede observar que el diseño de USD2 está contenido en el diseño anterior y simplemente se añaden más participantes a los grupos ya existentes G9, G12, G13 y G14.

Gr	Grupo Original	S1		S2	
		Período 1		Período 2	
		ITL	TDD	ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G6.5(1), G6.7(1), G6.14(1), G7.1(8), G8.1 (31), G9.2(2)	MR_NS	-	-	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.4(1), G6.9(1), G9.7(2)	BSK_S	-	-	MR_NS
G3	G16.2(22), G9.6(3)	BSK_NS	-	-	MR_S
G4	G9.1(3), G14.2(4), G15.2(3), G17.2(23)	MR_NS	-	-	BSK_NS
G5	G9.5(3), G14.4(0), G15.4(1), G17.1(30)	BSK_NS	-	-	MR_NS
G6	G9.3(3)	MR_S	-	-	BSK_NS
G7	G9.4(2), G10.2 (9)	MR_S	-	-	BSK_S
G8	G9.8(2)	BSK_S	-	-	MR_S
G9	G5.1(18), G11.4(15), G11.8(3)	BSK_S	-	-	-
G10	G5.2(23)	-	-	-	BSK_S
G11	G10.1 (11)	-	MR_S	BSK_S	-
G12	G11.3(16), G11.7(2)	BSK_NS	-	-	-
G13	G11.1(15), G11.5(3)	-	BSK_NS	-	-
G14	G11.2(17), G11.6(4)	-	BSK_S	-	-

Tabla 6.26 Diseño agregado (Paso 10, añadir USD)

Gr	Grupo Original	S1		S2	
		Período 1		Período 2	
		ITL	TDD	ITL	TDD
G1	G1.1(8), G2.1(11), G3.1(6), G4.1(9), G6.5(1), G6.7(1), G6.14(1), G7.1(8), G8.1 (31), G9.2(2)	MR_NS	-	-	BSK_S
G2	G7.2(9), G16.1(23), G6.2(1), G6.4(1), G6.9(1), G9.7(2)	BSK_S	-	-	MR_NS
G3	G16.2(22), G9.6(3)	BSK_NS	-	-	MR_S
G4	G9.1(3), G14.2(4), G15.2(3), G17.2(23)	MR_NS	-	-	BSK_NS
G5	G9.5(3), G14.4(0), G15.4(1), G17.1(30)	BSK_NS	-	-	MR_NS
G6	G9.3(3)	MR_S	-	-	BSK_NS
G7	G9.4(2), G10.2 (9)	MR_S	-	-	BSK_S
G8	G9.8(2)	BSK_S	-	-	MR_S
G9	G5.1(18), G11.4(15), G11.8(3), G12.4(19)	BSK_S	-	-	-
G10	G5.2(23)	-	-	-	BSK_S
G11	G10.1 (11)	-	MR_S	BSK_S	-
G12	G11.3(16), G11.7(2), G12.3(17)	BSK_NS	-	-	-
G13	G11.1(15), G11.5(3), G12.1(15)	-	BSK_NS	-	-
G14	G11.2(17), G11.6(4), G12.2(13)	-	BSK_S	-	-

Tabla 6.27 Diseño agregado (Paso 11, añadir USD2)

6.2.2.2 Exploración de los Resultados de los Experimentos Individuales

La Tabla 6.28 muestra los factores estudiados en cada uno de los experimentos. Se han estudiado los principales factores y la interacción entre ellos. Debido que no todos los diseños son iguales se tienen que analizar de manera distinta.

Es importante tener en cuenta que, por restricciones de las pruebas estadísticas, se ha tenido que, o bien, transformar la variable respuesta, o utilizar otra prueba estadística:

- **Transformación de la variable.** Los experimentos en los que al transformar la variable se cumplen las restricciones de la prueba estadística son: USD-UG y NoviSad.
- **Elección de otra prueba estadística.** En USD2 a pesar de transformar la variable no se cumplen las restricciones de la prueba estadística y por lo tanto se ha utilizado la prueba

no paramétrica Kruskal-Wallis con 4 niveles: ITL-BSK_NS, ITL-BSK_S, TDD-BSK_NS y TDD-BSK_NS.

Exp	Variables					
	Enfoque de Desarrollo	Programa	Granularidad	Enfoque de Desarrollo * Programa	Enfoque de Desarrollo * Granularidad	Programa* Granularidad
OY1	x	N/A	N/A	N/A	N/A	N/A
FSH	x	N/A	N/A	N/A	N/A	N/A
FSK	x	N/A	N/A	N/A	N/A	N/A
FSO	x	N/A	N/A	N/A	N/A	N/A
EB	x	N/A	N/A	N/A	N/A	N/A
UPM	x	x	N/A	x	N/A	N/A
PT	x	x	N/A	x	N/A	N/A
UPV	x	N/A	N/A	N/A	N/A	N/A
ER	x	x	x	x	x	x
UB	x	x	N/A	x	N/A	N/A
USD	x	N/A	x	N/A	x	N/A
PAF	x	x	N/A	x	N/A	N/A
ETAPA	x	x	N/A	x	N/A	N/A
USD2	x	N/A	x	N/A	x	N/A
OY2	x	N/A	x	N/A	N/A	x
NS	x	x	N/A	x	N/A	N/A

Tabla 6.28 Pruebas estadísticas y factores analizados individualmente

- En los experimentos FSH, FSK, FSO, EB y UPV no se puede estudiar ni el programa, ni la interacción programa por enfoque de desarrollo, porque Programa y Enfoque de desarrollo están confundidos.
- Por otro lado, en OY1, USD, USD2 y OY2, tampoco se puede estudiar el programa, ni la interacción programa por enfoque de desarrollo porque sólo se utilizó un programa.
- La Granularidad y la interacción enfoque de desarrollo por granularidad y programa por granularidad solo se puede estudiar en ER, USD, USD2 y OY2 que son los experimentos en los que aparece como factor.

Los resultados de los análisis individuales de cada experimento se muestran en el Apéndice E.2. La Tabla 6.29 muestra los casos en los que se ha obtenido significación estadística (p -valor $<0,05$) y/o significación práctica ($r>0,23$).

Los casos en los que no existe ni significación estadística ni práctica aparecen representados con un guión, mientras que las variables que no se han considerado en ese análisis aparecen representados con un N/A.

Exp	Enfoque de Desarrollo	Programa	Granularidad	Enfoque de Desarrollo * Programa	Enfoque de Desarrollo * Granularidad	Programa* Granularidad
OY1	-	N/A	N/A	N/A	N/A	N/A
FSH	Estadística Práctica	N/A	N/A	N/A	N/A	N/A
FSK	Práctica	N/A	N/A	N/A	N/A	N/A
FSO	Estadística Práctica	N/A	N/A	N/A	N/A	N/A
EB	Estadística Práctica	N/A	N/A	N/A	N/A	N/A
UPM	N/A	Estadística Práctica	N/A	Práctica	N/A	N/A
PT	Estadística Práctica	Estadística Práctica	N/A	Práctica	N/A	N/A
UPV	Estadística Práctica	N/A	N/A	N/A	N/A	N/A
ER	Estadística Práctica	-	Estadística Práctica	Estadística Práctica	Estadística Práctica	Práctica
UB	-	Estadística Práctica	N/A	Práctica	N/A	N/A
USD-UG	-	N/A	Estadística Práctica	N/A	Práctica	N/A
USD-G	-	N/A	-	N/A	Práctica	N/A
PAF	-	Práctica	N/A	Práctica	N/A	N/A
ETAPA	Estadística Práctica	-	N/A	Estadística Práctica	N/A	N/A
USD2	-	N/A	N/A	-	Estadística Práctica	N/A
OY2	Estadística Práctica	N/A	Estadística	N/A	Práctica	N/A
NS	-	Estadística Práctica	N/A	Práctica	N/A	N/A

Tabla 6.29 Significación estadística o práctica en cada experimento (2/2)

6.2.2.3 Identificación del Conjunto de Variables de Diseño

Considerando el diseño agregado mostrado en la Tabla 6.27 se obtiene que las variables a incluir en el análisis son:

- Enfoque de Desarrollo, es el factor principal de la familia.
- Programa, viene impuesta por el diseño agregado.
- Período, viene impuesta por el diseño agregado.
- Enfoque de Desarrollo*Programa, viene impuesta por el diseño agregado.

Por otro lado, de los análisis individuales, las variables que se deben incluir son:

- Granularidad. Existen dos opciones de estudiarla: mezclada con el programa o como un factor independiente. Nótese que existen experimentos que la tienen como factor, otros que no y otros en los que está confundida con el programa. Finalmente, debido a que en aquellos experimentos que la estudian como factor es significativa, se la estudiará con la segunda opción, es decir, se opta por que sea un factor independiente.

- Enfoque de Desarrollo*Granularidad. Existe significación estadística y/o práctica en todos los experimentos, en los que se ha podido comprobar.
- Programa*Granularidad. Existe significación estadística y/o práctica en el único experimento, en el que se ha podido comprobar.

6.2.3 Identificación de las Variables Contextuales

En este paso se identifican aquellos elementos cuyos valores podrían provocar diferencias en los resultados experimentales.

6.2.3.1 Identificación de Diferencias Contextuales

Los elementos de la ejecución experimental a tener en cuenta a la hora de la identificación de diferencias contextuales aparecen en la Tabla 6.31 junto con los valores que toman para cada experimento de la familia. En gris se muestran aquellos que se considera podrían provocar diferencias en los resultados.

EJECUCIÓN EXPERIMENTAL			FS	OY1	EB	UPM	PT	UPV	ER	UB	USD	PAF	ETAPA	USD2	OY2	NoviSad
Diseño	Duración de la tarea	Limitada	-	-	x	-	-	x	-	-	-	-	x	x	x	x
		Ilimitada	x	x	-	x	x	-	x	x	x	x	-	-	-	-
Entrenamiento	Combinación entrenamiento y sesión	Secuencial	-	x	-	-	-	-	-	x	x	-	-	x	x	x
		Intercalado	x	-	x	x	x	x	x	-	-	x	x	-	-	-
	Aspectos de transmisión de los tratamientos		x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Duración del entrenamiento		x	x	x	x	x	x	x	x	x	x	x	x	x	x

Tabla 6.30 Elementos de la ejecución experimental

- La *duración de la tarea* podría afectar a los resultados de la agregación porque los participantes que trabajan con tiempo limitado podrían tener peores resultados ya que estarían trabajando bajo presión de conseguir terminar las tareas a tiempo y por tanto podrían aplicar el enfoque de desarrollo peor.
- La *combinación entre el entrenamiento y las sesiones* podría tener implicaciones en los resultados porque los participantes que lo recibieron de manera secuencial podrían no haber tenido tiempo suficiente para estudiar el último tratamiento enseñado, o viceversa, podrían haber olvidado el primero, provocando un conocimiento desigual de los mismos.
- La *transmisión del tratamiento* no afectará a los resultados ya que todos los entrenadores tienen experiencia alta en la enseñanza de TDD.
- La *duración del entrenamiento* no afecta a los resultados de la agregación porque este elemento está adaptado a los conocimientos y experiencia de los participantes y a las condiciones contextuales de cada experimento.

Los elementos de la población para tener en cuenta a la hora de la identificación de diferencias contextuales aparecen en la Tabla 6.32 junto con los valores que toman para cada experimento de la familia. En gris se muestran aquellos que se considera podrían provocar diferencias en los resultados.

POBLACIÓN			FS	OY1	EB	UPM	PT	UPV	ER	UB	USD	PAF	ETAPA	USD2	OY2	NoviSad
Sujetos	Tipo de sujetos	Profesional	X	-	X	-	X	-	X	-	-	X	X	-	-	-
		Estudiante de Máster	-	X	-	X	-	X	-	X	X	-	-	X	X	-
		Estudiante de Grado	-	X	-	-	-	-	-	-	X	-	-	X	X	X
Objetos	Complejidad	Más fácil	X	-	X	-	X	-	X	-	X	X	X	X	-	X
		Igual	-	X	-	X	-	X	-	X	-	-	-	-	-	-
	Dominio	No real	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		Java	X	X	-	X	X	X	-	X	-	X	-	-	X	X
	Lenguaje	C++	-	-	X	-	-	-	X	-	-	-	-	-	-	-
		Genexus Evolution 1	-	-	-	-	-	-	-	-	-	-	X	-	-	-
		C#	-	-	-	-	-	-	-	-	X	-	-	X	-	-
			-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Granularidad	BSK_S	X	X	X	X	X	X	X	X	X	-	-	X	X	-
		BSK_NS	-	-	-	-	-	-	X		X	X	X	X	X	X
		MR_S	-	-	-	-	-	-	X	X	-	-	-	-	X	-
		MR_NS	X	-	X	X	X	X	X	-	-	X	X	-	X	X
		MF_NS	X	-	X	X	X	-	-	-	-	-	-	-	-	-
		SS_NS	-	-	-	-	-	-	-	-	-	X	X	-	-	-
		S_NS	-	-	-	X	-	-	-	-	-	-	-	-	-	-
			-	-	-	-	-	-	-	-	-	-	-	-	-	-

Tabla 6.31 Elementos de la Población

- El *tipo de sujetos* que participa en el experimento podría afectar a los resultados, porque los profesionales pueden tener más conocimientos de programación, debido a que llevan más tiempo trabajando con programas reales. Aunque para los experimentos realizados en universidades hay diferentes tipos de estudiantes, esta variable se simplificará en estudiantes y profesionales. Debido a que existen experimentos que contemplan varios tipos de participantes y no es posible distinguir el tipo ya que esa información no se recopiló.
- La *complejidad* de los objetos experimentales no afecta a los resultados del experimento, ya que en ninguno de los casos la complejidad de los objetos es más difícil que las tareas que realizaban los sujetos en su entorno diario.
- El *dominio* de los objetos experimentales no afecta a los resultados del experimento porque en todos los experimentos se utilizaron tareas acondicionadas para el experimento (software de juguete).
- El *lenguaje de programación* en el que se realiza las tareas podría afectar a los resultados, ya que las características propias del lenguaje podrían hacer que ITL o TDD se comporten de distinta forma.
- La *granularidad* de las tareas podría afectar a los resultados, sin embargo, esta ya se ha considerado en el paso 2 del proceso (sección 6.2.2) incluyéndola como factor.

Los elementos de la instrumentación para tener en cuenta a la hora de la identificación de diferencias contextuales aparecen en la Tabla 6.33 junto con los valores que toman para cada

experimento de la familia. En gris se muestran aquellos que se considera podrían provocar diferencias en los resultados.

INSTRUMENTACIÓN		FS	OY1	EB	UPM	PT	UPV	ER	UB	USD	PAF	ETAPA	USD2	OY2	NoviSad
Guías Suplementarias (Stubs)	MR-Java	x	-	-	x	x	x	-	x	-	x	-	-	x	x
	BSK-Java	x	x	-	x	x	x	-	x	-	x	-	-	x	x
	MF-Java	x	-	-	x	x	-	-	-	-	-	-	-	-	-
	SS-Java	-	-	-	-	-	-	-	-	-	x	-	-	-	-
	S-Java	-	-	-	x	-	-	-	-	-	-	-	-	-	-
	MR-C++	-	-	x	-	-	-	x	-	-	-	-	-	-	-
	BSK-C++	-	-	x	-	-	-	x	-	-	-	-	-	-	-
	MF-C++	-	-	x	-	-	-	-	-	-	-	-	-	-	-
	BSK-C#	-	-	-	-	-	-	-	-	x	-	-	x	-	-
	Ninguna	-	-	-	-	-	-	-	-	-	-	x	-	-	-
Formularios	Ninguno	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Transmisión de tratamientos	Transparencias v1	x	x	x	x	-	-	-	x	x	-	-	x	x	x
	Transparencias v2	-	-	-	-	x	x	x	-	-	x	x	-	-	-
Infraestructura tecnológica	Junit, Eclipse	x	x	-	x	x	x	-	x	-	-	-	-	x	x
	BOOST, Eclipse, Vim	-	-	-	-	-	-	x	-	-	-	-	-	-	-
	Junit, IntelliJ, Eclipse	-	-	-	-	-	-	-	-	-	x	-	-	-	-
	GTEST, Eclipse CDT (linux)	-	-	x	-	-	-	-	-	-	-	-	-	-	-
	VS Testing Tool, Visual Studio	-	-	-	-	-	-	-	-	x	-	-	x	-	-
	GeneXus, Genexus unit testing facilities	-	-	-	-	-	-	-	-	-	-	x	-	-	-
	Besouro, EclEmma	-	-	-	-	-	-	-	-	-	-	-	-	x	-

Tabla 6.32 Elementos de la Instrumentación

- Las *guías suplementarias*, no afectan a los resultados ya que, están adaptadas a los objetos experimentales y el lenguaje de programación utilizado en cada experimento.
- Los *formularios* no tienen implicaciones en los resultados porque no se utilizaron formularios.
- La *transmisión del tratamiento (transparencias)*, podrían afectar a los resultados dado que existen dos versiones distintas de las transparencias utilizadas.
- La *infraestructura tecnológica*, no afecta a los resultados porque este elemento está adaptado al lenguaje de programación (IDE o framework específico para el lenguaje) y las preferencias de los participantes (más habituados con algunos IDEs).

Los elementos de los experimentadores para tener en cuenta a la hora de la identificación de diferencias contextuales aparecen en la Tabla 6.34 junto con los valores que toman para cada experimento de la familia. En gris se muestran aquellos que se considera podrían provocar diferencias en los resultados.

- El *diseñador* del experimento no debería afectar a los resultados, porque siempre hay un diseñador común a varios experimentos. Además, su posible impacto ya se considera en el diseño en sí mismo.
- El *entrenador* del experimento podría afectar a los resultados de la agregación porque, aunque mantuvieron comunicación entre los experimentadores, para intentar la menor

variación posible, las características personales y profesionales de cada persona puede hacer que la información se transmita de forma diferente.

EXPERIMENTADORES		FS	OY1	EB	UPM	PT	UPV	ER	UB	USD	PAF	ETAPA	USD2	OY2	NoviSad
Diseñador	Juristo	x	x	x	x	x	x	x	-	x	x	x	x	x	x
	Dieste	x	-	x	x	x	x	x	-	-	x	x	-	-	-
	Vegas	-	-	-	x	-	-	-	-	-	x	x	-	-	-
	Turhan	-	x	-	x	-	-	-	x	x	-	-	x	x	x
	Fucci	-	x	-	-	-	-	-	x	-	-	-	-	-	-
	Kuhrmann	-	-	-	-	-	-	-	-	x	-	-	x	-	-
	Karac	-	-	-	-	-	-	-	-	-	-	-	-	x	x
	Mandić	-	-	-	-	-	-	-	-	-	-	-	-	-	x
	Ramač	-	-	-	-	-	-	-	-	-	-	-	-	-	x
Entrenador	Turhan	x	-	x	x	-	-	-	-	-	-	-	-	-	-
	Scanniello	-	-	-	-	-	-	-	x	-	-	-	-	-	-
	Kuhrmann	-	-	-	-	-	-	-	-	x	-	-	x	-	-
	Dieste	-	-	-	-	x	x	x	-	-	x	x	-	-	-
	Fucci	-	x	-	-	-	-	-	-	-	-	-	-	-	-
	Karac	-	-	-	-	-	-	-	-	-	-	-	-	x	-
	Mandić	-	-	-	-	-	-	-	-	-	-	-	-	-	x
Monitor	Ramač	-	-	-	-	-	-	-	-	-	-	-	-	-	x
	Dieste	x	-	x	x	-	x	x	-	-	-	x	-	-	-
	Fucci	x	x	x	-	-	-	-	-	-	-	-	-	-	-
	Tosun	x	-	x	-	x	-	x	-	-	-	-	-	-	-
	Scanniello	-	-	-	-	-	-	-	x	-	-	-	-	-	-
	Romano	-	-	-	-	-	-	-	x	-	-	-	-	-	-
	Kuhrmann	-	-	-	-	-	-	-	-	x	-	-	x	-	-
	Santos	-	-	-	-	-	-	-	-	-	x	-	-	-	-
	Karac	-	-	-	-	-	-	-	-	-	-	-	-	x	x
	Mandić	-	-	-	-	-	-	-	-	-	-	-	-	-	x
	Ramač	-	-	-	-	-	-	-	-	-	-	-	-	-	x
Medidor	Uyaguari	x	-	x	x	x	x	x	x	x	x	-	-	-	-
	Fucci	-	x	-	-	-	-	-	-	-	-	-	-	-	-
	Palomeque	-	-	-	-	-	-	-	-	-	-	x	-	-	-
	Karac	-	-	-	-	-	-	-	-	-	-	-	x	x	x

Tabla 6.33 Elementos de los Experimentadores

- El *monitor* del experimento no afecta a los resultados porque todos sabían el procedimiento a seguir durante el desarrollo del experimento.
- El *medidor* del experimento no afecta a los resultados porque para la agregación se han considerado los datos obtenidos por el medidor Uyaguari, y para los experimentos en los que el medidor es distinto, ellos siguieron las guías descritas en (Uyaguari, 2017).

6.2.3.2 Identificación de Variables Contextuales

En esta actividad, se va a transformar las diferencias encontradas a variables a incluir en el paso 5. La Tabla 6.35 muestra el conjunto final de variables. En este caso hay una correspondencia 1 a 1 entre diferencias y variables, ya que no hay variables confundidas.

Variable	Valores	FS	OY1	EB	UPM	PT	UPV	ER	UB	USD	PAF	ETAPA	USD2	OY2	NoviSad
Duración de la tarea	Limitada	-	-	X	-	-	X	-	-	-	-	X	X	X	X
	Ilimitada	X	X	-	X	X	-	X	X	X	X	-	-	-	-
Combinación entrenamiento y sesión	Secuencial	-	X	-	-	-	-	-	X	X	-	-	X	X	X
	Intercalado	X	-	X	X	X	X	X	-	-	X	X	-	-	-
Tipo de sujetos	Profesional	X	-	X	-	X	-	X	-	-	X	X	-	-	-
	Estudiante	-	X	-	X	-	X	-	X	X	-	-	X	X	X
Lenguaje de Programación	Java	X	X	-	X	X	X	-	X	-	X	-	-	X	X
	C++	-	-	X	-	-	-	X	-	-	-	-	-	-	-
	Genexus Evolution 1	-	-	-	-	-	-	-	-	-	-	X	-	-	-
	C#	-	-	-	-	-	-	-	-	X	-	-	X	-	-
Transmisión de tratamientos	Transparencias v1	X	X	X	X	-	-	-	X	X	-	-	X	X	X
	Transparencias v2	-	-	-	-	X	X	X	-	-	X	X	-	-	-
Entrenador	Turhan	X	-	X	X	-	-	-	-	-	-	-	-	-	-
	Scanniello	-	-	-	-	-	-	-	X	-	-	-	-	-	-
	Kuhrmann	-	-	-	-	-	-	-	-	X	-	-	X	-	-
	Dieste	-	-	-	-	X	X	X	-	-	X	X	-	-	-
	Fucci	-	X	-	-	-	-	-	-	-	-	-	-	-	-
	Karac	-	-	-	-	-	-	-	-	-	-	-	-	X	-
	Mandić- Ramač	-	-	-	-	-	-	-	-	-	-	-	-	-	X

Tabla 6.34 Variables para el análisis

6.2.3.3 Refinamiento del Conjunto Inicial de Variables Contextuales

La última actividad de este paso es conocer cómo se distribuyen las variables del conjunto determinado en el paso anterior. La Tabla 6.36 y Tabla 6.37 muestran cuántos participantes (P) y observaciones (O) hay para cada variable y también para cada experimento.

Se puede observar que las variables tipo de sujetos (Profesionales: 156 y Estudiantes: 467 observaciones), lenguaje de programación (Java: 418, C++: 58, Genexus Evolution 1: 8 y C#: 139 observaciones) y transmisión de los tratamientos (ITL: 467 y TLD: 156 observaciones), están bastante desbalanceadas, por lo tanto, podría ocurrir que los resultados obtenidos no fuesen válidos y que fuese necesario eliminar algunas de estas variables del modelo.

6.2.4 Identificación de las Variables de Participantes

En este paso se usará la información recogida de los participantes sobre su experiencia o conocimientos.

6.2.4.1 Identificación de las Diferencias entre los Participantes

La Tabla 6.38 muestra las características variables de los participantes, es decir, la experiencia que tienen en programación, con el lenguaje utilizado, en pruebas unitarias y con el framework de pruebas utilizado en cada experimento. Debido a que el valor de cada una de estas variables es diferente para cada participante solo se muestra el tipo de valor que toma y en que experimentos se ha medido. A continuación, se detalla cada una de ellas:

Variable	Valores		FS	OY1	EB	UPM	PT	UPV	ER	UB	USD	PAF	ETAPA	USD2	OY2	NoviSad	TOTAL
Duración de la tarea	Limitada	P	-	-	9	-	-	33	-	-	-	-	4	64	45	53	208
		O	-	-	18	-	-	66	-	-	-	-	8	64	90	101	347
	Ilimitada	P	24	41	-	6	17	-	20	9	75	4	-	-	-	-	196
		O	48	41	-	12	34	-	40	18	75	8	-	-	-	-	276
Combinación entrenamiento y sesión	Secuencial	P	-	41	-	-	-	-	-	9	75	-	-	64	45	53	287
		O	-	41	-	-	-	-	-	18	75	-	-	64	90	101	389
	Intercalado	P	24	-	9	6	17	33	20	-	-	4	4	-	-	-	117
		O	48	-	18	12	34	66	40	-	-	8	8	-	-	-	234
Tipo de sujetos	Profesional	P	24	-	9	-	17	-	20	-	-	4	4	-	-	-	78
		O	48	-	18	-	34	-	40	-	-	8	8	-	-	-	156
	Estudiante	P	-	41	-	6	-	33	-	9	75	-	-	64	45	53	326
		O	-	41	-	12	-	66	-	18	75	-	-	64	90	101	467
Lenguaje de Programación	Java	P	24	41	-	6	17	33	-	9	-	4	-	-	45	53	232
		O	48	41	-	12	34	66	-	18	-	8	-	-	90	101	418
	C++	P	-	-	9	-	-	-	20	-	-	-	-	-	-	-	29
		O	-	-	18	-	-	-	40	-	-	-	-	-	-	-	58
	Genexus Evolution 1	P	-	-	-	-	-	-	-	-	-	-	4	-	-	-	4
		O	-	-	-	-	-	-	-	-	-	-	8	-	-	-	8
	C#	P	-	-	-	-	-	-	-	-	75	-	-	64	-	-	139
		O	-	-	-	-	-	-	-	-	75	-	-	64	-	-	139
Transmisión de tratamientos	Transparencias v1	P	24	41	9	6	-	-	-	9	75	-	-	64	45	53	326
		O	48	41	18	12	-	-	-	18	75	-	-	64	90	101	467
	Transparencias v2	P	-	-	-	-	17	33	20	-	-	4	4	-	-	-	78
		O	-	-	-	-	34	66	40	-	-	8	8	-	-	-	156
Entrenador	Turhan	P	24	-	9	6	-	-	-	-	-	-	-	-	-	-	39
		O	48	-	18	12	-	-	-	-	-	-	-	-	-	-	78
	Scanniello	P	-	-	-	-	-	-	-	9	-	-	-	-	-	-	9
		O	-	-	-	-	-	-	-	18	-	-	-	-	-	-	18
	Kuhrmann	P	-	-	-	-	-	-	-	-	75	-	-	64	-	-	139
		O	-	-	-	-	-	-	-	-	75	-	-	64	-	-	139
	Dieste	P	-	-	-	-	17	33	20	-	-	4	4	-	-	-	78
		O	-	-	-	-	34	66	40	-	-	8	8	-	-	-	156
	Fucci	P	-	41	-	-	-	-	-	-	-	-	-	-	-	-	41
		O	-	41	-	-	-	-	-	-	-	-	-	-	-	-	41
	Karac	P	-	-	-	-	-	-	-	-	-	-	-	-	45	-	45
		O	-	-	-	-	-	-	-	-	-	-	-	-	90	-	90
	Mandić- Ramač	P	-	-	-	-	-	-	-	-	-	-	-	-	-	53	53
		O	-	-	-	-	-	-	-	-	-	-	-	-	-	101	101

Tabla 6.35 Número de observaciones para cada variable (2/2)

- **Experiencia en Programación.** Recoge información sobre la experiencia que tienen los participantes en programación. Esta variable tiene los siguientes 4 niveles:
 - 1. Sin experiencia (< 2 años)
 - 2. Novato (2-5 años)
 - 3. Intermedio (5-10 años)
 - 4. Experto (> 10 años)
- **Experiencia con el lenguaje de programación.** Recoge información sobre la experiencia que tienen los participantes con el lenguaje de programación utilizado en cada uno de los experimentos. Esta variable tiene los siguientes 4 niveles:

- 1. No familiarizado (< 2 años)
- 2. Novato (2-5 años)
- 3. Intermedio (5-10 años)
- 4. Experto (> 10 años)

Exp	Programación	Lenguaje	Pruebas Unitarias	Framework de Pruebas
OY1	x	x	x	x
FSH	x	x	x	x
FSK	x	x	x	x
FSO	x	x	x	x
EB	x	x	x	x
UPM	x	x	x	x
PT	x	x	x	x
UPV	x	x	x	x
ER	x	x	x	x
UB	x	x	x	x
USD-UG	x	x	x	x
USD-G	x	x	x	x
PAF	x	x	x	x
ETAPA	x	x	x	x
USD2	x	x	x	x
OY2	x	x	x	x
NS	x	x	x	x

Tabla 6.36 Distribución de las variables de los participantes

- Experiencia en pruebas unitarias. Recoge información sobre la experiencia que tienen los participantes realizando pruebas unitarias. Esta variable tiene los siguientes 4 niveles:
 - 1. Sin experiencia (< 2 años)
 - 2. Novato (2-5 años)
 - 3. Intermedio (5-10 años)
 - 4. Experto (> 10 años)
- Experiencia con el framework de pruebas. Recoge información sobre la experiencia que tienen los participantes con el framework de pruebas utilizado en cada uno de los experimentos. Esta variable tiene los siguientes 4 niveles:
 - 1. No familiarizado (< 2 años)
 - 2. Novato (2-5 años)
 - 3. Intermedio (5-10 años)
 - 4. Experto (> 10 años)

6.2.4.2 Creación de un Conjunto Inicial de las Variables de los Participantes

Se tendrán en cuenta todas las variables comentadas en la actividad anterior, ya que, están disponibles en todos los experimentos. La Tabla 6.39 muestra el conjunto final de variables con sus valores y la escala usada para cada variable.

Elemento	Valor	Escala
Experiencia en Programación	Escala de Likert (con 4 respuestas)	Ordinal
Experiencia con el Lenguaje	Escala de Likert (con 4 respuesta)	Ordinal
Experiencia en Pruebas Unitarias	Escala de Likert (con 4 respuestas)	Ordinal
Experiencia con el Framework de Pruebas	Escala de Likert (con 4 respuestas)	Ordinal

Tabla 6.37 Conjunto final de las variables de participantes

6.2.4.3 Refinar el Conjunto Inicial de las Variables de los Participantes

Por último, es necesario conocer como se distribuyen las variables que se han elegido anteriormente. La Figura 6.1, Figura 6.2, Figura 6.3 y Figura 6.4 muestra el comportamiento de cada una de las variables.

La Figura 6.1 muestra que en OY1, UPV, UB y NoviSad la mayoría de los participantes son novatos en conocimientos de programación. Por el contrario, únicamente en FSH la mayor parte de los participantes son expertos. Finalmente, en UPM, PT, UPV, Ericsson y UB hay participantes sin experiencia. En general, el mayor número de participantes está concentrado entre los valores novato e intermedio. Esta variable no está confundida con ninguna otra porque todos los participantes tienen un amplio rango de experiencia y no puede sustituir a ninguna de las otras variables identificadas.

Figura 6.1 Experiencia en Programación

La Figura 6.2 muestra que en once experimentos (61%): FSK, OY1, EB, UPV, UB1, USDU, USDG, UB2, USD2, OY2 y NoviSad la mayor parte de los participantes son novatos en conocimientos del lenguaje de programación usado en cada uno de esos experimentos. En ningún caso el número de participantes expertos supera al resto de las categorías. En FSO, el número de expertos es igual que el número de novatos e intermedios. Finalmente, en EB, USDG, ETAPA y Novisad no hay participantes sin experiencia. En general, el mayor número de participantes está concentrado entre los valores novato e intermedio. Esta variable no está confundida con ninguna otra porque todos

los participantes tienen un amplio rango de experiencia y no puede sustituir a ninguna de las otras variables identificadas.

Figura 6.2 Experiencia con el Lenguaje de Programación

La Figura 6.3 muestra que once experimentos (61,11%): OY1, EB, UPM, PT, UPV, ER, UB1, UB2, ETAPA (todos los participantes), OY2 y NoviSad la mayor parte de los participantes no tienen experiencia en el desarrollo de pruebas unitarias, además en otros cinco experimentos: FSH, FSK, USDU, USDG y USD2 la mayor parte de sujetos son novatos. Únicamente en PAF la mayoría de los participantes tienen conocimientos intermedios. Finalmente, en FSO el número de intermedios es igual que el número de novatos. Por lo tanto, se puede observar que en la mayoría de los experimentos los participantes no tienen (o tienen muy pocos) conocimientos en pruebas unitarias. Esta variable no está confundida con ninguna otra porque todos los participantes tienen un amplio rango de experiencia y no puede sustituir a ninguna de las otras variables identificadas.

Figura 6.3 Experiencia en Pruebas Unitarias

La Figura 6.4 muestra que trece experimentos (72,22%): FSK, OY1, EB (todos los participantes), UPM, PT, UPV (todos los participantes), ER, UB1, UB2, ETAPA (todos los participantes), USD2, OY2 y NoviSad la mayor parte de los participantes no tienen experiencia con el framework de pruebas utilizado en cada uno de los experimentos, además en otros tres experimentos: FSO, USDU y USDG la mayor parte de sujetos son novatos y en FSH el número de novatos es igual que el número de participantes sin experiencia. Finalmente, únicamente en PAF la mayoría de los participantes tienen conocimientos intermedios. Por lo tanto, se puede observar que en la mayoría de los experimentos los participantes no tienen (o tienen muy pocos) conocimientos con el framework de pruebas. Esta variable no está confundida con ninguna otra porque todos los participantes tienen un amplio rango de experiencia y no puede sustituir a ninguna de las otras variables identificadas.

Figura 6.4 Experiencia con el framework de pruebas

Se puede observar que la variable experiencia con el framework de pruebas está desbalanceada ya que existen pocos datos para la categoría expertos. En caso de que ocurran errores en la convergencia de las matrices a la hora de analizar, se tendría que considerar su eliminación del modelo.

6.2.5 Conducir Análisis

6.2.5.1 Identificación del Modelo de Análisis

Como ya se ha mencionado hay varios experimentos ejecutados en Finlandia, España, Italia y Dinamarca. Una posible estructura jerárquica podría ser de 4 niveles. El nivel más bajo corresponde a las medidas repetidas tomadas de cada participante, el segundo nivel a los participantes, el tercer nivel a los experimentos y el cuarto nivel a los sitios. Sin embargo, esta estructura no está equilibrada, ya que los sitios PT, ETAPA y NoviSad tienen un experimento cada uno. Por lo tanto, se ha optado por una jerarquía de 3 niveles como la que se muestra en la Figura 6.5 y la Figura 6.6. El nivel 1 corresponde a medidas repetidas, el nivel 2 a los participantes y el nivel 3 a los experimentos. La definición de las variables que se analizarán se muestra en la Tabla 6.40.

Figura 6.5 Modelo jerárquico familia TDD (1/2)

Figura 6.6 Modelo jerárquico familia TDD (2/2)

Variable	Nivel	Valores	Escala
Calidad	1	0-100	Intervalo
Enfoque de desarrollo	1	ITL TDD	Nominal
Objeto experimental	1	BSK MR	Nominal
Período	1	1-2	Ordinal
Granularidad	1	Grano Fino (S)	Nominal
		Grano Grueso (NS)	
Experiencia en Programación	2	1-4	Ordinal
Experiencia con el Lenguaje de Programación	2	1-4	Ordinal
Experiencia en pruebas unitarias	2	1-4	Ordinal
Experiencia con el framework de pruebas	2	1-4	Ordinal
Duración de la tarea	3	Limitada Ilimitada	Nominal
Combinación entre entrenamiento y sesiones	3	Secuencial Intercalada	Nominal
Tipo de sujetos	3	Profesional Estudiante	Nominal
Lenguaje de Programación	3	Java C++ Genexus Evolution 1 C#	Nominal
Transmisión de tratamientos	3	ITL TLD	Nominal
Entrenador	3	Turhan Scanniello Kuhrmann Dieste Fucci Karac Mandić- Ramač	Nominal

Tabla 6.38 Definición de las variables de análisis

Las preguntas de investigación a responder con el análisis son:

- RQ_1 : ¿Depende la calidad del código generado del enfoque de desarrollo utilizado, el programa, la granularidad, el período y/o la interacción de enfoque de desarrollo*programa, enfoque de desarrollo*granularidad y programa*granularidad?
- RQ_2 : ¿Está relacionada la calidad del código generado con la experiencia de los participantes en programación, la experiencia con el lenguaje de programación usado en el experimento, la experiencia en pruebas unitarias y/o la experiencia con el framework de pruebas usado en el experimento?
- RQ_3 : ¿Está relacionada la calidad del código generado con las diferencias entre los experimentos representadas por las variables duración de la tarea, tipo de sujetos, el lenguaje de programación utilizado en el experimento, transmisión del tratamiento, el entrenador y la combinación entre entrenamiento y sesiones?
- RQ_4 : ¿El enfoque de desarrollo se comporta de forma diferente en los diferentes experimentos?
- RQ_5 : ¿La experiencia de los participantes en programación, la experiencia con el lenguaje de programación usado en el experimento, la experiencia en pruebas unitarias

y/o la experiencia con el framework de pruebas usado en el experimento interactúan con el enfoque de desarrollo?

- *RQ₆*: ¿Las diferencias entre los experimentos representadas por las variables duración de la tarea, tipo de sujetos, el lenguaje de programación utilizado en el experimento, transmisión del tratamiento, el entrenador y la combinación entre entrenamiento y sesiones interactúan con el enfoque de desarrollo?

6.2.5.2 Análisis del Modelo

La Tabla 6.41 muestra los resultados del AIC para el modelo nulo, incluyendo solo las variables de diseño (sin distinguir entre experimentos): enfoque de desarrollo, programa, granularidad, período y las interacciones enfoque de desarrollo*programa, enfoque de desarrollo*granularidad y programa*granularidad. En todos los casos, se prueban las 5 matrices de covarianza: identidad, diagonal, auto-regresiva de primer orden, simetría compuesta y no estructurada. Dado que hay 3 variables intra-participantes (enfoque de desarrollo, período y programa), se prueba como efectos repetidos cada una de ellas. El modelo con mejor ajuste (AIC más bajo) aparece en gris claro y corresponde a los efectos repetidos del enfoque de desarrollo y la estructura de covarianza diagonal.

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Enfoque de Desarrollo	Identidad	5986,513	Si	
	CS	5988,513	Si	No converge
	Diagonal	5981,610	Si	
	No estructurada	5983,610	Si	No converge
	AR(1)	5988,513	Si	La matriz no está definida positivamente
Programa	Identidad	5986,513	Si	
	CS	5988,513	Si	No converge
	Diagonal	5988,198	Si	
	No estructurada	5990,198	Si	No converge
	AR(1)	5988,513	Si	La matriz no está definida positivamente
Período	Identidad	5986,513	Si	
	CS	5988,513	Si	No converge
	Diagonal	5984,148	Si	
	No estructurada	5986,148	Si	No converge
	AR(1)	5988,132	Si	La matriz no está definida positivamente

Tabla 6.39 AICs para el modelo básico familia TDD

Este modelo nulo se refina aún más como se explica en la Sección 4.6.2. Los resultados detallados del cálculo del AIC se muestran en el Apéndice E.2. La Tabla 6.42 muestra un resumen de estos, donde se refleja el AIC para el mejor modelo obtenido cada vez que se agregan nuevas variables al modelo.

En la Tabla 6.42 se puede observar que el AIC baja siempre que se pasa de un paso al siguiente hasta llegar al paso de interacciones entre variables de distinto nivel. Concretamente, cuando se pasa del modelo nulo al modelo básico en el que se añade la estructura jerárquica, el AIC se reduce en 20,453. Estos resultados indican que es necesario reconocer el hecho de que existen 16

conjuntos de datos y, por lo tanto, no se pueden analizar los datos como si provinieran de un solo experimento.

Modelo	Efecto Repetido	Matriz de covarianza	AIC
Nulo (Sin jerarquía)	Enfoque de desarrollo	Diagonal	5981,610
Añadiendo jerarquía	Enfoque de desarrollo	Diagonal	5961,157
Añadiendo variables de experimento y participante	Enfoque de desarrollo	Diagonal	5620,467
Especificando enfoque de desarrollo como efecto aleatorio	Enfoque de desarrollo	Diagonal	5606,013
Enfoque de desarrollo*Exp. en Programación	Enfoque de desarrollo	Diagonal	5605,060
Enfoque de desarrollo*Exp. en Programación, Enfoque de desarrollo*Exp. con Lenguaje	Enfoque de desarrollo	Diagonal	5606,936
Enfoque de desarrollo*Exp. en Programación, Enfoque de desarrollo*Tipo de Sujeto	Enfoque de desarrollo	Diagonal	5603,930
Enfoque de desarrollo*Exp. en Programación, Enfoque de desarrollo*Tipo de Sujeto, Enfoque de desarrollo*Duración de la tarea	Enfoque de desarrollo	Diagonal	5604,057
Enfoque de desarrollo*Exp. en Programación, Enfoque de desarrollo*Tipo de Sujeto, Enfoque de desarrollo*Duración de la tarea, Enfoque de desarrollo*Entrenador	Enfoque de desarrollo	Diagonal	5604,166
Enfoque de desarrollo*Exp. en Programación, Enfoque de desarrollo*Tipo de Sujeto, Enfoque de desarrollo*Duración de la tarea, Enfoque de desarrollo*Entrenador, Enfoque de desarrollo*Exp. con el framework de pruebas	Enfoque de desarrollo	Diagonal	5605,310

Tabla 6.40 Resultados del análisis para el modelo refinado

Cuando se incorporan las variables de los niveles identificados, en este caso participantes y experimento, el AIC disminuye hasta 5620,467, es decir, se reduce en 340,690 con respecto al modelo básico. Esto significa que las diferencias entre los participantes y experimento se tienen que reconocer en el modelo.

Al especificar el enfoque de desarrollo como factor aleatorio (para expresar la posibilidad de que el enfoque de desarrollo se comporta de manera diferente para cada experimento), el AIC disminuye a 5606,013, es decir, se reduce en 14,454 con respecto al modelo en el que se incluyen las variables. Esto significa que el enfoque de desarrollo no se comporta igual en todos los experimentos.

En la Figura 6.7 se muestra como se van añadiendo las interacciones entre distintos niveles una a una, si al añadir una interacción se reduce el AIC a ese modelo se le incorporan nuevas interacciones. Al añadir la interacción Enfoque de desarrollo*Experiencia en programación, el AIC disminuye a 5605,060 así que esta interacción se mantiene en el modelo. Sin embargo, al añadir a este último modelo la interacción Enfoque de desarrollo*Experiencia con el lenguaje programación el AIC aumenta; se vuelve al modelo anterior el que contenía una única interacción y se añade otra interacción. El AIC sigue subiendo hasta que se añade la interacción Enfoque de desarrollo*Tipo de sujeto, que disminuye a 5603,930. A este modelo se le añade las siguientes interacciones, al igual que el caso anterior, el AIC sube hasta que se añade la interacción Enfoque de desarrollo*Duración

de la tarea. Aunque este AIC es ligeramente mayor que el del modelo del que se partía (0,127), sigue siendo menor que el del paso anterior, en el que se tiene el enfoque de desarrollo como efecto aleatorio, con estas tres interacciones es AIC se reduce de 5606,013 a 5604,057, es decir, se reduce en 1,956, por tanto, se mantienen estas tres interacciones en el modelo y se le vuelven a añadir las interacciones una por una. Al añadir la interacción Enfoque de desarrollo*Entrenador, se obtiene un AIC de 5604,166, al igual que antes este es mayor que el del paso anterior (aumenta 0,109), pero menor que el del modelo del que se partía (se reduce en 1,847), por lo tanto, esta interacción se añade al modelo. A este modelo se le vuelven a añadir las interacciones y de la misma manera que en el caso anterior al añadir la interacción enfoque de desarrollo*experiencia con el framework de pruebas, el AIC es mayor que el del paso anterior (aumenta 1,144), pero menor que el del modelo del que se partía (se reduce en 0,703). A partir de este modelo, para cualquier interacción que se añade el AIC es mayor que el del modelo 4. Por lo tanto, finalmente, las interacciones para este modelo son enfoque de desarrollo*experiencia en programación, enfoque de desarrollo*tipo de sujeto, enfoque de desarrollo*duración de la sesión, enfoque de desarrollo*entrenador y enfoque de desarrollo*experiencia con el framework de pruebas.

Figura 6.7 Modelo 5: Incorporar interacciones de entre variables de distinto nivel

El modelo final obtenido para esta familia de experimentos incluye:

- Enfoque de desarrollo, Programa, Granularidad, Período y las interacciones Enfoque de desarrollo*Programa, Enfoque de desarrollo*Granularidad y Programa*Granularidad como factores fijos.
- El participante como una variable aleatoria.
- El experimento como una variable aleatoria.
- Las variables a los dos niveles son:
 - De participante: experiencia en programación, experiencia con el lenguaje de programación utilizado en el experimento, experiencia en pruebas unitarias y la experiencia con el framework de pruebas utilizado en el experimento.

- De experimento: duración de la tarea, tipo de sujeto, lenguaje, transmisión del tratamiento, entrenador y combinación entre el entrenamiento y las sesiones experimentales.
- El enfoque de desarrollo como un factor aleatorio.
- Las interacciones Enfoque de desarrollo*Experiencia en programación, Enfoque de desarrollo*Tipo de sujeto, Enfoque de desarrollo*Duración de la tarea, Enfoque de desarrollo*Entrenador y Enfoque de desarrollo*Experiencia con el framework de pruebas como factores fijos.

Una vez se ha elegido el mejor modelo, se realiza el análisis de este. Nótese que antes de llegar al modelo elegido, sobre el que finalmente se realiza el análisis se ha calculado el AIC de un total de 405 modelos: 5 matrices de covarianza x 3 variables de medidas repetidas x 27 modelos:

1. Modelo nulo
2. Modelo básico con jerarquía
3. Incorporando variables de análisis
4. Técnica como efecto aleatorio
5. Interacción Enfoque de Desarrollo*Experiencia en Programación
6. Interacciones: Enfoque de Desarrollo*Experiencia en Programación y Enfoque de Desarrollo*Experiencia con el Lenguaje de Programación
7. Interacciones: Enfoque de Desarrollo*Experiencia en Programación y Enfoque de Desarrollo*Experiencia en Pruebas Unitarias
8. Interacciones: Enfoque de Desarrollo*Experiencia en Programación y Enfoque de Desarrollo*Experiencia con el framework
9. Interacciones: Enfoque de Desarrollo*Experiencia en Programación y Enfoque de Desarrollo*Tipo de Sujeto
10. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Experiencia con el Lenguaje de Programación
11. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Experiencia en Pruebas Unitarias
12. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Experiencia con el framework
13. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Lenguaje
14. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Aspectos de transmisión del tratamiento
15. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Entrenador
16. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Orden entre entrenamiento y sesiones
17. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto y Enfoque de Desarrollo*Duración de la Sesión

18. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión y Enfoque de Desarrollo*Experiencia con el Lenguaje de Programación.
19. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión y Enfoque de Desarrollo*Experiencia en Pruebas Unitarias
20. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión y Enfoque de Desarrollo*Experiencia con el framework
21. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión y Enfoque de Desarrollo*Lenguaje
22. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión y Enfoque de Desarrollo*Aspectos de transmisión del tratamiento
23. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión y Enfoque de Desarrollo*Entrenador
24. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión y Enfoque de Desarrollo*Orden entre entrenamiento y sesiones
25. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión, Enfoque de Desarrollo*Entrenador y Enfoque de Desarrollo*Experiencia con el Lenguaje de Programación
26. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión, Enfoque de Desarrollo*Entrenador y Enfoque de Desarrollo*Experiencia en Pruebas Unitarias
27. Interacciones: Enfoque de Desarrollo*Experiencia en Programación, Enfoque de Desarrollo*Tipo de Sujeto, Enfoque de Desarrollo*Duración de la Sesión, Enfoque de Desarrollo*Entrenador y Enfoque de Desarrollo*Experiencia con el framework

6.2.5.3 Interpretación de los Resultados

La Tabla 6.41, Tabla 6.42, la Figura 6.8 y la Figura 6.9 muestran los resultados del análisis para el modelo final, con las que se podrán responder las preguntas de investigación planteadas:

- *RQ₁*: ¿Depende la calidad del código generado del enfoque de desarrollo, el programa, la granularidad, el período y/o la interacción de enfoque de desarrollo*programa, enfoque de desarrollo*granularidad y programa*granularidad?
 - El enfoque de desarrollo no influye en la calidad, ya que ambos se comportan de la misma manera.
 - El programa influye en la calidad. MR se comporta peor que BSK
 - La granularidad influye en la calidad. Grano grueso se comporta peor que grano fino.

Origen	gl numerador	gl denominador	F	Sig.
Intersección	1	37,377	1,068	,308
Enfoque desarrollo	1	141,563	,143	,706
Programa	1	155,593	60,718	,000
Granularidad	1	381,582	21,606	,000
Período	1	73,445	12,800	,001
Enfoque desarrollo * Programa	1	248,028	2,444	,119
Enfoque desarrollo * Granularidad	1	410,784	2,966	,086
Programa * Granularidad	1	343,884	,307	,580
Experiencia en Programación	1	371,469	9,572	,002
Experiencia con el lenguaje	1	360,827	3,741	,054
Experiencia en pruebas unitarias	1	401,559	3,058	,081
Experiencia con el framework de pruebas	1	415,508	,056	,813
Tipo de sujeto	1	22,400	2,907	,102
Lenguaje	1	24,001	2,380	,136
Transmisión del tratamiento	1	19,165	6,793	,017
Entrenador	1	13,631	1,348	,266
Combinación entrenamiento y sesiones	1	20,630	,130	,723
Duración de la tarea	1	17,757	4,151	,057
Enfoque desarrollo * Exp. Programación	1	317,028	,979	,323
Enfoque desarrollo * Tipo de sujeto	1	55,258	6,641	,013
Enfoque desarrollo * Duración de la tarea	1	36,118	5,150	,029
Enfoque desarrollo * Entrenador	1	31,934	2,192	,149
Enfoque desarrollo * Exp. con framework pruebas	1	345,256	,859	,355

Tabla 6.41 Resultados del análisis para el modelo básico

- El período influye en la calidad. El primer período se comporta peor que el segundo.
 - La interacción Enfoque de desarrollo*Programa no influye en la calidad.
 - La interacción Enfoque de desarrollo*Granularidad no influye en la calidad.
 - La interacción Programa*Granularidad no influye en la calidad.
- RQ₂: ¿Está relacionada la calidad del código generado con la experiencia de los participantes en programación, la experiencia con el lenguaje de programación usado en el experimento, la experiencia en pruebas unitarias y/o la experiencia con el framework de pruebas usado en el experimento?
 - La experiencia en programación influye en la calidad. La Figura 6.7 muestra que los participantes con mayor experiencia generan un código de mayor calidad.

Figura 6.8 Resultados para la Experiencia en Programación

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Programa	0,000	MR BSK	29,130 51,959	MR<BSK	0,000	-0,35
Granularidad	0,000	NS S	34,021 47,068	NS<S	0,000	-0,31
Período	0,001	P1 P2	30,966 50,124	P1<P2	0,001	-0,15
Experiencia en programación	0,002	Sin experiencia Novato Intermedio Experto	29,370 35,308 47,917 52,404	Creciente	N/A	2,61
Transmisión del tratamiento	0,017	Transparencias v1 Transparencias v2	40,109 46,204	v1<v2	0,017	-0,06
Enfoque de desarrollo *	0,013	ITL=>Prof	36,138	ITL=>Prof<Est	N/A	-0,08
Tipo de sujeto		ITL=>Est TDD=>Est TDD=>Prof	41,429 42,614 46,398	TDD=>Est<Prof	N/A	-0,05
Tipo de sujeto*Enfoque de desarrollo	0,013	Prof=>ITL	36,138	Prof=>ITL<TDD	N/A	-0,15
		Prof=>TDD Est=>ITL Est=>TDD	46,398 41,429 42,614	Est=>ITL=TDD	N/A	-0,03
Enfoque de desarrollo *	0,029	ITL=>Ilimitada	38,539	ITL=>Ilimitada<Limitada	N/A	-0,06
Duración de la tarea		ITL=>Limitada TDD=>Limitada TDD=>Ilimitada	41,320 41,280 46,445	TDD=>Limitada<Ilimitada	N/A	-0,07
Duración de la tarea*	0,029	Ilimitada=>ITL	38,539	Ilimitada=>ITL<TDD	N/A	-0,12
Enfoque de desarrollo		Ilimitada=>TDD Limitada=>TDD Limitada=>ITL	46,445 41,280 41,320	Limitada=>TDD=ITL	N/A	-0,02

Tabla 6.42 Resultados del análisis

- La experiencia con el lenguaje de programación, la experiencia en pruebas unitarias y la experiencia con el framework de pruebas no influye en la calidad.
- RQ₃: ¿Está relacionada la calidad del código generado con las diferencias entre los experimentos representadas por las variables duración de la tarea, tipo de sujetos, el lenguaje de programación utilizado en el experimento, transmisión del tratamiento, el entrenador y la combinación entre entrenamiento y sesiones?
 - La transmisión del tratamiento influye en la calidad. La versión 1 de las transparencias se comportan peor que la versión 2.
 - El tipo de sujetos, el lenguaje de programación utilizado en el experimento, el entrenador, la combinación entre entrenamiento y sesiones y la duración de la tarea no influye en la calidad.
- RQ₄: ¿El enfoque de desarrollo se comporta de forma diferente en los diferentes experimentos?
 - Si, ya que, al añadir el enfoque de desarrollo como efecto aleatorio, el AIC disminuye. La Figura 6.8 muestra el comportamiento del enfoque de desarrollo para cada experimento de la familia.

Figura 6.9 Comportamiento del enfoque de desarrollo a través de los distintos experimentos

- **RQ₅:** ¿La experiencia de los participantes en programación, la experiencia con el lenguaje de programación usado en el experimento, la experiencia en pruebas unitarias y/o la experiencia con el framework de pruebas usado interactúan con el enfoque de desarrollo?
 - No, aunque el AIC disminuye con las interacciones Enfoque de desarrollo*Experiencia en programación y Enfoque de desarrollo*Experiencia con el framework de pruebas estas no son significativas
- **RQ₆:** ¿Las diferencias entre los experimentos representadas por las variables duración de la tarea, tipo de sujetos, el lenguaje de programación utilizado en el experimento, transmisión del tratamiento, el entrenador y la combinación entre entrenamiento y sesiones interactúan con el enfoque de desarrollo?
 - Las interacciones Enfoque de desarrollo*Tipo de sujeto y Enfoque de desarrollo*Duración de la tarea y Enfoque de desarrollo*Entrenador son significativas. Con respecto Enfoque de desarrollo*Tipo de sujeto, cuando los participantes aplican ITL, estudiantes se comportan mejor que profesionales, sin embargo, cuando aplican TDD, ocurre lo contrario, profesionales se comportan mejor que estudiantes. Finalmente, con respecto a Enfoque de desarrollo*Duración de la tarea, cuando los participantes aplican TDD con la duración ilimitada se comporta mejor que limitada, sin embargo, en cuando aplican ITL la duración limitada se comporta mejor que ilimitada. Sin embargo, dado que ambos casos el tamaño del efecto es despreciable o pequeño estas dos variables no interactúan con el enfoque de desarrollo.

De todo esto se puede concluir que:

- Los programas se comportan de forma diferente, aunque no hay interacción con el método de desarrollo, lo que indica una simple variabilidad entre los programas.

- Los programas de grano fino consiguen una mejor calidad en general, independientemente del método de desarrollo y del programa. Esto hace pensar que en general son más sencillos de desarrollar.
- El período influye en la calidad, lo que indica la existencia de efecto de aprendizaje.
- El único tipo de experiencia que influye en los resultados es la experiencia en programación. A mayor experiencia, mayor calidad. El resto de las experiencias (con el lenguaje de programación utilizado, con el framework de pruebas utilizado y en pruebas unitarias) no influyen.
- Los profesionales consiguen mejor calidad con TDD que con ITL, sin embargo, para los estudiantes no hay diferencia entre ambos tratamientos.
- Si la duración de la tarea es ilimitada se consigue mayor calidad con TDD que con ITL, sin embargo, para una duración limitada no hay diferencias entre ambos tratamientos.

Capítulo 7

IMPACTO DEL USO DE ESTUDIANTES EN EXPERIMENTOS DE INGENIERÍA DEL SOFTWARE

En este capítulo se estudia el impacto de realizar experimentos en Ingeniería del Software, utilizando estudiantes como participantes.

Los resultados de este capítulo se encuentran en la publicación:

- Sira Vegas, Patricia Riofrío, Natalia Juristo. Does Subject Type Influence Software Engineering Experiment Results? *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, Buenos Aires, Argentina, Mayo 2017, pp. 210-212.

7.1 Introducción

Las simplificaciones de la realidad son imprescindibles en los experimentos de laboratorio. Una simplificación muy común es el uso de estudiantes en lugar de desarrolladores profesionales en experimentos de laboratorio IS. Varios autores han llegado a hallazgos similares con respecto a las tasas de uso de estudiantes y profesionales en experimentos de SE. Höfer y Tichy (Höfer & Tichy, 2007) informan que el 60% de los estudios experimentales en SE utilizan estudiantes, mientras que el 22% emplea profesionales y solo el 14% utiliza ambos grupos de participantes. Sjøberg y col. (Sjøberg, et al., 2005) concluyen que, de 103 artículos que reportan experimentos controlados realizados entre 1993 y 2002, el 87% usa estudiantes y solo el 9% emplea profesionales. Además, los estudiantes universitarios se usan con más frecuencia que los estudiantes de posgrado. Pero los experimentos realizados con estudiantes sufren de la interacción de la relación causal con la amenaza de validez externa de las unidades³ (Shadish, et al., 2001).

³ Un efecto identificado para un tipo de unidad podría no ser válido para estudios con otro tipo de unidad (Shadish, et al., 2001).

En el Capítulo 2 de esta tesis, se pudo comprobar que en el 89% de las familias se identifican variables moderadoras relacionadas con los participantes, de estas en una gran mayoría (38,89%) se identifica el Tipo de sujeto como variable moderadora. No obstante, a pesar del interés de investigar exhaustivamente el impacto del uso de estudiantes en experimentos de IS, no se ha hecho ningún esfuerzo previo para mejorar nuestro conocimiento.

Se realizó una investigación inicial sobre la viabilidad de realizar dicho estudio, y se encontró una cantidad importante de experimentos que han comparado el desempeño de estudiantes y profesionales. Por ejemplo⁴ (Berander, 2004), (Carver, et al., 2003), (Höst, et al., 2000), (McMeekin, et al., 2009), (Porter & Votta, 1998), (Remus, 1989), (Runeson., 2003), (Svahnberg, et al., 2008). Con la excepción de (Carver, et al., 2003) y (Svahnberg, et al., 2008), todos son estudios experimentales. Estos estudios experimentales difieren con respecto al enfoque utilizado para comparar los tipos de sujetos y los resultados.

En algunos experimentos, el *tipo de sujeto se incluye como factor* en el experimento. En el contexto de la priorización de requisitos, (Berander, 2004) descubre que los estudiantes tienen un desempeño peor que los profesionales, sin embargo, los estudiantes que tienen experiencia en proyectos no tienen peor desempeño. En el contexto de la evaluación de impacto en el tiempo de entrega, (Höst, et al., 2000) no encuentra diferencias entre estudiantes y profesionales. En el contexto de las técnicas de inspección, (McMeekin, et al., 2009) encuentran que los profesionales se desempeñan mejor que los estudiantes. En el contexto de la toma de decisiones de programación de producción, (Remus, 1989) encuentra que los estudiantes universitarios tienen un desempeño peor que los profesionales.

Otros experimentos *comparan sus resultados con los de estudios anteriores realizados por otros investigadores que realizaron el experimento con otro tipo de sujeto*. En el contexto del Proceso de Software Personal (PSP), (Runeson., 2003) encuentra que existen diferencias significativas al comparar estudiantes de pregrado y posgrado, pero las diferencias son pequeñas entre los profesionales industriales y los estudiantes de posgrado.

En otros experimentos, *el mismo equipo de investigadores replica el mismo experimento* en la academia y la industria. En el contexto de las técnicas de lectura de inspección, (Porter & Votta, 1998) concluyen que no hay diferencia entre los resultados logrados por los estudiantes y los profesionales. En el contexto de los métodos de análisis de seguridad.

Se ha identificado la necesidad de sintetizar la evidencia sobre el impacto del uso de profesionales o estudiantes en experimentos de SE mediante el examen exhaustivo de sus resultados. Para este propósito, se ha revisado la literatura para estudiar el impacto en los resultados del uso de profesionales versus estudiantes y las características de los sujetos en los experimentos de SE publicados. Hemos realizado un estudio de mapeo sistemático de conformidad con (Petersen, et al., 2008). El mapeo sistemático es una metodología para investigar la literatura en un campo de particular interés con el fin de determinar la naturaleza, el alcance y el número de estudios primarios publicados (Petersen, et al., 2008). Los estudios de mapeo clasifican los estudios primarios para dar una visión general del campo de investigación en consideración.

⁴ Este conjunto de estudios se ha incluido solo para ilustrar el estado actual de la cuestión. No tiene la intención de ser exhaustivo.

7.2 Método de Investigación

Este estudio de mapeo sistemático busca respuestas a las siguientes preguntas de investigación:

- *RQ₁*: ¿Cuál es la escala de uso de estudiantes y profesionales en experimentos de IS? Interesa conocer la frecuencia con la que se usa cada tipo de participantes en los experimentos de IS, y la cantidad de participantes que se usan típicamente para cada uno de los tipos.
- *RQ₂*: ¿Cómo se caracterizan los participantes en los experimentos?
Las diferencias y similitudes entre los participantes de un experimento se pueden reconocer de dos maneras diferentes:
 - Indicando el tipo de participante. Típicamente, estudiantes/profesionales, pero pueden ser más específicos. En el caso de los estudiantes, podrían dividirse aún más en licenciatura/máster/doctorado, etc.
 - Enumerando un conjunto de características que son relevantes para el experimento. Por ejemplo, experiencia en un lenguaje de programación dado, conocimiento de una herramienta dada, etc.

Interesa conocer cuáles son las características utilizadas en los experimentos de SE para describir participantes.

- *RQ₃*: ¿Cómo se incorporan las características/tipos de participantes en el experimento?
Dependiendo de cómo se incorporen las características/tipos de participantes en el análisis de los datos experimentales, la evaluación del impacto del uso de diferentes participantes será diferente.
- *RQ₄*: ¿Cuál es el impacto del uso de características/tipos de participantes en los resultados experimentales?
Este es el objetivo principal de este capítulo. Tiene la intención de responder a la pregunta de si los resultados (in)consistentes entre los diferentes tipos de participantes se deben a las características del participante.

Adicionalmente, se van a estudiar los resultados de la familia de TDD del capítulo anterior.

7.2.1 Proceso de Búsqueda

Para identificar todos los artículos que reportan experimentos realizados con personas, se realizó una búsqueda manual de todos los artículos publicados en las principales revistas y conferencias de IS. Se utilizó una búsqueda manual para asegurar que no se perdía ningún experimento. (Dieste, et al., 2009) encontró que los autores usan diferentes términos para referirse a los experimentos en SE. Además, se quería examinar no solo los artículos que informaban sobre experimentos de IS, sino también el creciente número de artículos que utilizan experimentos de IS para evaluar propuestas teóricas. Como su contribución clave no es el experimento en sí, es posible que este tipo de documentos no contenga el término de búsqueda o los términos que identifican los experimentos.

Debido a que la búsqueda fue manual, se limitó su alcance, examinando el período de enero de 2011 a junio de 2016 de las siguientes seis revistas y congresos: IEEE Transactions on Software Engineering (TSE), Empirical Software Engineering Journal (EMSE), ACM Transactions on Software Engineering and Methodology (TOSEM), International Conference on Software Engineering (ICSE),

European Software Engineering Conference/Foundations on Software Engineering (ESEC/FSE), y International Symposium on Empirical Software Engineering and Measurement (ESEM)⁵.

Aunque cualquier resultado para los documentos identificados en las fuentes elegidas puede considerarse representativo de la tendencia actual en la experimentación, esta conjetura es una declaración radical que se va a evitar. Por lo tanto, los resultados obtenidos reflejan el estado de este conjunto de fuentes, que pueden no ser absolutamente consistentes con el estado real de la cuestión.

7.2.2 Proceso de Selección

Se examinó un total de 1646 artículos publicados que reportan (o incluyen como validación) experimentos que satisfacen los siguientes criterios de inclusión/exclusión:

- Experimentos controlados o cuasi experimentos.* Se excluyeron otros estudios empíricos, como estudios exploratorios, estudios de casos, encuestas, etc.
- Se comparen al menos dos tratamientos.* Se excluyeron los cuasi experimentos que exploran solo un tratamiento sin un grupo de control.
- Los participantes se utilizan para aplicar tratamientos.* Se excluyeron los experimentos en los que las unidades experimentales no son humanos o se compara el rendimiento de participantes humanos con una herramienta.
- La asignación del tratamiento a los participantes es aleatoria⁶.* Se excluyeron los cuasi experimentos en los que los participantes no pueden asignarse aleatoriamente a los tratamientos.

Se identificaron 159 artículos que reportan experimentos con participantes. La Tabla 7.1 muestra el número de artículos identificados por fuente y año, con el número total de artículos publicados en cada fuente y año entre paréntesis.

Año Fuente	2011	2012	2013	2014	2015	2016	TOTAL
TSE	6 (48)	8 (82)	4 (94)	5 (63)	4 (62)	5 (30)	32 (379)
EMSE	5 (28)	3 (24)	3 (32)	10 (57)	11 (52)	6 (40)	38 (233)
TOSEM	1 (18)	1 (18)	4 (35)	5 (37)	4 (22)	2 (8)	17 (138)
ICSE	2 (62)	10 (87)	3 (85)	14 (99)	8 (84)	8 (101)	45 (518)
ESEM	8 (33)	3 (24)	3 (24)	2 (23)	4 (20)		20 (124)
ESEC/FSE	0 (34)	1 (34)	1 (51)	2 (61)	3 (74)		7 (254)
TOTAL	22 (223)	26 (269)	18 (321)	38 (340)	34 (314)	21 (179)	159 (1646)

Tabla 7.1 Artículos que han reportado experimentos con participantes

En términos relativos, EMSE y ESEM son las fuentes donde se han encontrado la mayoría de los experimentos con participantes (16%), seguidos de TOSEM (12%). ICSE y TSE tienen aproximadamente

⁵ La búsqueda en 2016 incluye 6 números de TSE, 3 números de EMSE, 1 número de TOSEM e ICSE. Tenga en cuenta que en el momento en que se escribió este documento, los documentos de ESEM y FSE 2016 aún no estaban disponibles.

⁶ En el caso de diseños de medidas repetidas, los sujetos aplican todos los tratamientos o, en el caso de diseños cruzados, los sujetos se asignan aleatoriamente a secuencias de tratamientos

el mismo número de artículos que informan experimentos con participantes (9% y 8% respectivamente). Solo el 3% de los artículos que aparecen en ESEC/FSE reportan experimentos con participantes.

Con respecto a la evolución en el tiempo, con la excepción de 2013 (solo el 6% de los artículos publicados informaron experimentos con participantes), parece haber una ligera tendencia al alza: 10% en 2011 y 2012, 11% en 2014 y 2015 y 12% en 2016.

7.2.3 Extracción de Datos

Los datos se extrajeron como parte de un proceso iterativo de tres rondas, tal y como se indica a continuación.

Primero, se recogieron datos sobre el tipo de participante utilizado en cada experimento. Se analizaron los resultados de los 159 artículos examinados y se seleccionó los informes de los artículos: (1) experimentos con ambos tipos de participantes (profesionales y estudiantes), o (2) réplicas de un solo experimento, donde las diferentes réplicas representan diferentes tipos de participantes. Además, se eliminaron los documentos que informaban el mismo experimento (solo se encontró un duplicado)⁷.

De los 41 documentos restantes (en adelante, estudios primarios), se recopilaron datos sobre las características de los sujetos y cómo se incorporaron al experimento. Se obtuvieron 22 estudios primarios que proporcionan información sobre la influencia del tipo de participantes (o cualquiera de las características) en la variable de respuesta en estudio. Esto incluye estudios primarios que: (1) incluyen análisis separados para diferentes categorías de participantes o para cualquiera de las características del participante, o (2) incorporan el tipo o cualquiera de las características de los participantes en el análisis, incluso si se analizan todos los participantes juntos.

Finalmente, de los 22 estudios primarios seleccionados, se reunió información para estudiar el impacto del participante en los resultados.

7.2.4 Análisis

La respuesta a RQ₁, RQ₂ y RQ₃ se basa en los 41 estudios primarios, de los 159 trabajos seleccionados, que utilizan estudiantes y profesionales como sujetos experimentales. Mientras que la respuesta a RQ₄ se basa en los 22 estudios primarios finales.

7.3 RQ₁: Escala de Uso de Estudiantes y Profesionales

De los 159 trabajos seleccionados, 96 usan solo estudiantes, 22 usan solo profesionales y 41 usan ambos. Esto significa que 137 documentos utilizan estudiantes y 63 profesionales. Los números encontrados para los experimentos realizados solo con estudiantes son consistentes con los reportados por (Höfer & Tichy, 2007). Sin embargo, los números de experimentos realizados solo con profesionales o ambos tipos de participantes se intercambian con respecto a los dados por (Höfer & Tichy, 2007).

La Tabla 7.2 muestra el número y el porcentaje de trabajos por categoría de participantes de los 41 estudios primarios. Se puede observar que los estudiantes de pregrado, máster y doctorado se utilizan

⁷ No se ha tenido en cuenta los experimentos en los que los resultados para diferentes tipos de participantes se informan en documentos separados.

igualmente para representar a la población de estudiantes. Sin embargo, los profesionales de la industria generalmente se utilizan para representar a la población de profesionales en lugar de investigadores. Esto tiene sentido, ya que todos los estudios primarios se centran en temas relevantes para los profesionales de la industria.

		# Artículos	% Artículos
Estudiantes	Pregrado	23	56%
	Máster	18	44%
	PhD	16	39%
	Posgrado	18	44%
	No especificado	1	2%
Profesionales	Profesionales de la industria	34	83%
	Investigadores	12	29%
No especificado		1	2%

Tabla 7.2 Número de artículos y experimentos por categoría de participantes

La Tabla 7.3 resume los resultados con respecto al número de participantes utilizados en los experimentos por categoría. Como se esperaba, el número medio de estudiantes utilizados (24,5) es mayor que el número de profesionales (15,3). Lo mismo ocurre con el número medio de profesionales de la industria utilizados versus investigadores (17,2 y 3,1 respectivamente). Sin embargo, se utiliza un número menor de estudiantes de pregrado en comparación con los otros tipos de estudiantes. Aunque el número medio de estudiantes de pregrado es mayor que el resto de los participantes se puede observar que la mediana es mucho más baja que la media. Esto se debe a que hay un experimento con un alto número de estudiantes de pregrado (161)⁸.

		Media	Des. Típica	Min.	Max.	Mediana
Estudiantes	Pregrado	21,0	34,2	1	161	8,5
	Máster	17,3	14,4	1	52	18,0
	PhD	12,4	8	1	32	11,5
	Posgrado	12,3	9,2	2	44	9,0
	No especificado	24,5	16	14	48	18
	Total	24,5	24,1	2	167	19
Profesionales	Profesionales de la industria	17,2	20,5	1	96	10
	Investigadores	3,1	3,7	1	15	2
	Total	15,3	19,3	1	96	9
No especificado		36	N/A	36	36	36
Total		30,4	29,3	1	169	21

Tabla 7.3 Número de participantes usados en experimentos

7.4 RQ₂: Caracterización de los Participantes

Se ha encontrado que los participantes se caracterizan de acuerdo con tres tipos diferentes de información: información general común a la muestra, como el origen de los sujetos y las características de la muestra, e información particular sobre cada sujeto que se recopila típicamente mediante un cuestionario.

⁸ Si se eliminan estos datos, la media cae hasta 14,9.

El origen de los sujetos se menciona en todos los estudios primarios excepto en 7 ([S13], [S32], [S17], [S8], [S16], [S40] y [S35]). Además, 6 estudios primarios no mencionan el origen de uno de los tipos de participantes: [S24] para estudiantes y [S37], [S3], [S6], [S36] y [S38] para profesionales. Es muy común que se omita el nombre de la empresa donde trabajan los profesionales (en algunos casos los autores explican que no pueden divulgar la información), lo que significa que para la mayoría de las empresas se proporciona información general como "PYME en el área de XX".

Las *características de la muestra* encapsulan características comunes a todos los participantes (o tipos de participantes) que permiten a los investigadores explicar si se cumple un conjunto de requisitos mínimos (necesarios o recomendados). En algunos casos, estas características de muestra se dan para un tipo de participantes, ya que los otros tipos están garantizados por cómo han sido elegidos. Por ejemplo, [S15] establece que "se requiere experiencia en proyectos de seguridad en la industria" para los estudiantes involucrados en el experimento, ya que todos los profesionales de la empresa elegida tienen la garantía de tenerla.

La *información particular* sobre los participantes que se ha recopilado en el conjunto de estudios primarios se puede clasificar en: demográfica (edad, género, títulos obtenidos, etc.), experiencia (ya sea en la tecnología en estudio o relacionada), habilidades (no exactamente igual que el anterior) y preferencias ((dis)gustos, o intereses). La Figura 7.1 muestra el tipo de información recopilada por los estudios primarios. Es interesante ver que la mayoría de los estudios (27) recopilan información sobre la experiencia, seguida de habilidades y demografía (12) y, por último, solo unos pocos recopilan información sobre las preferencias de los sujetos (2). Además, se ha descubierto que los experimentos tienden a recopilar más información que la estrictamente utilizada para fines de análisis. En algunos casos, esta información se utiliza para describir o caracterizar la muestra utilizada, pero no siempre. En otros, simplemente no se usa (o al menos no se informa).

Figura 7.1 Caracterización de los participantes

7.5 RQ₃: Incorporación del Tipo de Participante / Características en el Experimento

El análisis del impacto del tipo/características de participantes en los resultados de los experimentos será diferente dependiendo de cómo se haya incorporado al análisis del experimento. Las columnas de la Tabla 7.4 muestran los diferentes tipos de análisis que se han encontrado en los estudios primarios (tenga en cuenta que algunos estudios primarios realizan más de un tipo de análisis).

La columna 1 en la Tabla 7.4 muestra que en aproximadamente la mitad de los estudios primarios (17 de 41) los datos correspondientes a todos los sujetos se analizan juntos, sin hacer distinción entre tipos/características. En estos casos, **no es posible estudiar el impacto del tipo/características de los participantes** (aproximadamente la mitad de los estudios primarios).

Como se muestra en la columna 5, hay tres estudios primarios adicionales para los cuales no se puede evaluar el impacto del tipo/características del participante. En este caso, no realizan análisis estadísticos ([S17] y en el experimento II de [S18]) o el tratamiento se confunde con el tipo de sujeto ([S19]).

Todos los sujetos juntos			Análisis separado por tipo de participantes	Otros
Sin distinción	Tipo de participante	Características		
[S25]	[S10]	[S1]	[S1]	[S17] [S18] (II) [S19]
[S14]	[S37]	[S7]	[S3]	
[S32]	[S6]	[S10]	[S15]	
[S41]	[S2]	[S13]	[S18] (I)	
[S5]	[S31]	[S26]	[S33]	
[S29]	[S4]	[S33]	[S6]	
[S8]	[S39]	[S37]	[S16]	
[S30]	[S34]	[S4]	[S34]	
[S36]		[S12]		
[S11]		[S22]		
[S40]		[S38]		
[S35]		[S28]		
[S20]		[S21]		
[S9]				
[S23]				
[S27]				
[S24]				

Tabla 7.4 Incorporación de tipo/características de participantes al análisis en el experimento

Para los estudios primarios enumerados en las columnas 2-4 de la Tabla 7.4, es posible estudiar el impacto del tipo/características del participante, se han identificado tres situaciones diferentes, correspondientes a tres formas diferentes de analizar el impacto:

- Los experimentos ejecutan análisis separados para diferentes tipos de participantes (estudios primarios correspondientes a la columna 4 de la Tabla 7.4). Por lo general, los tipos utilizados son estudiantes versus profesionales, pero pueden ser más detallado (como en el caso de [S34] que diferencia entre estudiantes de pregrado, máster, PhD y profesionales). En esta situación, el impacto se evaluará examinando la consistencia de los resultados entre diferentes tipos de participantes y relacionando las posibles (in)consistencias con las características de los participantes.

- Los datos correspondientes a todos los sujetos se analizan juntos, incorporando el tipo de participante como factor en el análisis del experimento (estudios primarios que aparecen en la columna 2 de la Tabla 7.4). En esta situación, el impacto se evaluará examinando si el tipo de participante o su interacción con los tratamientos son significativos y relacionando la significancia con las características de los participantes.
- Los datos correspondientes a todos los participantes se analizan juntos, incorporando las características del sujeto como factor en el análisis del experimento (estudios primarios enumerados en la columna 3 de la Tabla 7.4). En esta situación, el impacto se evaluará examinando directamente las características significantes y relacionándolas con los tipos de participantes.

La siguiente sección analiza el impacto del tipo de sujeto en los resultados experimentales, de acuerdo con las tres situaciones que se han identificado.

7.6 RQ₄: ¿Cuál es el impacto del uso de características / tipos de participantes en los resultados experimentales?

7.6.1 Impacto del Tipo de Participantes al Analizarlos por Separado

La Tabla 7.5 y la Tabla 7.6 muestran el impacto de los tipos de participantes en los resultados experimentales. La columna 1 enumera los estudios primarios. Las columnas 2 a 5 describen los elementos clave de los experimentos: actividad de desarrollo abordada por cada experimento, tecnología en estudio, tratamientos explorados y variables de respuesta examinadas. La columna 6 hace referencia al experimento (recuerde que muchos de los artículos describen más de un experimento). Las columnas 7 a 10 presentan los resultados (dirección del efecto) de cada experimento para los diferentes tipos de participantes utilizados (estudiantes de pregrado, estudiantes graduados, investigadores y profesionales). La columna 11 indica el tamaño del efecto. La última columna muestra si el tipo de participante influye o no en los resultados del experimento. Las celdas que aparecen en color gris denotan las variables de respuesta sobre las cuales influye el tipo de sujeto.

En 3 de los 8 experimentos examinados ([S3], [S18] y [S33]), el tipo de participante no influye en los resultados. En los 5 experimentos restantes, el tipo de sujeto influye en los resultados en al menos una variable de respuesta. Específicamente:

- Para la variable respuesta claridad en [S15], comprensión en [S34] y corrección en [S6], los tratamientos se comportan de manera diferente para profesionales y estudiantes. Hay una **diferencia en la dirección del efecto**.
- Para la variable de respuesta comprensión en [S1] y precisión en [S16], la dirección del efecto es la misma tanto para profesionales como para estudiantes, pero hay una **diferencia en el tamaño del efecto**.

Es interesante notar que no se ha encontrado influencia del tipo de participantes para las variables de eficiencia (tiempo) de [S34], [S6] y [S16].

Se van a estudiar con más detalle las características de cada tipo de participante en los 8 experimentos para ver si se puede encontrar una explicación para los resultados.

[S3] muestra resultados consistentes para diferentes tipos de participantes. Los estudiantes tienen habilidades en el desarrollo web que van desde principiantes hasta profesionales, y profesionales de nivel medio a alto. Sin embargo, según los autores, la capacitación para cada experimento se personalizó según las necesidades de los sujetos experimentales. Además, los objetos experimentales utilizados con los estudiantes son diferentes de los utilizados con los profesionales (más complejos). Por lo tanto, la coincidencia en los resultados podría deberse a la *configuración experimental ajustada a cada tipo de participante*.

En [S18] se encuentran resultados consistentes para diferentes tipos de participantes. Los estudiantes de posgrado tienen experiencia en programación Java de 2 a 5 años, y los desarrolladores fueron seleccionados de comunidades de desarrolladores en línea y compañías de software con el único requisito de tener experiencia en Java. Según los autores, "algunos participantes pueden no estar completamente cualificados", ya que no podían "verificar las cualificaciones de algunos desarrolladores". Por lo tanto, la coincidencia en los resultados podría deberse a que *ambos tipos de participantes representan la misma población*.

En [S33], los resultados para profesionales y estudiantes son consistentes, aunque en uno de los experimentos, los estudiantes de máster muestran un resultado diferente al de otros estudiantes de máster. Según los autores, esto podría deberse al hecho de que estos estudiantes de maestría tienen menos experiencia en modelado. Por lo tanto, la coincidencia en los resultados podría deberse nuevamente a *ambos tipos de participantes que representan la misma población*.

En [S15], solo en una de las variables de respuesta (claridad) se encuentra una diferencia entre profesionales y estudiantes. Según los autores, la coincidencia en los resultados podría deberse al hecho de que "la autoevaluación de ambas muestras no discrimina entre estudiantes y profesionales" y, por lo tanto, *ambos tipos de participantes representan la misma población*. Los autores no consideran que la diferencia en los resultados para la variable respuesta claridad sea importante, ya que es una variable cualitativa (refleja opiniones).

En [S34], hay resultados inconsistentes entre estudiantes de pregrado y maestría, y estudiantes de doctorado y profesionales para la comprensión. En línea con las conclusiones extraídas por los autores, estos resultados podrían deberse a *diferentes tipos de participantes que tienen diferentes características* (conocimiento de diseño de alto y bajo nivel de OO, y desarrollo y mantenimiento de software).

En [S6], hay resultados inconsistentes en la corrección para diferentes tipos de participantes. Hay diferencias en los tipos de participantes con respecto a los años de programación en Java y los años que llevan usando Junit. Por lo tanto, siguiendo el razonamiento de los autores, las diferencias en los resultados se deben a *diferentes tipos de participantes que tienen diferentes características*.

[S1] muestra resultados inconsistentes para diferentes tipos de sujetos. Los autores afirman que la experiencia de los participantes en los diagramas UML podría explicar las diferencias en el tamaño del efecto encontrado. En este estudio, los participantes en el experimento II serían los más experimentados, seguidos por los de los experimentos IV y V, luego por el experimento II y finalmente por el experimento I. Por lo tanto, las diferencias en los resultados podrían explicarse por *diferentes tipos de participantes que tienen diferentes características*.

En [S16], los estudiantes superan a los profesionales. Nuevamente, la diferencia en los resultados puede considerarse como *diferentes tipos de participantes que tienen diferentes características*. Según los autores, "los estudiantes estaban más actualizados con el tema de los casos de uso".

Estudio	Actividad Desarrollada	Tecnología bajo estudio	Tratamientos	Variable Respuesta	Exp.	Dirección del efecto					Tamaño del efecto	Influencia del tipo de participante
						Estudiantes				Profesionales		
						Pregrado	Máster	Posgrado	PhD	Industria		
[S15]	Verificación & Validación	Métodos para el análisis de seguridad	Árboles de fallas (FT) Árboles de fallas integrados por componentes (CFT)	Precisión Consistencia Mantenibilidad	I	-	FT=CFT	-	-	-	-	No
					II	-	-	-	-	FT=CFT	-	
				Claridad	I	-	FT>CFT	-	-	-	-	Si
					II	-	-	-	-	FT=CFT	-	
[S3]	Comprensión de programas	Técnica para comprender el comportamiento de programas de aplicaciones web	CLEMATIS Otro	Tiempo	I	CLEMATIS<Otro		-	9	-	-	No
					II	-		-	-	CLEMATIS<Otro	-	
				Exactitud	I	CLEMATIS>Otro		-	9	-	-	No
					II	-		-	-	CLEMATIS>Otro	-	
[S18]	Depuración	Enfoque de corrección de errores automático	PAR GenProg Parches humanos	Aceptabilidad	I	-	-	PAR>GenProg PAR=Humano	-	PAR>GenProg PAR=Humano	-	No
[S1]	Requisitos	Diagramas de Secuencia UML	Con UML SD Sin UML SD	Comprensión	I	SD=NSD	-	-	-	-	Despreciable	Si
					II	-	SD>NSD	-	-	-	Grande	
					III	-	SD>NSD	-	-	-	Pequeño	
					IV	-	SD>NSD	-	-	-	Mediano	
					V	-	-	-	-	SD>NSD	Mediano	
[S33]	Requisitos	Análisis de Modelos UML	Con análisis de modelos UML	Comprensibilidad	I, II	-	UML=NUML	-	-	-	-	No
					III	-	UML> NUML	-	-	-	-	
					IV	-	-	-	-	UML=NUML	-	
			Sin análisis de modelos UML	Modificabilidad	I, II, III	-	UML=NUML	-	-	-	-	No
					IV	-	-	-	-	UML=NUML	-	

Tabla 7.5 Resumen: el tipo de participante se analiza por separado (1/2)

⁹ Hay un investigador en el grupo de estudiantes, pero esto no es apreciable en el análisis y el efecto parece haber sido diluido por el resto del grupo.

Estudio	Actividad Desarrollada	Tecnología bajo estudio	Tratamientos	Variable Respuesta	Exp.	Dirección del efecto					Tamaño del efecto	Influencia del tipo de participante
						Estudiantes				Profesionales		
						Pregrado	Máster	Posgrado	PhD	Industria		
[S16]	Requisitos	Métodos de especificación de eventos	Ad-hoc H4U	Exactitud	I	Ad-hoc<H4U	-	-	-	-	Grande	Si
					II	-	-	-	-	Ad-hoc<H4U	Mediano	
				Velocidad de revisión	I	Ad-hoc≠H4U	-	-	-	-	Grande	No
					II	-	-	-	-	Ad-hoc≠H4U	Grande	
[S34]	Comprensión de programas	Documentación para comprender patrones de diseño codificado	Indocumentado	Tiempo	I	-	-	-	-	SC=GD=TD	-	No
					II	-	SC=GD=TD	-	-	-	-	
					III	SC=GD=TD	-	-	-	-	-	
					IV	-	-	-	SC=GD=TD	-	-	
			Diagramas de clase UML	Comprensión	I	-	-	-	-	(TD=GD) > SC	Grande	Si
					II	-	SC=GD=TD	-	-	-	-	
					III	SC=GD=TD	-	-	-	-	-	
					IV	-	-	-	(TD=GD)>SC	-	Grande	
[S6]	Comprensión y mantenimiento de programas	Prueba de hediondez de código	Prueba de hediondez	Precisión	I	-	NTS<TS	-	-	-	Mediano-Grande	Si
					II	NTS<TS	-	-	-	-	Mediano-Grande	
					III	NTS<TS	-	-	-	-	Mediano-Grande	
					IV	-	-	-	-	NTS=TS	-	
			Sin prueba de hediondez	Tiempo	I	-	NTS=TS	-	-	-	-	No
					II	NTS=TS	-	-	-	-	-	
					III	NTS=TS	-	-	-	-	-	
					IV	-	-	-	-	NTS=TS	-	

Tabla 7.6 Resumen: el tipo de participante se analiza por separado (2/2)

7.6.2 Impacto del Tipo de Participantes al Analizarlos Conjuntamente

La Tabla 7.7 y la Tabla 7.8 muestran el impacto de los tipos de participantes en los resultados experimentales. Las primeras cinco columnas son las mismas que en la Tabla 7.5. Las columnas 7-12 indican los tipos de participantes utilizados en los experimentos. Las columnas 13 y 14 muestran si el tipo de participantes influye o no en los resultados del experimento (efecto principal e interacción con los tratamientos, respectivamente). Las celdas que aparece en color gris denotan las variables respuesta sobre las cuales influye el tipo de participante.

Se ha encontrado que en 3 de los 8 experimentos examinados ([S10], [S2] y [S4]), el tipo de participante no influye en los resultados. Tenga en cuenta que en el caso de [S4] no se ha explorado la interacción entre el tipo de participante y los tratamientos. En los 5 experimentos restantes, el tipo de participante influye en los resultados en al menos una variable de respuesta. Específicamente:

- En [S31] se encuentran **resultados contradictorios**. Existe una interacción entre el tipo de participante y los tratamientos para 5 de las 20 métricas medidas. En 11 de las 15 métricas restantes hay influencia del tipo de participante. Para las otras 4, no se observa influencia.
- Para las variables respuesta de corrección y tiempo en [S6] y comprensión y tiempo en [S34], hay una **influencia del tipo de participante, pero la interacción no es significativa**.
- Para las variables respuesta de corrección de parches y tiempo de depuración en [S37], y tareas completadas y tiempo en [S39] hay una **influencia del tipo de participante, pero la interacción no ha sido explorada**.

Se estudia con más detalle las características de cada tipo de participante en los experimentos para ver si se puede encontrar una explicación para los resultados. Se analizará [S10], [S37] y [S4] en la Sección 7.6.3, ya que estos estudios primarios incluyen el tipo de participante como una de varias características del participante exploradas en el análisis.

Los estudios primarios [S6] y [S34] se discutieron en la Sección 7.6.1. La conclusión fue que las diferencias en los resultados se deben a *diferentes tipos de participantes que tienen diferentes características*. Sin embargo, es interesante notar que el análisis realizado en la Sección 7.6.1 no mostró diferencias entre estudiantes y profesionales para la variable respuesta de tiempo, mientras que el análisis realizado en esta sección si las ha identificado. Una posible explicación para esto es que cuando los datos se agrupan, el tamaño de la muestra aumenta, junto con el poder de las pruebas estadísticas.

En [S2], la única información proporcionada sobre los participantes es que los profesionales tienen experiencia en las pruebas, y los estudiantes son un 65% superiores al seguir un curso de maestría sobre verificación y validación. Los autores atribuyen la falta de influencia del tipo de participante en los resultados a "una cantidad similar de conocimiento de dominio" (el objeto experimental es un editor de texto genérico). *No se puede llegar a ninguna conclusión*; ya que no se describen más los participantes.

En [S31], los participantes se describen de acuerdo con la experiencia en programación, la experiencia en programación en java, la experiencia en pruebas unitarias, la experiencia en JUnit y TDD. Los autores atribuyen los resultados a una mayor experiencia en programación y pruebas unitarias de los profesionales. Esto significa que *diferentes tipos de participantes tienen características diferentes*.

En [S39], el tipo de participante influye en el uso de expresiones lambda versus iteradores en términos de tareas completadas y tiempo. Los diferentes tipos de participantes tienen diferentes perfiles.

Mientras que los estudiantes de primer año y segundo año no conocen ni iteradores ni lambdas, los estudiantes de tercer año y senior si los conocen a ambos. Finalmente, los profesionales son desarrolladores altamente experimentados. Por lo tanto, la diferencia en los resultados puede atribuirse a *diferentes tipos de participantes que tienen características diferentes*.

La última fila de la Tabla 7.8 muestra los resultados para la familia de experimentos sobre TDD. Este estudio reporta 16 experimentos con profesionales y estudiantes de varios tipos, que finalmente se han categorizado simplemente como estudiantes y profesionales. De acuerdo con los resultados mostrados en la sección 6.2.5.3, la interacción del enfoque de desarrollo*tipo de sujeto es significativa. Concretamente, en la Figura 7.2 y la Figura 7.3 se puede observar que los profesionales aplican peor ITL que TDD, por el contrario, para los estudiantes no existen diferencias entre ambos tratamientos. Sin embargo, el tamaño del efecto es pequeño por lo que se puede concluir que la interacción no influye en los resultados. En la siguiente sección se va a estudiar que ocurre con la experiencia en programación (una de las características de los participantes) que también es significativa.

Figura 7.2 Interacción Enfoque de Desarrollo*Tipo de Sujeto

Figura 7.3 Interacción Tipo de Sujeto*Enfoque de Desarrollo

Estudio	Actividad Desarrollada	Tecnología bajo estudio	Tratamientos	Variable Respuesta	Exp.	Tipo de participante utilizado						Influencia del tipo de participante	
						Estudiantes				Profesionales		Efecto principal	Interacción
						Pregrado	Máster	Posgrado	PhD	Inv.	Ind.		
[S10]	Comprensión de programas	Impacto de la información contextual en cómo dividir / expandir identificadores	Información interna	F-medida	I	X	X	-	X	X	-	No	No
			Información externa		II	X	X	-	X	X	-		
[S37]	Depuración	Corrección automática de errores como ayuda a la depuración	Parches de alta calidad	Corrección de parche	I	-	-	X	-	-	-	Si	?
			Parches de baja calidad		II	-	-	-	-	-	X		
			Ubicaciones de errores		III	X ¹⁰	-	-	-	-	X		
				Tiempo de depuración	I	-	-	X	-	-	-	Si (II)	?
					II	-	-	-	-	-	X		
					III	X ¹⁰	-	-	-	-	X		
[S6]	Comprensión y mantenimiento de programas	Prueba de hediondez de código	Prueba de hediondez	Precisión	I	-	X	-	-	-	-	Si	No
					II	X	-	-	-	-	-		
					III	X	-	-	-	-	-		
					IV	-	-	-	-	-	X		
			Sin prueba de hediondez	Tiempo	I	-	X	-	-	-	-	Si	No
					II	X	-	-	-	-	-		
					III	X	-	-	-	-	-		
					IV	-	-	-	-	-	X		
[S2]	Pruebas	Técnicas de prueba	Prueba explanatoria Prueba basada en casos de prueba	Número de defecto Falsos defectos reportados	I	-	X	-	-	-	X	No	No

Tabla 7.7 Resumen: el tipo de participante se analiza conjuntamente (1/2)

¹⁰ Hay 2 estudiantes de pregrado, pero es inapreciable en el análisis y el efecto parece haber sido diluido por el resto del grupo. No se utiliza ninguna prueba estadística para evaluar qué grupos se comportan de manera diferente o si existe una interacción. Esto se verifica solo visualmente.

Estudio	Actividad Desarrollada	Tecnología bajo estudio	Tratamientos	Variable Respuesta	Exp.	Tipo de participante utilizado						Influencia del tipo de participante	
						Estudiantes				Profesionales		Efecto principal	Interacción
						Pregrado	Máster	Posgrado	PhD	Inv.	Ind.		
[S31]	Codificación	Test-driven development	Test-Driven Development Test Last Development	Calidad interna ¹¹	I	-	X	-	-	-	-	Si (11) No (4)	Si (5) No (15)
					II	-	-	-	-	-	X		
[S4]	Codificación	Opiniones sobre antipatrones lingüísticos	LAs No LAs	Percepción de código	I			X			X	No	?
[S39]	Codificación	Uso de lambdas	Expresiones Lambda Iteradores	Tareas completadas	I	X	-	-	-	-	X	Si	?
				Tiempo	I	X	-	-	-	-	X		
				Errores del compilador	I	X	-	-	-	-	X	No	?
				Tiempo trabajando en errores	I	X	-	-	-	-	X		
[S34]	Comprensión de programas	Documentación para comprender patrones de diseño codificado	Indocumentado Diagramas de clase UML Comentarios en el código fuente	Comprensión Tiempo	I	-	-	-	-	-	X	Si	No
					II	-	X	-	-	-	-		
					III	X	-	-	-	-	-		
					IV	-	-	-	X	-	-		
TDD	Codificación	Desarrollo de código	TDD ITL	Calidad	16	X	X	X	-	-	X	No	No

Tabla 7.8 Resumen: el tipo de participante se analiza conjuntamente (2/2)

¹¹ Se utilizan 20 métricas diferentes.

7.6.3 Impacto de las Características de los Participantes en los Resultados

La Tabla 7.9, la Tabla 7.10 y la Tabla 7.11 muestran el impacto de las características del participante en los resultados experimentales. Las primeras cinco columnas son las mismas que en la Tabla 7.5. La columna 6 muestra el número de experimentos reportados en cada estudio primario. Las columnas 7 a 12 recuerdan el tipo de participantes utilizados en el experimento (estudiantes de pregrado, estudiantes graduados, investigadores y profesionales). Las columnas 13 y 14 enumeran los resultados de los experimentos con respecto al impacto de las características del participante en las variables respuesta. La columna 13 indica que características del participante influyen, mientras que la columna 14 especifica que características del participante no influyen. El experimento donde ocurre aparece entre paréntesis. Si ocurre en todos los experimentos, no se muestra nada. Las características que aparecen en **negrita** y subrayadas corresponden al tipo de participante.

En la Tabla 7.9, la Tabla 7.10 y la Tabla 7.11 se pueden ver que en 5 estudios ([S26], [S4], [S38], [S28] y [S21]) ninguna característica influye en los resultados, lo que significa que los resultados del experimento se mantienen independientemente de los valores de la característica. Para los 8 estudios restantes, una o más características influyen en los resultados. Tenga en cuenta que las características se refieren a la experiencia y el conocimiento relacionados con la tecnología que se estudia.

Como ya se ha mencionado, para evaluar si el impacto en los resultados se debe al tipo de sujeto o las características del sujeto, es necesario identificar las características que tienen los diferentes tipos de participante. Desafortunadamente, esto no es posible para todos los estudios:

- Algunos estudios no proporcionan los valores para las características estudiadas dentro de cada tipo de participante. Esto ocurre en [S7], [S10], [S13], [S21], [S37] y [S12]. Esto hace imposible evaluar en qué medida las características están relacionadas con los tipos de participantes y, por lo tanto, *no se puede llegar a ninguna conclusión*.
- En dos estudios ([S1] y [S33]), las características están anidadas dentro de los tipos de participantes y *no tiene sentido relacionarlas con los tipos de participantes*.

Para el resto de los estudios, es posible evaluar el impacto en los resultados, ya que:

- Proporcionan los valores para las características dentro de cada tipo de participante. Esto ocurre en [S4], [S22] y [S38].
- Aunque no proporcionan los valores para las características dentro de cada tipo de participante, dan una pista que permite adivinarlos. Esto ocurre en [S26] y [S28]. Por ejemplo, en [S26], se utilizan 44 profesionales y 5 estudiantes. La característica estudiada es la experiencia en el lenguaje de programación. Los autores establecen dos grupos: < 2 años y ≥ 2 años. Además, mencionan que 33 reportan experiencia ≥ 2 años. Esto significa que hay profesionales que tienen las mismas características que los estudiantes y, por lo tanto, es la característica y no el tipo de participante lo que influye en los resultados.

Es importante tener en cuenta que cuando las características no son significativas, no tiene sentido relacionar sus valores con los tipos de participantes, por lo tanto, se estudiarán las dos últimas situaciones, en los estudios para los cuales cualquier característica es significativa.

El único estudio que cumple ambas condiciones es [S22]. Para una de las variables respuesta examinadas, la precisión, la experiencia de desarrollo es significativa. Este estudio reporta un experimento con estudiantes de máster y profesionales, donde la experiencia de desarrollo se mide en una escala ordinal que cubre los siguientes períodos: 1-3 años de experiencia, 4-6 años de experiencia, 7-9 años de experiencia y > 9 años de experiencia. Según la información proporcionada en el artículo sobre cuántos estudiantes y profesionales pertenecen a cada grupo de experiencia, se puede ver que se pueden encontrar ambos tipos de participantes en cada categoría. Esto significa que los *tipos de participantes son heterogéneos* y no representan la variabilidad en la muestra.

Finalmente, para la familia de experimentos sobre TDD. La experiencia en programación se mide en una escala Likert/ordinal que tiene las siguientes categorías: Sin experiencia (< 2 años), Novato (2-5 años), Intermedio (5-10 años) y Experto (> 10 años). De acuerdo con la gráfica 7.2 se puede observar que en cada categoría existen ambos tipos de participantes. Por lo tanto, ambos tipos de participantes son heterogéneos y no representan variabilidad en la muestra.

Figura 7.4 Experiencia en programación según el tipo de sujeto

7.6.4 Discusión

De acuerdo con los resultados obtenidos, se puede concluir que las razones de las (in)consistencias en los resultados se deben a:

- *Ambos tipos de participantes que representan la misma población.* Todos los tipos de participantes son iguales con respecto a las características relevantes para el experimento. Esto sucede en [S15], [S18] y [S33].
- *La configuración experimental está ajustada,* de modo que las posibles diferencias entre los tipos de participantes se descartan o compensan. Este es el caso de [S3].
- *Los diferentes tipos de participantes tienen diferentes características.* En este caso, se confunden tanto el tipo de participantes como las características correspondientes, y no se

puede distinguir cuál está teniendo efecto. Este es el caso de [S34], [S6], [S1] y [S16], [S31] y [S39].

- *Los tipos de participantes son heterogéneos* y no representan la variabilidad que existe en la muestra. Este es el caso de [S22] y la familia de experimentos de TDD.
- *No se puede llegar a ninguna conclusión.* El estudio primario no proporciona suficiente información para permitir ningún razonamiento sobre sus resultados. Este es el caso de [S2], [S7], [S10], [S13], [S21], [S37] y [S12].
- *No tiene sentido relacionar las características de los participantes con el estudio de tipos de participantes*, ya que:
 - Las características están anidadas dentro de los tipos de participantes. Este es el caso de [S1] y [S33].
 - Ninguna de las características examinadas es significativa [S4], [S38], [S26] y [S28].

Estos resultados sugieren que, para este conjunto de estudios primarios, no es el tipo de participante lo que influye en la (in)consistencia de los resultados experimentales, sino las características que tienen relacionadas con la experiencia en la tecnología evaluada en el experimento.

Sin embargo, es importante tener en cuenta que los resultados solo son válidos para el conjunto de estudios primarios cubierto. Por lo tanto, si el uso de estudiantes (o diferentes tipos de participantes) afecta los resultados de SE es un tema que necesita más investigación empírica.

Mientras tanto, se recomienda a los investigadores que tengan en cuenta no solo el tipo de participante sino también las características relevantes para el tema del experimento al diseñarlo.

Estudio	Actividad Desarrollada	Tecnología bajo estudio	Tratamientos	Variable Respuesta	N. Exp.	Estudiantes				Profs		Características de los participantes	
						Preg.	Máster	Posg.	PhD	Inv.	Ind.	Influencia	Sin Influencia
[S7]	Comprensión de programas	Nomenclatura de identificadores	Notación de guión bajo Notación de Camello	Rendimiento Eficiencia	5	X	-	X	-	X	-	Edad (EIII) Experiencia (EIV) Tratamientos*Formación (EIII) Tratamientos*Edad(EIII) Tratamientos*Experiencia (EIV) Género (EIII) Tratamientos*Formación (EI)	El documento solo analiza las características (de un total de 21) que fueron significativas al menos una vez
[S13]	Reutilización del framework OO	Ayuda a comprender el framework	FUDA Documentación del Framework	Tiempo	1	X	-	X	-	-	X	Experiencia con framework objetivo	-
[S26]	Reutilización	Buscadores de códigos	Portfolio Google Code Search Koders	Confianza	2	-	-	X	-	-	X	-	Experiencia con el lenguaje de programación (EI)
[S10]	Comprensión de programas	Impacto de la información contextual en cómo dividir / expandir identificadores	Información interna Información externa	F-medida	2	X	X	-	X	X	-	Conocimiento de Linux (EI) Dominio del Inglés	Experiencia en lenguaje C Tipo de sujeto Tratamientos*Conocimiento Linux Tratamientos*Dominio Inglés Tratamientos* Experiencia en Lenguaje C Tratamientos* Tipo de Sujeto
[S1]	Requisitos	Diagramas de secuencia UML	Con UML SD Sin UML SD	Comprensión	5	X	X	-	X	-	X	Habilidad (EIII) Tratamientos *Habilidad	Habilidad (EI, EII, EIV, EV)

Tabla 7.9 Resumen: para las características de los participantes (1/3)

Estudio	Actividad Desarrollada	Tecnología bajo estudio	Tratamientos	Variable Respuesta	N. Exp.	Estudiantes				Profs.		Características de los participantes	
						Preg.	Máster	Posg.	PhD	Inv.	Ind.	Influencia	Sin Influencia
[S33]	Análisis	Modelos de Análisis UML	Con modelos UML	Comprensibilidad	4	-	X	-	-	-	X	-	Habilidad Tratamientos*Habilidad
			Sin modelos UML	Modificabilidad								Habilidad (EII)	Habilidad (EI, EIII, EIV) Tratamientos* Habilidad
[S37]	Depuración	Corrección automática de errores como ayuda a la depuración	Parches de alta calidad	Corrección de parche	3	X	-	X	-	-	X	Experiencia Java <u>Tipo de sujeto</u>	Tratamientos* Experiencia Java
			Parches de baja calidad	Tiempo de depuración								<u>Tipo de sujeto</u>	Experiencia Java Tratamientos* Experiencia Java
[S4]	Codificación	Opiniones sobre antipatrones lingüísticos	Las No Las	Percepción de código	1	-	-	X	-	-	X	-	Lenguaje de programación principal Exp. en programación <u>Ocupación</u>
[S12]	Diseño	Artefactos para comunicar decisiones de arquitectura	Diagramas Texto	Respuestas correctas	1	X	X	-	X	-	X	Distancia del idioma Preferencia de medios Habilidad de modelado Experiencia	Exclusión de medios Familiarizado con diagramas Inclusión de medios Tiempo medio por pregunta
[S22]	Mantenimiento	Trazabilidad	Con trazabilidad Sin trazabilidad	Tiempo	1	-	X	-	-	-	X	-	Experiencia de desarrollo
				Precisión								Experiencia de desarrollo	-
[S38]	Requisitos	Transparencia	Especificación de requisitos Modelo de caso de uso	Tiempo Precisión Confianza	1	X	X	-	X	-	X	-	Experiencia *Tratamientos
[S28]	Pruebas y Depuración	Generación de pruebas automatizadas	Con resúmenes Sin resúmenes	Corrección de errores	1	X	X	-	X	X	X	-	Exp. en programación Exp. en programación *tratamientos

Tabla 7.10 Resumen: para las características de los participantes (2/3)

Estudio	Actividad Desarrollada	Tecnología bajo estudio	Tratamientos	Variable Respuesta	N. Exp.	Estudiantes				Profs.		Características de los participantes	
						Preg.	Máster	Posg.	PhD	Inv.	Ind.	Influencia	Sin Influencia
[S21]	Depuración	Técnicas de depuración de concurrencia	DSP Agenda completa	Dificultad del error Tiempo de diagnóstico	1	X	X	-	X	-	X	-	Experiencia en depuración
TDD	Codificación	Desarrollo de código	TDD ITL	Calidad	16	X	X	X	-	-	X	Experiencia en Programación <u>Tratamientos*</u> <u>Tipo de sujeto</u>	Tipo de Sujeto Experiencia con el lenguaje Experiencia en pruebas unitarias Experiencia con el framework de pruebas Experiencia en Programación* Tratamientos Experiencia con el framework de pruebas * Tratamientos Experiencia con el lenguaje de programación * Tratamientos Experiencia en pruebas unitarias * Tratamientos

Tabla 7.11 Resumen: para las características de los participantes (3/3)

Capítulo 8

CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

Este capítulo recoge las conclusiones obtenidas, tras el desarrollo de esta investigación, específicamente se presentan los resultados obtenidos para cada uno de los objetivos planteados en la sección 3.2.

8.1 Conclusiones

Tal y como se comentó en el Capítulo 2 de esta tesis el problema a resolver es la carencia de un proceso para la agregación de familias de experimentos que permita el estudio sistemático de variables moderadoras. Posteriormente, derivado de este problema se identificaron los siguientes subproblemas:

1. La identificación de similitudes y diferencias entre los experimentos de una familia no se realiza de forma precisa.
2. El establecimiento del conjunto de variables moderadoras no se realiza de forma sistemática.
3. El análisis de datos no incorpora de forma conjunta todas las variables identificadas.

Asociados a la resolución de los problemas identificados se establecieron 4 objetivos cuyas conclusiones se detallan a continuación.

O₁: Creación de un modelo de caracterización de experimentos.

El modelo de caracterización surgió a través de la adaptación del esquema propuesto por Gómez (Gómez, 2012) para identificar tipos de replicaciones. Esta adaptación se realizó de manera incremental, conforme se fue evaluando una a una las 46 familias de experimentos encontradas en la literatura y las dos familias adicionales en las que se validó el proceso propuesto en esta investigación, ya que ahora su propósito es la caracterización de experimentos.

El modelo de caracterización mostrado en la Tabla 4.3 permite recabar la mayor información posible sobre un experimento, además al describir cada uno de los experimentos de la familia bajo el mismo esquema se facilita considerablemente la comparación entre los mismos. De esta manera se ha podido

comprobar que en todas las familias se realizan cambios entre los experimentos, incluso en aquellas en las que los autores, dicen realizar replicaciones exactas.

A partir del modelo de caracterización se creó una aplicación web, para almacenar familias de experimentos. Esta aplicación contiene los mismos elementos que el modelo de caracterización, pero a su vez, sirve como repositorio para almacenar tanto los ficheros de los datos obtenidos en los experimentos, como los ficheros de análisis de datos generados. Esto facilita la planificación y realización de un nuevo experimento dentro de una familia y su posterior agregación, ya que todos los datos están disponibles.

O₂: Identificación de variables moderadoras

Se han definido dos tipos de variables: de diseño y moderadoras. A su vez estas últimas se dividen en de participante y contextuales. Las variables de diseño están asociadas al diseño experimental, según se ha comprobado con las 46 familias del mapeo sistemático, en muy pocas ocasiones se cambia el diseño experimental, sin embargo, hay otros elementos del diseño que cambian con mayor frecuencia, como, por ejemplo: el número de sesiones, el número de grupos o los objetos experimentales usados en cada sesión. Por otro lado, en las familias del Capítulo 5 y 6, existe al menos un experimento cuyo diseño experimental es diferente al de los demás. Para agregar experimentos con distinto diseño se ha definido un conjunto de 6 reglas que permiten crear un diseño global o agregado.

Por otro lado, el esquema de caracterización de una situación experimental permite comparar los experimentos y, por lo tanto, se pueden identificar de manera sistemática las diferencias entre los mismos. Reconocer el impacto que pueden tener estas diferencias en los resultados, permite convertirlas a variables moderadores contextuales que se pueden estudiar estadísticamente. Sin embargo, es necesario conocer la distribución de estas variables y determinar que no estén confundidas, ya que, al momento de analizarlas en el caso de obtener peores resultados al pasar de un modelo a otro o tener problemas con el modelo mixto, se puede considerar la eliminación de estas variables al refinar el modelo final.

Finalmente, para la identificación de variables moderadoras de participante, se utiliza la información demográfica recogida durante la ejecución de cada uno de los experimentos. Esta información permite estudiar las diferencias que existen entre los participantes. Una vez identificadas las diferencias, es necesario comprobar que las variables estén medidas en las mismas escalas y que se hayan recogido en todos los experimentos. Además, es importante refinar el conjunto final de variables, en función de la posible existencia de conflictos: ya sea, que las variables estén confundidas con otras o que exista desbalanceo. También se debe comparar estas variables con las contextuales, ya que también podrían estar confundidas con estas.

O₃: Analizar los experimentos conjuntamente

El análisis estadístico propuesto es el de un modelo jerárquico que puede tener varios niveles. En la jerarquía más simple los participantes están anidados dentro de experimentos (diseños entre-sujetos). Para conseguir un modelo válido (final) se va evolucionando desde un modelo muy simple a uno más complicado a través de los siguientes 5 pasos:

1. Modelo básico sin distinguir entre experimentos
2. Añadir los niveles jerárquicos al modelo (participantes, experimentos, etc)
3. Añadir variables de cada uno de los niveles identificados (participantes (nivel 1), contextuales (nivel 2), etc.)

4. Especificar el factor como aleatorio
5. Explorar interacciones entre las variables contextuales y el factor principal y entre las variables de participante y el factor principal.

Para cada uno de estos pasos es necesario comprobar 5 tipos diferentes de matrices de covarianza. Además, cada matriz se verifica para cada una de las variables de medida repetidas estudiadas. Para finalmente, obtener un modelo cuyo AIC sea el más bajo, este será el modelo final en cada uno de los pasos.

En la familia del Capítulo 5, se tuvo que eliminar una de las variables de experimento (DOTSEM) porque daba un error con la matriz y tampoco fue posible incluir todas las interacciones. Por tanto, el modelo final obtenido se obtiene con la matriz de covarianza Diagonal y el Período como medida repetida e incluye la Técnica, Programa, Período y la interacción Técnica*Programa (variables de diseño), la experiencia absoluta y relativa con el lenguaje de programación de los participantes (variables de participante), los conocimientos de programación (variable de experimento) y las interacciones Técnica*Experiencia relativa y Técnica*Conocimiento de programación.

Por otro lado, en la familia del Capítulo 6, en el modelo final obtenido, no fue necesario eliminar ninguna de las variables que podrían presentar conflictos, sin embargo, en el último paso, se incluyeron 5 interacciones la interacción del factor principal con las variables: la experiencia en programación, el tipo de sujeto, la duración de la tarea, el entrenador y la experiencia con el framework de pruebas. Por lo tanto, el modelo final obtenido, se obtiene con la matriz de covarianza Diagonal y el enfoque de desarrollo como medida repetida. Además, estudia como variables de diseño: el Enfoque de desarrollo, el Programa, la Granularidad, el Período y las interacciones: Enfoque de desarrollo por Programa, Programa por Granularidad y Enfoque de Desarrollo por Granularidad. Como variables de participante: la experiencia en programación, con el lenguaje de programación del experimento, en pruebas unitarias y con el framework de pruebas utilizado en el experimento. Como variables de experimento: Tipo de sujeto, Lenguaje de programación en el que se realizó experimento, Transmisión del Tratamiento, el Entrenador, la combinación entre el entrenamiento y las sesiones experimentales y la duración de la tarea. El Enfoque de desarrollo como factor aleatorio. Y finalmente, las interacciones Enfoque de Desarrollo por la experiencia en programación, por el tipo de sujeto, por la duración de la tarea, por el entrenador y por la experiencia con el framework de pruebas.

Finalmente, con respecto a la consecución del objetivo general de esta investigación: *formalizar y crear un proceso de agregación de familias de experimentos, de modo que facilite y permita la inclusión de las modificaciones o alteraciones realizadas entre los experimentos de la familia*. Se puede concluir, que el proceso aquí descrito permite agregar diferentes experimentos de una familia independientemente del diseño utilizado y las diferencias entre los mismos; la única restricción es que los constructos sean similares. Además, este proceso permite estudiar si las diferencias o similitudes tienen algún impacto en la variable respuesta estudiada. Finalmente, dado que la evaluación del modelo es un proceso incremental, se garantiza que el modelo final es válido y explica correctamente las variables incluidas en el mismo.

O₄: Estudiar la adecuación de la etiqueta estudiantes y profesionales en las familias de experimentos

De 41 estudios en los que se realizan experimentos tanto con estudiantes y profesionales, se pudo estudiar la influencia del tipo de sujeto cuando se realizan análisis por separado (8 estudios primarios), cuando se analizan conjuntamente y se incorpora la variable tipo de sujeto (8 estudios primarios) y

cuando se incorporan las características de los participantes (13 estudios primarios). Cuando no existen diferencias entre ambos tipos de participantes se pudo comprobar que esto se debe a que ambos tipos de participantes representan a la misma población, o a que la configuración experimental está ajustada a cada tipo de participante, o a que los participantes son heterogéneos. Por el contrario, cuando existen diferencias se pudo comprobar que diferentes tipos de participantes tienen diferentes características. Por otro lado, existen estudios en los que no se pudo llegar a ninguna conclusión ya que no se proporciona suficiente información. Finalmente, existen estudios en los que no se puede relacionar las características con el tipo de sujeto porque las características están anidadas dentro el tipo o porque ninguna de las características fue significativa. Por el contrario, para la familia de TDD se pudo comprobar que ambos tipos de participantes representan diferentes muestras.

Por lo tanto, se puede concluir que no es el tipo de participante lo que influye en la (in)consistencia de los resultados experimentales, sino las características que tienen relacionadas con la experiencia en la tecnología evaluada en el experimento, ya que, en una gran mayoría de los estudios los participantes tienen que cumplir ciertos requerimientos para participar en el experimento o son entrenados para conseguir equiparar sus conocimientos.

8.2 Contextualización del Proceso Propuesto

El proceso propuesto en esta tesis tiene las siguientes ventajas. Nótese que estas ventajas están directamente relacionadas con el hecho de suplir las carencias encontradas en el Capítulo 2:

- El esquema de caracterización propuesto permite la descripción exhaustiva de los detalles de cada experimento, evitando que se omitan detalles relevantes que podrían estar influyendo en la variable respuesta.
- La identificación de similitudes y diferencias entre los experimentos de una familia se realiza de forma sistemática.
- El establecimiento del conjunto de variables a analizar se realiza de forma sistemática.
- El análisis de datos incorpora de forma conjunta todas las variables identificadas.
- El análisis de datos se realiza de manera incremental con lo cual se puede garantizar que el modelo final conseguido es válido.

Existen ciertas circunstancias que pueden afectar al proceso descrito. Estas circunstancias están relacionadas con:

- La disponibilidad de información sobre las familias de experimentos.
- El diseño de la familia de experimentos.
- El proceso en sí.

A continuación, se describe, asociadas a cada paso del proceso dichas circunstancias:

- Es necesario disponer de toda la información de cada uno de los experimentos que componen una familia. Por ejemplo, si se quiere aplicar el proceso a las familias del mapeo sistemático de la literatura, es posible que se pierdan variables relevantes porque en muchas ocasiones se omite información relativa a los experimentos, esto trae como consecuencia que la caracterización de estos no esté completa. Si la caracterización está incompleta habrá

variables de las que se desconozca el valor y por tanto no se podrán estudiar. Este problema está asociado con la disponibilidad de la información.

- En las actividades de identificación de variables moderadoras, si hay varias variables confundidas, a la hora de analizarlas, si resultan significativas no se podrá determinar cual de estas variables es la que está influyendo en la variable respuesta. Este problema está asociado con el diseño de la familia.
- En las actividades de refinamiento de variables moderadoras, si las variables no están balanceadas y finalmente hay que eliminarlas del análisis de datos, porque ocurran problemas con la prueba de análisis estadístico (modelo lineal mixto), se podría estar perdiendo información relevante para los resultados obtenidos. Este problema está asociado con el diseño de la familia.
- Si tanto la variable respuesta como las variables moderadoras identificadas no están medidas con la misma métrica en todos los experimentos, y no fuera posible transformarlas, no se podrían incluir en el análisis y esto podría hacer que se pierda información relevante. Este problema está asociado con el diseño de la familia.
- Finalmente, el proceso es largo y tedioso de aplicar, sobre todo el paso de la obtención de un modelo final válido que dependiendo del número de variables se puede incrementar considerablemente. Este problema está asociado con el proceso.

Paso	Actividad	Circunstancia	Motivo
1. Identificación de los Grupos de Agregación	1.1 Caracterización de experimentos de la familia	Pérdida de información	Falta de información
	1.2 Identificación de los grupos de agregación	Múltiples grupos de agregación	Diseño de la familia
2. Identificación de Variables de Diseño	2.2 Exploración de los resultados de los análisis individuales	Múltiples análisis: tantos como experimentos compongan la familia	Proceso largo y tedioso
3. Identificación de Variables Contextuales	3.1 Identificación de diferencias contextuales	Pérdida de información	Falta de información
	3.2 Creación de un conjunto inicial de variables contextuales	Pérdida de variables	Falta de información
	3.3 Refinamiento del conjunto inicial de variables contextuales	Pérdida de variables	Diseño de la familia
4. Identificación de Variables de Participantes	4.1 Identificación de diferencias entre los participantes	Pérdida de información	Falta de información
	4.2 Creación de un conjunto inicial de variables de participantes	Pérdida de variables	Falta de información
	4.3 Refinamiento del conjunto inicial de variables de participante	Pérdida de variables	Diseño de la familia
5. Realización del análisis	5.2 Análisis del modelo	Múltiples análisis: 25 modelos como mínimo (5 matrices de covarianza x 5 modelos)	Proceso largo y tedioso

Tabla 8.1 Circunstancias que afectan al proceso asociadas a cada paso de este

8.3 Líneas Futuras

En esta investigación no se han podido estudiar las siguientes líneas de investigación, que se consideran muy relevantes en IS. A continuación, se enumeran las más importantes:

- El método propuesto ha sido estudiado en dos familias una de tamaño grande y una de tamaño pequeño. Para que los resultados obtenidos sean más robustos, sería conveniente, validar el método estudiado con más familias de experimentos, principalmente, familias de tamaño medio, que son las más realizadas en IS.
- Las dos familias aquí estudiadas fueron diseñadas y ejecutadas por las directoras de esta tesis, por lo tanto, se posee pleno conocimiento sobre las características de los experimentos y su posible influencia. Sería conveniente, repetir este proceso en las familias de las que no se conocen los detalles tan profundamente, con el objetivo de revisar si los resultados son extensibles a familias de las que se tiene poca información. En concreto, sería deseable hacerlo con las familias estudiadas en el Capítulo 2.
- En todos los modelos explorados se ha mantenido fija la matriz de covarianza de las variables aleatorias. Sería interesante comprobar los resultados con las distintas matrices en las variables aleatorias.
- El análisis estadístico aquí propuesto estudia la significancia estadística de las variables establecidas como efectos fijos. Sería interesante, ampliar los resultados estudiando cuanta variabilidad está asociada a cada uno de los niveles del modelo jerárquico, es decir, conocer cual es el nivel en el que existe mayor variabilidad (que hace que el factor principal no se comporte de la misma manera) y que porcentaje de variabilidad tienen el resto de los niveles.
- El proceso aquí descrito (excepto el paso 5 relativo al análisis) se podrá replantear, para la planificación de los experimentos de una familia. De esta manera por ejemplo se podría evitar el desbalanceo entre las variables, o la falta de información en los experimentos.
- El proceso aquí propuesto debería ser evaluado por otras personas para poder extender su validez.

A.1 Operacionalización

Tabla A.1 Grupos de variación para los elementos de la operacionalización

Tabla A.2 Grupos de variación para los elementos de la operacionalización

Figura A.2 Distribución de cambios para la Ejecución Experimental

A.3 Instrumentación

Figura A.3 Distribución de cambios para la Instrumentación

Grupo	Estudio Primario	Total
Poco	[EP2], [EP5], [EP9], [EP12], [EP13], [EP15], [EP16], [EP17], [EP19], [EP20], [EP21], [EP23], [EP24], [EP26], [EP27], [EP28], [EP29], [EP30], [EP32], [EP33], [EP35], [EP37], [EP38], [EP39], [EP43], [EP44], [EP45], [EP46]	28 (61%)
Medio	[EP1], [EP3], [EP7], [EP8], [EP11], [EP22], [EP25], [EP34], [EP36], [EP40], [EP41], [EP42]	12 (26%)
Mucho	[EP4], [EP6], [EP10], [EP14], [EP18], [EP31]	6 (13%)

Tabla A.3 Grupos de variación para los elementos de la instrumentación

A.4 Población

Figura A.4 Distribución de cambios para la Población

Grupo	Estudio Primario	Total
Poco	[EP2], [EP3], [EP7], [EP9], [EP13], [EP15], [EP16], [EP17], [EP22], [EP24], [EP27], [EP29], [EP30], [EP31], [EP35], [EP37], [EP38], [EP43], [EP44], [EP46]	20 (43%)
Medio	[EP4], [EP6], [EP11], [EP14], [EP18], [EP20], [EP21], [EP23], [EP25], [EP26], [EP28], [EP32], [EP33], [EP34], [EP40], [EP42], [EP45]	17 (37%)
Mucho	[EP1], [EP5], [EP8], [EP10], [EP12], [EP19], [EP36], [EP39], [EP41]	9 (20%)

Tabla A.4 Grupos de variación para los elementos de la Población

A.5 Experimentador

Figura A.5 Distribución de cambios para los Experimentadores

Grupo	Estudio Primario	Total
Poco	[EP2], [EP3], [EP4], [EP5], [EP6], [EP7], [EP8], [EP9], [EP11], [EP12], [EP13], [EP15], [EP16], [EP18], [EP20], [EP21], [EP22], [EP23], [EP24], [EP25], [EP26], [EP27], [EP29], [EP30], [EP32], [EP33], [EP34], [EP35], [EP37], [EP40], [EP42], [EP43], [EP44], [EP45], [EP46]	35 (76%)
Medio	[EP14], [EP38], [EP39]	3 (7%)
Mucho	[EP1], [EP10], [EP17], [EP19], [EP28], [EP31], [EP36], [EP41]	8 (17%)

Tabla A.5 Grupos de variación para los elementos de Experimentador

Apéndice B Aplicación web para familias de Experimentos

B.1 Listado de usuarios

Name	Email	Rol	Action
Patricia Riofrio	pattyriofo@gmail.com	ADMIN
Santiago Rodriguez	santiagorodriguez@gmail.com	CREATOR
Kennith Wilkinson III	kertzmann.deshawn@example.org	VIEWER
Dr. Juanita Shanahan	gwaters@example.org	VIEWER
Mrs. Arlie Hartmann	johnston.jamar@example.org	VIEWER
Garth Bins	dameon.kilback@example.net	VIEWER
Raoul Wunsch	branson.bartoletti@example.org	VIEWER
Cathrine Howe	xschaefee@example.com	VIEWER
Janice Emmerich MD	alfonso21@example.org	VIEWER
Garry Haley	arempel@example.net	VIEWER

Figura B.6 Listado de usuarios de la aplicación

B.2 Creación de nuevo usuario

New user

Name

Email

Password

User Rol

State

Figura B.7 Creación de un nuevo usuario

B.3 Listado de familias de experimentos

Name	Description	Research Goal	Action
TDD	Lorem ipsum	Lorem ipsum	
quos	Est inventore veniam numquam sit quasi alias voluptatem.	Et nam distinctio dolores et quod aut deserunt.	
veniam	Rerum dolore aliquid qui ut.	Nemo debitis rerum et similique velit.	
minima	Ut beatae commodi assumenda molestiae.	Neque expedita labore vel fuga non.	
laudantium	Aut et voluptates ut ut.	Ut et necessitatibus exercitationem debitis.	
velit	Impedit asperiores accusamus ut autem necessitatibus quasi.	Dolorem officiis officia similique ut eaque maiores sed.	
incidunt	Laboriosam dicta totam harum id doloreque repellat quaerat.	Accusantium et sunt consectetur eos expedita qui vitae.	
non	Adipisci aut officia laboriosam sequi.	Id nobis numquam voluptatum quis et aut aliquid eligendi.	
magni	Earum tempora vel sapiente magnam minima suscipit laboriosam.	Itaque quia quae dicta.	
commodi	Numquam pariatur rerum totam.	Est aut recusandae corrupti quam molestiae quos ut.	

Figura B.8 Listado de familias de experimentos

B.4 Listado de experimentos de una familia

Experiments for family: TDD | [New Experiment](#)

FSecure (FS) [Edit identification info](#)

INSTITUTION:	FSecure	Operationalization	Data Analysis
TYPE:	INDUSTRY	Design	Population
PUBLICATION:	An Industry Experiment on the Effects of Test-Driven Development on External Quality and Productivity	Instrumentation	Results
SITES:	Country City Date From Date To		

Basilicata2015 (U. di Basilicata) [Edit identification info](#)

INSTITUTION:	University of Basilicata	Operationalization	Data Analysis
TYPE:	ACADEMIA	Design	Population
PUBLICATION:	An External Replication on the Effects of Test-driven Development Using a Multi-site Blind Analysis Approach	Instrumentation	Results
SITES:	Country City Date From Date To		

Figura B.9 Listado de los experimentos de una familia

B.5 Detalle de la operacionalización de un experimento

Factor: Development Approach

Treatment	literature	process	Instruction
Test-Driven Development	Process of writing unit tests before source code so that developers can better understand what the code needs to do	1. Decompose the specifications into small, manageable programming tasks. 2. Write low-level functional tests associated with each task before the production code. 3. Get instant feedback from tests to decide whether the task has been fully implemented as intended and whether it interferes with the existing code. 4. Write minimal pieces of code, necessary to make the tests pass, for the associated task. 5. Have an up-to-date and complete list of tests in place that are frequently run. 6. Focus on low-level design and incrementally refactor the production code.	It is more important to do a task well
Test Last Development	Writing code first and immediately writing tests before moving on to a new functionality	Both TDD and ITLD follow small steps, such as decomposing the specification into small programming tasks, coding, testing and refactoring. Slice is not taught	It is more important to do a task well

Figura B.10 Detalle de la Operacionalización de un experimento

B.6 Detalle de la Población de un experimento

Subjects Properties

Subjects' type during experiment						Characterization	Additional information	Action
Students				Professionals				
Undergraduate	Graduate	Master	PhD	Practitioner	Researcher			
0	0	0	0	24	0		7 (O) 11 (KL) 6 (H)	

More than 80% had 6 or more years of programming experience. Around half had 2 to 10 years of experience in Java. All were familiar with JUnit, and half had no experience in TDD. the majority of the participants had a bachelor's degree in computer science, computer engineering or electrical engineering.

Objects Properties

Experiment	Complexity	Domain	Language	Action
experimental tasks easier than their normal tasks	NO REAL	Java		

Figura B.11 Detalle de la Población de un experimento

B.7 Edición de un experimento (Operacionalización)

eseapp

Welcome, Patty

DASHBOARD

Users

Families

TDD: FSecure | [Back to experiments](#)

Operationalization Design Instrumentation Data Analysis Population Results

Cause | Effect | Data measurement

Deactivate edition

Factor: Development Approach

Treatment

Treatment name	literature	process	instruction
Test-Driven Development	Process of writing unit tests before source code so that developers can better	1. Decompose the specifications into small, measurable, verifiable	it is more important to do a task well
Test Last Development	Writing code first and immediately writing tests before making any new	Both TDD and ITLD follow small steps, such as decomposing the specification	it is more important to do a task well

Add Treatment +

Figura B.12 Edición de un experimento

Apéndice C Caracterización de los Experimentos

C.1 Experimentos de Técnicas de Prueba

C.1.1 UPM-2

Dimensión	Elemento	Valor
Identificación	Institución	UPM
	Sitio	Academia
	Publicación	(Juristo, et al., 2012)
Operacionalización	Factores	Técnica
	Tratamientos	Revisión
		Estructural
		Funcional
	Tratamiento acorde a la literatura	Lectura de código por abstracciones sucesivas (Linger, et al., 1979) Cobertura de sentencias (Beizer, 1990) Particiones en clases de equivalencia (Myers, et al., 2004)
	Tratamiento como instrucciones	Revisión de código Cobertura de decisiones Particiones en clases de equivalencia
	Proceso de aplicación del tratamiento	Procedimiento revisión de código Procedimiento cobertura de decisiones Procedimiento particiones en clases de equivalencia
	Variable respuesta	Efectividad
	Elaboración de la variable respuesta	- Revisión: Porcentaje de faltas reportadas - Estructural y Funcional: Porcentaje de faltas para los cuales se genera un caso de prueba
	Métrica	- Revisión: $\frac{FaltasReportadas}{FaltasTotal} * 100$
		- Estr. y Func.: $\frac{Faltas(CasoPrueba)}{FaltasTotal} * 100$
Medición	Instrumentos de medición	Ordenador, Folios, Bolígrafos (manual)
	Procedimiento de medida	Criterio Relajado
Ejecución experimental	Diseño	Diseño experimental
		Crossover
		Número de sesiones
		3
		Nº de factores por sesión
		1
		Nº tratamientos/factor/sesión
		3
		Variables de bloque
		Programa
		R-F-E
		R-E-F
	Orden de los tratamientos (grupos)	E-R-F
		E-F-R
		F-R-E
		F-E-R
		F-E-R
	Objetos exp. por sesión	1
	Nº unidades experimentales	108
	Número de sujetos	32
	Duración de la tarea	Ilimitado
Entrenamiento	Comb. entrenamiento y sesión	Secuencial
	Asp. de transmisión del trat.	Experiencia alta en enseñanza de técnicas de prueba
	Duración del entrenamiento	3 sesiones de 4 horas

Tabla C.1.6 Caracterización UPM-2 (1/2)

Dimensión		Elemento	Valor
Instrumentación		Guías suplementarias	Guías
		Formularios	Formularios
		Transmisión de los tratamientos	Instrucciones
		Objetos experimentales	Cmdline
			Nametbl
			Ntree
		Faltas	Instancia 1 Instancia 2
		Infraestructura tecnológica	Ordenador
Población	Sujetos	Tipo de sujetos	Estudiantes 5º año
		Experiencia con la técnica	No familiarizado
		Conocimiento en programación	Alta
		Motivación	Alta
	Objetos	Características variables	Exp. Relativa Exp. Absoluta
		Complejidad	Escala de Likert Número de años
		Dominio	Media-baja 15-29 Alta > 45
Experimentadores		Lenguaje de programación	Tareas de juguete C
		Diseñador	Juristo, Vegas
		Entrenador	Juristo, Moreno
		Monitor	Vegas, Moreno
		Medidor	Solari

Tabla C.1.7 Caracterización UPM-2 (2/2)

C.1.2 UPV

Dimensión		Elemento	Valor
Identificación		Institución	UPV
		Sitio	Academia
		Publicación	(Juristo, et al., 2012)
Operacionalización	Causa	Factores	Técnica
		Tratamientos	Estructural Funcional
		Tratamiento acorde a la literatura	Cobertura de sentencias (Beizer, 1990) Particiones en clases de equivalencia (Myers, et al., 2004)
		Tratamiento como instrucciones	Cobertura de decisiones Particiones en clases de equivalencia
		Proceso de aplicación del tratamiento	Procedimiento cobertura de decisiones Procedimiento particiones en clases de equivalencia
		Variable respuesta	Efectividad
	Efecto	Elaboración de la variable respuesta	- Estr. y Func.: Porcentaje de faltas para los cuales se genera un caso de prueba
		Métrica	- Estr. y Func.: $\frac{Faltas(CasoPrueba)}{FaltasTotal} * 100$
	Medición	Instrumentos de medición	Ordenador, Folios, Bolígrafos (manual)
		Procedimiento de medida	Criterio Relajado

Tabla C.1.8 Caracterización UPV (1/2)

Dimensión		Elemento	Valor	
Ejecución experimental	Diseño	Diseño experimental	Medidas repetidas cruzando tareas	
		Sesiones	2	
		Factores por sesión	1	
		Tratamientos x factor x sesión	1	
		Variables de bloque	Programa	
		Orden de los tratamientos (grupos)	E-F (cmdline-ntree)	
			E-F (nametbl-cmdline)	
			E-F (ntree-nametbl)	
			E-F (cmdline-nametbl)	
			E-F (nametbl-ntree)	
			E-F (ntree-cmdline)	
		Objetos experimentales por sesión	3	
		Unidades experimentales	62	
	Número de sujetos	33		
	Duración de la tarea	Limitado		
Entrenamiento	Combinación entrenamiento y sesiones	Intercalado		
	Aspectos de transmisión del tratamiento	Experiencia en enseñanza de técnicas de prueba alta		
	Duración del entrenamiento	2 sesiones de 2 horas		
Instrumentación	Guías suplementarias	Guías		
	Formularios	Formularios		
	Transmisión de tratamientos	Instrucciones		
	Objetos experimentales	Programas	Cmdline Nametbl Ntree	
		Faltas	Instancia 1	
			Instancia 2	
	Infraestructura tecnológica	Ordenador		
Población	Sujetos	Tipo de sujetos	Estudiantes 5º año	
		Características comunes	Experiencia con la técnica	Familiarizado
			Conocimiento en programación	Alta
			Motivación	Baja
	Objetos	Características variables	Exp. Relativa	Escala de Likert
			Exp. Absoluta	Número de años
		Complejidad		Media-baja 15-29 Alta > 45
Experimentadores	Diseñador		Juristo, Vegas	
	Entrenador		Letelier	
	Monitor		Letelier	
	Medidor		Solari	

Tabla C.1.9 Caracterización UPV (2/2)

C.1.3 ORT

Dimensión	Elemento	Valor
Identificación	Institución	ORT
	Sitio	Academia
	Publicación	(Juristo, et al., 2012)
Operacionalización	Factores	Técnica
	Tratamientos	Estructural Funcional
	Tratamiento acorde a la literatura	Cobertura de sentencias (Beizer, 1990) Particiones en clases de equivalencia (Myers, et al., 2004)
	Tratamiento como instrucciones	Cobertura de decisiones Particiones en clases de equivalencia
	Proceso de aplicación del tratamiento	Procedimiento cobertura de decisiones Procedimiento particiones en clases de equivalencia
	Variable respuesta	Efectividad
	Elaboración de la variable respuesta	- Estr. y Func.: Porcentaje de faltas para los cuales se genera un caso de prueba
	Métrica	- Estr. y Func.: $\frac{Faltas(CasoPrueba)}{FaltasTotal} * 100$
	Medición	Instrumentos de medición
		Procedimiento de medida
	Ejecución experimental	Ordenador, Folios, Bolígrafos (manual) Criterio Relajado
		Diseño experimental
		Crossover
		Sesiones
		2
		Factores por sesión
		1
		Tratamientos por factor por sesión
		2
		Variables de bloque
		Programa
		Orden de los tratamientos (grupos)
		E-F (cmdline-nametbl) E-F (nametbl- cmdline) F-E (cmdline-nametbl) F-E (nametbl- cmdline)
		Objetos experimentales por sesión
		2
		Unidades experimentales
		152
		Número de sujetos
		78
		Duración de la tarea
		Ilimitado
	Entrenamiento	Combinación entre el entrenamiento y las sesiones
		Secuencial
		Aspectos de transmisión del tratamiento
		Experiencia en enseñanza de técnicas de prueba alta
		Duración del entrenamiento
		3 sesiones de 4 horas
Instrumentación	Guías suplementarias	Guías
		Formularios
		Formularios
	Transmisión de tratamientos	Instrucciones
		Objetos experimentales
		Programas
		Cmdline Nametbl
	Faltas	Instancia 1
		Instancia 2
	Infraestructura tecnológica	
	Ninguna	

Tabla C.1.10 Caracterización ORT (1/2)

Dimensión		Elemento	Valor	
Población	Sujetos	Tipo de sujetos	Estudiantes de 2º año	
		Características comunes	Experiencia con la técnica	No familiarizado
			Conocimiento en programación	Baja
			Motivación	Alta
	Objetos	Características variables	Exp. Relativa	Escala de Likert
			Exp. Absoluta	Años
		Complejidad	Media-baja 15-29	
Alta > 45				
Dominio	Tareas de juguete			
		Lenguaje de programación	C	
Experimentadores	Diseñador		Solari	
	Entrenador		Solari	
	Monitor		Solari	
	Medidor		Solari	

Tabla C.1.11 Caracterización ORT (2/2)

C.2 Experimentos de Test Driven Development (TDD)

C.2.1 FSecure

Dimensión		Elemento	Valor	
Identificación		Institución	FSH, FSK, FKO	
		Sitio	Industria	
		Publicación	(Tosun, et al., 2017)	
Operacionalización	Causa	Factores	Enfoque de Desarrollo	
		Tratamientos	ITL TDD	
		Tratamiento acorde a la literatura	Sección 6.1.1	
		Tratamiento como instrucciones		
		Proceso de aplicación del tratamiento		
		Efecto	Variable Respuesta	Cantidad de Trabajo Calidad
	Elaboración de la variable respuesta		Sección 6.1.2	
	Métrica			
	Medición	Instrumentos de medida	Ad-hoc	
		Procedimiento de medición	Sintáctica (Manual)	
Ejecución experimental	Diseño	Diseño experimental	Intra-sujetos	
		Sesiones	3	
		Factores por sesión	1	
		Tratamientos por factor por sesión	1	
		Variables de bloque	Ninguna	
		Orden de los tratamientos (grupos)	Fijo: ITL-TDD-TDD	
		Objetos experimentales por sesión	1	
		Unidades experimentales	72	
		Número de sujetos	24	
		Duración de la tarea	Ilimitada	
	Entrenamiento	Combinación entrenamiento y sesión	Intercalado	
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD	
		Duración del entrenamiento	Adaptado	
Instrumentación		Guías suplementarias	BSK-Java MR-Java MF-Java	
		Formularios	Ninguno	
		Transmisión de tratamientos (slices)	TLD	
		Objetos experimentales	BSK MR MF	
		Infraestructura tecnológica	Junit, Eclipse	
Población	Sujetos	Caract. comunes	Tipo de sujetos	Profesional
		Características variables	Exp Programación	Escala de Likert
			Exp con Lenguaje	Escala de Likert
			Exp pruebas unitarias	Escala de Likert
			Exp framework pruebas	Escala de Likert
	Objetos	Complejidad	Más fácil que las tareas que realizan habitualmente	
		Dominio	No real	
		Lenguaje	Java	
		Granularidad	BSK_S MR_NS MF_NS	

Tabla C.2.12 Caracterización FSecure (1/2)

Dimensión	Elemento	Valor
Experimentadores	Diseñador	Juristo, Dieste
	Entrenador	Turhan
	Monitor	Dieste, Fucci, Tosun
	Medidor	Uyaguari, Fucci

Tabla C.2.13 Caracterización FSecure (2/2)

C.2.2 Elektrobit

Dimensión		Elemento	Valor	
Identificación		Institución	EB	
		Sitio	Industria	
		Publicación	No	
Operacionalización	Causa	Factores	Enfoque de Desarrollo	
		Tratamientos	ITL TDD	
		Tratamiento acorde a la literatura	Sección 6.1.1	
		Tratamiento como instrucciones		
		Proceso de aplicación del tratamiento		
		Efecto	Variable Respuesta	Cantidad de Trabajo Calidad
	Elaboración de la variable respuesta		Sección 6.1.2	
	Métrica			
	Medición	Instrumentos de medida	Ad-hoc	
		Procedimiento de medición	Sin intervención (Manual)	
Ejecución experimental	Diseño	Diseño experimental	Intra-sujetos	
		Sesiones	3	
		Factores por sesión	1	
		Tratamientos por factor por sesión	1	
		Variables de bloque	Ninguna	
		Orden de los tratamientos (grupos)	Fijo: ITL-TDD-TDD	
		Objetos experimentales por sesión	1	
		Unidades experimentales	27	
		Número de sujetos	9	
		Duración de la tarea	Limitada	
	Entrenamiento	Combinación entrenamiento y sesión	Intercalado	
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD	
		Duración del entrenamiento	Adaptado	
Instrumentación		Guías suplementarias	BSK-C++ MR-C++ MF-C++	
		Formularios	Ninguno	
		Transmisión de tratamientos (slices)	TLD	
		Objetos experimentales	BSK MR MF	
		Infraestructura tecnológica	GTEST, Eclipse CDT (linux)	
	Población	Sujetos	Caract. comunes	Tipo de sujetos
Características variables			Exp Programación	Escala de Likert
			Exp con Lenguaje	Escala de Likert
			Exp pruebas unitarias	Escala de Likert
			Exp framework pruebas	Escala de Likert

Tabla C.214 Caracterización Elektrobit (1/2)

Dimensión	Elemento	Valor
Población	Objetos	Complejidad
		Más fácil que las tareas que realizan habitualmente
		Dominio
		No real
Experimentadores		Lenguaje
		C++
		Granularidad
		BSK_S MR_NS MF_NS
	Experimentadores	Diseñador
		Juristo, Dieste
		Entrenador
		Turhan
		Monitor
		Dieste, Fucci, Tosun
		Medidor
		Uyaguari

Tabla C.2.15 Caracterización Elektrobit (2/2)

C.2.3 Universidad Politécnica de Madrid

Dimensión	Elemento	Valor
Identificación	Institución	UPM
	Sitio	Academia
	Publicación	No
Operacionalización	Causa	Factores
		Enfoque de Desarrollo
		Efecto de Aprendizaje
		Tratamientos
		Enfoque de ITL
		Desarrollo TDD
	Efecto	Efecto de Primera
		Aprendizaje Segunda
		Tratamiento acorde a la literatura
		Tratamiento como instrucciones
Ejecución experimental	Diseño	Proceso de aplicación del tratamiento
		Sección 6.1.1
		Cantidad de Trabajo
		Calidad
		Número de casos de prueba
		Elaboración de la variable respuesta
		Métrica
		Sección 6.1.2
	Medición	Instrumentos de medida
		Ad-hoc
	Entrenamiento	Procedimiento de medición
		Sintáctica (Manual)
		Diseño experimental
		Medidas repetidas cruzando tareas
		Sesiones
		4
		Factores por sesión
		2
		Tratamientos por factor por sesión
		1, 1
		Variables de bloque
		Objeto experimental
		Orden de los tratamientos (grupos)
		Fijo: ITL-TDD-ITL-TDD
		Objetos experimentales por sesión
		4
		Unidades experimentales
		16
		Número de sujetos
		16
		Duración de la tarea
		Ilimitada
		Combinación entrenamiento y sesión
		Intercalado
		Aspectos de transmisión del tratamiento
		Exp alta en enseñanza de TDD
		Duración del entrenamiento
		Adaptado

Tabla C.2.16 Caracterización UPM (1/2)

Dimensión	Elemento	Valor
Instrumentación	Guías suplementarias	BSK-Java
		MR-Java
		MF-Java
		Sudoku-Java
	Formularios	Ninguno
	Transmisión de tratamientos (slices)	TLD
Población	Sujetos	Objetos experimentales
		BSK
		MR
		MF
		Sudoku
		Infraestructura tecnológica
		Junit, Eclipse
	Objetos	Caract. comunes Tipo de sujetos
		Estudiantes de Máster
		Exp Programación
		Escala de Likert
		Exp con Lenguaje
		Escala de Likert
		Exp pruebas unitarias
		Escala de Likert
		Exp framework pruebas
		Escala de Likert
Experimentadores	Complejidad	Igual a las tareas que realizan habitualmente
		Dominio
		No real
		Lenguaje
		Java
	Granularidad	BSK_S
		MR_NS
		MF_NS
		Sudoku_NS
		Diseñador
		Juristo, Dieste, Vegas, Turhan
	Entrenador	Turhan
		Monitor
		Dieste
		Medidor
		Uyaguari

Tabla C.2.17 Caracterización UPM (2/2)

C.2.4 PlayTech

Dimensión		Elemento	Valor
Identificación		Institución	PT
		Sitio	Industria
		Publicación	No
Operacionalización	Causa	Factores	Enfoque de Desarrollo
		Tratamientos	ITL
			TDD
		Tratamiento acorde a la literatura	Sección 6.1.1
		Tratamiento como instrucciones	
		Proceso de aplicación del tratamiento	
		Efecto	Variable Respuesta
			Calidad
			Número de casos de prueba
	Elaboración de la variable respuesta		Sección 6.1.2
	Métrica		
	Medición	Instrumentos de medida	Ad-hoc
Procedimiento de medición		Sintáctica (Manual)	

Tabla C.1.18 Caracterización PlayTech (1/2)

Dimensión	Elemento		Valor
Ejecución experimental	Diseño	Diseño experimental	Medidas repetidas cruzando tareas
		Sesiones	3
		Factores por sesión	1
		Tratamientos por factor por sesión	1
		Variables de bloque	Objeto experimental
		Orden de los tratamientos (grupos)	Fijo: ITL-TDD-TDD
		Objetos experimentales por sesión	2 o 1
		Unidades experimentales	51
		Número de sujetos	17
		Duración de la tarea	Ilimitada
Instrumentación	Entrenamiento	Combinación entrenamiento y sesión	Intercalado
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD
		Duración del entrenamiento	Adaptado
		Guías suplementarias	BSK-Java MR-Java MF-Java
		Formularios	Ninguno
		Transmisión de tratamientos (slices)	ITL
		Objetos experimentales	BSK MR MF
		Infraestructura tecnológica	Junit, Eclipse
		Caract. comunes	Tipo de sujetos
			Profesional
Población	Sujetos	Características variables	Exp Programación
			Escala de Likert
			Exp con Lenguaje
			Escala de Likert
			Exp pruebas unitarias
	Objetos	Complejidad	Escala de Likert
			Exp framework pruebas
			Escala de Likert
			Más fácil que las tareas que realizan habitualmente
			No real
Experimentadores	Experimentadores	Medidor	Java
			BSK_S
			MR_NS
			MF_NS
			Diseñador
			Juristo, Dieste
			Entrenador
			Dieste
			Monitor
			Tosun
			Medidor
			Uyaguari, Jaimes

Tabla C.2.19 Caracterización PlayTech (2/2)

C.2.5 Universidad Politécnica de Valencia

Dimensión		Elemento	Valor	
Identificación		Institución	UPV	
		Sitio	Academia	
		Publicación	No	
Operacionalización	Causa	Factores	Enfoque de Desarrollo	
		Tratamientos	ITL TDD	
		Tratamiento acorde a la literatura	Sección 6.1.1	
		Tratamiento como instrucciones		
		Proceso de aplicación del tratamiento		
		Efecto	Variable Respuesta	Cantidad de Trabajo Calidad
	Elaboración de la variable respuesta		Sección 6.1.2	
	Métrica			
	Medición	Instrumentos de medida	Ad-hoc	
		Procedimiento de medición	Sintáctica (Manual)	
Ejecución experimental	Diseño	Diseño experimental	Intra-sujetos	
		Sesiones	2	
		Factores por sesión	1	
		Tratamientos por factor por sesión	1	
		Variables de bloque	Ninguna	
		Orden de los tratamientos (grupos)	Fijo: ITL-TDD	
		Objetos experimentales por sesión	1	
		Unidades experimentales	62	
		Número de sujetos	31	
		Duración de la tarea	Limitada	
	Entrenamiento	Combinación entrenamiento y sesión	Intercalado	
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD	
	Duración del entrenamiento	Adaptado		
Instrumentación		Guías suplementarias	BSK-Java MR-Java	
		Formularios	Ninguno	
		Transmisión de tratamientos (slices)	ITL	
		Objetos experimentales	BSK MR	
		Infraestructura tecnológica	Junit, Eclipse	
Población	Sujetos	Caract. comunes	Tipo de sujetos	Estudiantes de máster
		Características variables	Exp Programación	Escala de Likert
			Exp con Lenguaje	Escala de Likert
			Exp pruebas unitarias	Escala de Likert
			Exp framework pruebas	Escala de Likert
	Objetos	Complejidad	Igual que las tareas que realizan habitualmente	
		Dominio	No real	
		Lenguaje	Java	
		Granularidad	BSK_S MR_NS	
Experimentadores		Diseñador	Juristo, Dieste	
		Entrenador	Dieste	
		Monitor	Dieste	
		Medidor	Uyaguari	

Tabla C.2.20 Caracterización UPV

C.2.6 Ericsson

Dimensión		Elemento	Valor	
Identificación		Institución	ER	
		Sitio	Industria	
		Publicación	No	
Operacionalización	Causa	Factores	Enfoque de Desarrollo Granularidad	
		Tratamientos	Enfoque de Desarrollo	ITL TDD
			Granularidad	Grano fino Grano grueso
			Tratamiento acorde a la literatura	
		Tratamiento como instrucciones		
		Proceso de aplicación del tratamiento		
		Efecto	Variable Respuesta	Cantidad de Trabajo Calidad Número de casos de prueba
	Elaboración de la variable respuesta			
	Métrica			
	Medición		Instrumentos de medida	Ad-hoc
		Procedimiento de medición	Sintáctica (Manual)	
	Ejecución experimental	Diseño	Diseño experimental	Medidas repetidas cruzando tareas
Sesiones			2	
Factores por sesión			2	
Tratamientos por factor por sesión			1, 2	
Variables de bloque			Objeto experimental	
Orden de los tratamientos (grupos)			Fijo: ITL-TDD	
Objetos experimentales por sesión			2	
Unidades experimentales			40	
Número de sujetos			20	
Duración de la tarea		Ilimitada		
Entrenamiento		Combinación entrenamiento y sesión	Intercalado	
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD	
	Duración del entrenamiento	Adaptado		
Instrumentación		Guías suplementarias	BSK-C++ MR-C++	
		Formularios	Ninguno	
		Transmisión de tratamientos (slices)	ITL	
		Objetos experimentales	BSK MR	
		Infraestructura tecnológica	BOOST, Eclipse, Vim	
Población	Sujetos	Caract. comunes	Tipo de sujetos	Profesional
		Características variables	Exp Programación	Escala de Likert
			Exp con Lenguaje	Escala de Likert
			Exp pruebas unitarias	Escala de Likert
			Exp framework pruebas	Escala de Likert

Tabla C.2.21 Caracterización Ericsson (1/2)

Dimensión	Elemento	Valor
Población	Objetos	Complejidad
		Más fácil que las tareas que realizan habitualmente
		Dominio
		No real
		Lenguaje
Experimentadores	Granularidad	C++
		BSK_S
		BSK_NS
		MR_S
	Diseñador	MR_NS
		Juristo, Dieste
		Entrenador
		Dieste
	Monitor	Dieste, Tosun
		Medidor
		Uyaguari

Tabla C.2.22 Caracterización Ericsson (2/2)

C.2.7 Universidad de Basilicata

Dimensión	Elemento	Valor
Identificación	Institución	UPV
	Sitio	Academia
	Publicación	(Fucci, et al., 2016)
Operacionalización	Causa	Factores
		Enfoque de Desarrollo
		Tratamientos
		ITL
		TDD
		Tratamiento acorde a la literatura
	Efecto	Tratamiento como instrucciones
		Sección 6.1.1
		Proceso de aplicación del tratamiento
	Medición	Variable Respuesta
		Calidad
		Número de casos de prueba
	Ejecución experimental	Elaboración de la variable respuesta
		Sección 6.1.2
		Métrica
Instrumentación	Diseño	Instrumentos de medida
		Ad-hoc
		Procedimiento de medición
		Sintáctica (Automática)
		Diseño experimental
		Crossover
		Sesiones
		2
		Factores por sesión
	Entrenamiento	1
		Tratamientos por factor por sesión
		2
	Instrumentación	Variables de bloque
		Ninguna
		Orden de los tratamientos (grupos)
		Todos: ITL-TDD y TDD-ITL
		Objetos experimentales por sesión
		1
		Unidades experimentales
		40
		Número de sujetos
		20
		Duración de la tarea
		Ilimitada
		Combinación entrenamiento y sesión
		Secuencial
		Aspectos de transmisión del tratamiento
		Exp alta en enseñanza de TDD
		Duración del entrenamiento
		Adaptado
	Instrumentación	Guías suplementarias
		BSK-Java
		MR-Java
		Formularios
		Ninguno
		Transmisión de tratamientos (slices)
		TLD
		Objetos experimentales
		BSK
		MR
		Infraestructura tecnológica
		Junit, Eclipse

Tabla C.2.23 Caracterización UB (1/2)

Población	Sujetos	Caract. comunes	Tipo de sujetos	Estudiantes de máster
		Características variables	Exp Programación	Escala de Likert
			Exp con Lenguaje	Escala de Likert
			Exp pruebas unitarias	Escala de Likert
			Exp framework pruebas	Escala de Likert
	Objetos	Complejidad		Igual que las tareas que realizan habitualmente
		Dominio		No real
		Lenguaje		Java
		Granularidad		BSK_S MR_NS
	Experimentadores	Diseñador	Turhan, Fucci	
Entrenador		Scanniello		
Monitor		Scanniello, Romano		
Medidor		Uyaguari		

Tabla C.2.24 Caracterización UB (2/2)

C.2.8 Universidad de Dinamarca

Dimensión		Elemento	Valor		
Identificación		Institución	USD		
		Sitio	Academia		
		Publicación	No		
Operacionalización	Causa	Factores	Enfoque de Desarrollo		
			Granularidad		
		Tratamientos	Enfoque de Desarrollo	ITL	
				TDD	
			Granularidad	Grano fino Grano grueso	
		Tratamiento acorde a la literatura		Sección 6.1.1	
		Tratamiento como instrucciones			
		Proceso de aplicación del tratamiento			
	Efecto	Variable Respuesta	Cantidad de trabajo Calidad Número de casos de prueba		
		Elaboración de la variable respuesta		Sección 6.1.2	
		Métrica			
		Medición	Instrumentos de medida	Ad-hoc	
	Procedimiento de medición		Sintáctica-Semántica (Manual)		
	Ejecución experimental	Diseño	Diseño experimental	Entre-sujetos	
			Sesiones	2	
Factores por sesión			2		
Tratamientos por factor por sesión			2, 2		
Variables de bloque			Nivel académico		
Orden de los tratamientos (grupos)			Ninguno		
Objetos experimentales por sesión			1		
Unidades experimentales			75		
Número de sujetos			75		
Duración de la tarea			Ilimitada		
Entrenamiento		Combinación entrenamiento y sesión	Secuencial		
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD		
		Duración del entrenamiento	Adaptado		

Tabla C.2.25 Caracterización USD (1/2)

Dimensión		Elemento		Valor
Instrumentación		Guías suplementarias		BSK-C#
		Formularios		Ninguno
		Transmisión de tratamientos (slices)		TLD
		Objetos experimentales		BSK
		Infraestructura tecnológica		VS Testing Tool, Visual Studio
Población	Sujetos	Caract. comunes	Tipo de sujetos	Estudiantes de grado y máster
		Características variables	Exp Programación	Escala de Likert
			Exp con Lenguaje	Escala de Likert
			Exp pruebas unitarias	Escala de Likert
			Exp framework pruebas	Escala de Likert
	Objetos	Complejidad		Más fácil que las tareas que realizan habitualmente
		Dominio		No real
		Lenguaje		C#
		Granularidad		BSK_S BSK_NS
	Experimentadores		Diseñador	
Entrenador			Kuhrmann	
Monitor			Kuhrmann	
Medidor			Uyaguari, Fucci, Karac	

Tabla C.2.26 Caracterización USD (2/2)

C.2.9 PAF

Dimensión		Elemento	Valor
Identificación		Institución	PAF
		Sitio	Industria
		Publicación	No
Operacionalización	Causa	Factores	Enfoque de Desarrollo
		Tratamientos	ITL
			TDD
			YW
		Tratamiento acorde a la literatura	Sección 6.1.1
		Tratamiento como instrucciones	
		Proceso de aplicación del tratamiento	
	Efecto	Variable Respuesta	Cantidad de trabajo
			Calidad
			Número de casos de prueba
		Elaboración de la variable respuesta	Sección 6.1.2
	Métrica		
Medición	Instrumentos de medida	Ad-hoc	
	Procedimiento de medición	Semántica (Automática)	
Ejecución experimental	Diseño	Diseño experimental	Medidas repetidas cruzando tareas
		Sesiones	3
		Factores por sesión	1
		Tratamientos por factor por sesión	1
		Variables de bloque	Objeto experimental
		Orden de los tratamientos (grupos)	Fijo: YW-ITL-TDD
		Objetos experimentales por sesión	3
		Unidades experimentales	27
		Número de sujetos	9
		Duración de la tarea	Ilimitada

Tabla C.2.27 Caracterización PAF (1/2)

Dimensión		Elemento	Valor		
Ejec Exp.	Entrenamiento	Combinación entrenamiento y sesión	Intercalada		
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD		
		Duración del entrenamiento	Adaptado		
Instrumentación		Guías suplementarias	BSK-Java MR-Java SS-Java		
		Formularios	Ninguno		
		Transmisión de tratamientos (slices)	ITL		
		Objetos experimentales	BSK MR SS		
		Infraestructura tecnológica	Junit, IntelliJ, Eclipse		
		Población	Sujetos	Caract. comunes Tipo de sujetos	Profesional
				Características variables	Exp Programación
Exp con Lenguaje	Escala de Likert				
Exp pruebas unitarias	Escala de Likert				
Exp framework pruebas	Escala de Likert				
Objetos	Complejidad		Más fácil que las tareas que realizan habitualmente		
	Dominio		No real		
	Lenguaje		Java		
	Granularidad		BSK_NS MR_NS SS_NS		
			Experimentadores	Diseñador	Juristo, Dieste, Vegas
Entrenador		Dieste			
Monitor	Santos				
Medidor	Uyaguari				

Tabla C.2.28 Caracterización PAF (2/2)

C.2.10 ETAPA

Dimensión		Elemento	Valor
Identificación		Institución	ETAPA
		Sitio	Industria
		Publicación	No
Operacionalización	Causa	Factores	Enfoque de Desarrollo
			ITL
		Tratamientos	TDD
			YW
		Tratamiento acorde a la literatura	
		Tratamiento como instrucciones	Sección 6.1.1
		Proceso de aplicación del tratamiento	
	Efecto		Cantidad de trabajo
		Variable Respuesta	Calidad
			Número de casos de prueba
		Elaboración de la variable respuesta	
		Métrica	Sección 6.1.2
	Medición	Instrumentos de medida	Ad-hoc
		Procedimiento de medición	Semántica (Manual)

Tabla C.2.29 Caracterización ETAPA (1/2)

Dimensión	Elemento	Valor
Ejecución experimental	Diseño	Diseño experimental
		Medidas repetidas cruzando tareas
		Sesiones
		3
		Factores por sesión
		1
		Tratamientos por factor por sesión
		1
		Variables de bloque
		Objeto experimental
Ejecución experimental	Entrenamiento	Orden de los tratamientos (grupos)
		Fijo: YW-ITL-TDD
		Objetos experimentales por sesión
		3
		Unidades experimentales
		36
		Número de sujetos
		12
		Duración de la tarea
		Limitada
Instrumentación		Combinación entrenamiento y sesión
		Intercalada
		Aspectos de transmisión del tratamiento
		Exp alta en enseñanza de TDD
		Duración del entrenamiento
		Adaptado
		Guías suplementarias
		Ninguna
		Formularios
		Ninguno
Población	Sujetos	Transmisión de tratamientos (slices)
		ITL
		Objetos experimentales
		BSK
		MR
		SS
		Infraestructura tecnológica
		GeneXus, Genexus unit testing facilities
		Caract. comunes Tipo de sujetos
		Profesional
Población	Objetos	Exp Programación
		Escala de Likert
		Características variables
		Exp con Lenguaje
		Escala de Likert
		Exp pruebas unitarias
		Escala de Likert
		Exp framework pruebas
		Escala de Likert
		Complejidad
Experimentadores		Más fácil que las tareas que realizan habitualmente
		Dominio
		No real
		Lenguaje
		Genexus Evolution 1
		Granularidad
		BSK_NS
		MR_NS
		SS_NS
		Diseñador
Experimentadores		Entrenador
		Juristo, Dieste, Vegas
		Monitor
		Dieste
		Medidor
		Dieste
		Palomeque

Tabla C.2.30 Caracterización ETAPA (2/2)

C.2.11 Universidad de Dinamarca 2

Dimensión		Elemento	Valor		
Identificación		Institución	USD2		
		Sitio	Academia		
		Publicación	No		
Operacionalización	Causa	Factores	Enfoque de Desarrollo		
		Tratamientos	Granularidad		
			Enfoque de Desarrollo	ITL	
			Granularidad	TDD	
		Tratamiento acorde a la literatura	Grano fino		
		Tratamiento como instrucciones	Grano grueso		
		Proceso de aplicación del tratamiento	Sección 6.1.1		
	Efecto	Variable Respuesta	Calidad		
		Elaboración de la variable respuesta	Número de casos		
		Métrica	Sección 6.1.2		
	Medición	Instrumentos de medida	Ad-hoc		
		Procedimiento de medición	Sintáctica-Semántica (Manual)		
	Ejecución experimental	Diseño	Diseño experimental	Entre-sujetos	
			Sesiones	2	
Factores por sesión			2		
Tratamientos por factor por sesión			2, 2		
Variables de bloque			Ninguna		
Orden de los tratamientos (grupos)			Ninguno		
Objetos experimentales por sesión			1		
Unidades experimentales			62		
Número de sujetos			31		
Duración de la tarea		Limitada			
Entrenamiento	Combinación entrenamiento y sesión	Secuencial			
	Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD			
	Duración del entrenamiento	Adaptado			
Instrumentación		Guías suplementarias	BSK-C#		
		Formularios	Ninguno		
		Transmisión de tratamientos (slices)	TLD		
		Objetos experimentales	BSK		
		Infraestructura tecnológica	VS Testing Tool, Visual Studio		
Población	Sujetos	Caract. comunes	Tipo de sujetos	Estudiantes de grado	
		Características variables	Exp Programación	Escala de Likert	
			Exp con Lenguaje	Escala de Likert	
			Exp pruebas unitarias	Escala de Likert	
			Exp framework pruebas	Escala de Likert	
	Objetos	Complejidad	Más fácil que las tareas que realizan habitualmente		
		Dominio	No real		
		Lenguaje	C#		
		Granularidad	BSK_S		
			BSK_NS		
Experimentadores		Diseñador	Juristo, Turhan, Kuhrmann		
		Entrenador	Kuhrmann		
		Monitor	Kuhrmann		
		Medidor	Karac		

Tabla C.2.31 Caracterización USD2

C.2.12 Universidad de Oulu 2

Dimensión	Elemento	Valor
Identificación	Institución	OY2
	Sitio	Academia
	Publicación	No
Operacionalización	Factores	Enfoque de Desarrollo Granularidad
	Causa	Enfoque de Desarrollo
		ITL
		TDD
	Tratamientos	Granularidad
	Tratamiento acorde a la literatura	
	Tratamiento como instrucciones	
	Proceso de aplicación del tratamiento	
	Efecto	Calidad
		Número de casos
		Conformidad del proceso
Ejecución experimental	Diseño	Número de Episodios
		Cobertura de código
		Elaboración de la variable respuesta
		Métrica
		Sección 6.1.1
		Sección 6.1.2
		Sección 6.1.2
		Sección 6.1.2
		Sección 6.1.2
	Medición	Instrumentos de medida
		Procedimiento de medición
	Entrenamiento	Ad-hoc
		Sintáctica-Semántica (Manual)
		Intra-sujetos
		1, 2 Períodos
		2
		2, 2
		Ninguna
		Fijo: ITL-TDD
		1
Instrumentación	Diseño	90
		45
		Limitada
		Combinación entrenamiento y sesión
		Secuencial
		Aspectos de transmisión del tratamiento
		Exp alta en enseñanza de TDD
		Duración del entrenamiento
		Adaptado
	Entrenamiento	Guías suplementarias
		BSK-Java
		MR-Java
		Formularios
		Ninguno
Población	Instrumentación	Transmisión de tratamientos (slices)
		TLD
		Objetos experimentales
		BSK
		MR
	Infraestructura tecnológica	JUnit, Exlipse, Besouro, EclEmma
		Caract. comunes
		Tipo de sujetos
		Estudiantes de grado y máster
		Exp Programación
	Sujetos	Escala de Likert
		Exp con Lenguaje
		Escala de Likert
		Exp pruebas unitarias
	Características variables	Escala de Likert
		Exp framework pruebas
		Escala de Likert

Tabla C.2.32 Caracterización OY2 (1/2)

Dimensión		Elemento	Valor
Población	Objetos	Complejidad	Más fácil que las tareas que realizan habitualmente
		Dominio	No real
		Lenguaje	Java
			BSK_S
	Granularidad		BSK_NS
			MR_S
			MR_NS
Experimentadores	Diseñador	Juristo, Turhan, Karac	
	Entrenador	Karac	
	Monitor	Karac	
	Medidor	Karac	

Tabla C.2.33 Caracterización OY2 (2/2)

C.2.13 Universidad de NoviSad

Dimensión		Elemento	Valor		
Identificación		Institución	NoviSad		
		Sitio	Academia		
		Publicación	No		
Operacionalización	Causa	Factores	Enfoque de Desarrollo		
		Tratamientos	Enfoque de Desarrollo	ITL	
				TDD	
			Tratamiento acorde a la literatura		Sección 6.1.1
		Tratamiento como instrucciones			
		Proceso de aplicación del tratamiento			
	Efecto	Variable Respuesta	Calidad		
			Número de casos		
			Conformidad del proceso		
			Número de Episodios		
			Cobertura de código		
			Elaboración de la variable respuesta		Sección 6.1.2
			Métrica		
	Medición	Instrumentos de medida	Ad-hoc		
Procedimiento de medición		Sintáctica-Semántica (Manual)			
Ejecución experimental	Diseño	Diseño experimental	Medidas repetidas cruzando tareas		
		Sesiones	1, 2 Períodos		
		Factores por sesión	1		
		Tratamientos por factor por sesión	2		
		Variables de bloque	Objeto experimental		
		Orden de los tratamientos (grupos)	Fijo: ITL-TDD		
		Objetos experimentales por sesión	2		
		Unidades experimentales	53		
		Número de sujetos	101		
		Duración de la tarea	Limitada		
	Entrenamiento	Combinación entrenamiento y sesión	Secuencial		
		Aspectos de transmisión del tratamiento	Exp alta en enseñanza de TDD		
		Duración del entrenamiento	Adaptado		

Tabla C.2.34 Caracterización NoviSad (1/2)

Dimensión		Elemento		Valor
Instrumentación		Guías suplementarias		BSK-Java MR-Java
		Formularios		Ninguno
		Transmisión de tratamientos (slices)		TLD
		Objetos experimentales		BSK MR
		Infraestructura tecnológica		Junit, Exlipse, Besouro, EclEmma
		Población	Sujetos	Caract. comunes
Características variables	Exp Programación			Escala de Likert
	Exp con Lenguaje			Escala de Likert
	Exp pruebas unitarias			Escala de Likert
	Exp framework pruebas			Escala de Likert
	Objetos	Complejidad		Igual que las tareas que realizan habitualmente
		Dominio		No real
		Lenguaje		Java
		Granularidad		BSK_NS MR_NS
Experimentadores		Diseñador		Juristo, Turhan, Karac, Mandić, Ramač
		Entrenador		Mandić, Ramač
		Monitor		Karac, Mandić, Ramač
		Medidor		Karac, Ramač

Tabla C.2.35 Caracterización NoviSad (2/2)

Apéndice D Identificación de Variables de Diseño

D.1 Experimentos de Técnicas de Prueba

D.1.1 UPM-1

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.1.1, se puede observar que el modelo con el menor AIC es un modelo con medidas repetidas y matriz "Diagonal", además no ha ocurrido ninguna anomalía durante la ejecución. Sin embargo, para poder conseguir normalidad en los residuos fue necesario transformar la variable respuesta con la fórmula $\sqrt{\max(Efectividad) - Efectividad_{sujeto}}$

Método	AIC	Normalidad	Advertencia
Variance Components	484,782	0,200	Ninguna
Identity	487,938	0,200	Ninguna
Compound symmetry	484,782	0,200	Ninguna
Diagonal	487,219	0,200	Ninguna
Unstructured	480,414	0,200	Ninguna
Ar(1)	486,540	0,200	Ninguna

Tabla D.1.36 Modelo Mixto UPM-1

Las Tablas D.1.2 y D.1.3 muestran el resultado de todas las variables estudiadas en UPM04.

Variable	p-valor	niveles	Media	Media Transf.	Dirección efecto	p-valor	Tamaño Efecto
Técnica	0,000	R	6,440	58,526	R<E	0,000	0,52
		E	3,890	84,868	R<F	0,000	0,57
		F	3,249	89,444	E<F	0,671	0,02
Programa	0,130	Nt	5,026	74,739	Nt<Cm	0,289	0,08
		Cm	4,313	81,398	Nt<Na	0,204	0,06
		Na	4,241	82,014	Cm<Na	1,000	0,05
Grupo	0,439	G3	5,187	73,095	G3<G1	1,000	-0,29
					G3<G6	1,000	-0,12
					G3<G2	1,000	0,09
					G3<G5	1,000	-0,39
		G1	5,183	73,137	G3<G4	1,000	-0,16
					G1<G6	1,000	-0,21
					G1<G2	1,000	-0,15
					G1<G5	1,000	-0,04
		G6	4,741	77,523	G1<G4	1,000	-0,02
					G6<G2	1,000	-0,18
					G6<G5	1,000	-0,32
					G6<G4	1,000	-0,09
		G5	4,210	82,276	G2<G5	1,000	-0,05
					G2<G4	1,000	-0,06
		G4	3,465	87,994	G5<G4	1,000	-0,14

Tabla D.1.37 Resultados UPM-1 (1/2)

Técnica * Programa	0,025	R=>Na	7,239	47,597	R=>Na<Cm	1,000	0,13
		R=>Cm	6,757	54,343	R=>Na<Nt	0,027	0,32
		R=>Nt	5,324	71,655	R=>Cm<Nt	0,143	0,38
		E=>Nt	5,152	73,457	E=>Nt<Cm	0,230	0,20
		E=>Cm	3,571	87,248	E=>Nt<Na	0,047	0,19
		E=>Na	2,948	91,309	E=>Cm<Na	1,000	0,04
		F=>Nt	4,600	78,840	F=>Nt<Cm	0,169	0,39
		F=>Cm	2,609	93,193	F=>Nt<Na	0,169	0,25
		F=>Na	2,538	93,559	F=>Cm<Na	1,000	0,11
Programa * Técnica	0,025	Na=>R	7,239	47,597	Na=>R<E	0,000	0,68
		Na=>E	2,948	91,309	Na=>R<F	0,000	0,60
		Na=>F	2,538	93,559	Na=>E<F	1,000	0,01
		Cm=>R	6,757	54,343	Cm=>R<E	0,001	0,60
		Cm=>E	3,571	87,248	Cm=>R<F	0,000	0,74
		Cm=>F	2,609	93,193	Cm=>E<F	0,851	0,21
		Nt=>R	5,324	71,655	Nt=>R<E	1,000	0,32
		Nt=>E	5,152	73,457	Nt=>R<F	1,000	0,26
		Nt=>F	4,600	78,840	Nt=>E<F	1,000	0,02

Tabla D.1.38 Resultados UPM-1 (2/2)

Cuando existe significancia estadística (p-valor), este valor se representa en color rojo. Por otro lado, la significancia práctica (tamaño del efecto) solo se ha considerado cuando sus valores son Mediano (0,24-0,36) y Grande (0,37-1); estos valores aparecen marcados en gris claro y gris oscuro, respectivamente.

D.1.2 UPM-2

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.1.4, se puede observar que el modelo con el menor AIC es un modelo con medidas repetidas y matriz "Diagonal", además no ha ocurrido ninguna anomalía durante la ejecución.

Método	AIC	Normalidad	Advertencia
Variance Components	457,070	0,003	Matriz Hessiana no definida positivamente
Identity	455,070	0,003	Ninguna
Compound symmetry	456,217	0,003	Ninguna
Diagonal	452,911	0,002	Ninguna
Unstructured	456,958	0,003	Ninguna
Ar(1)	456,727	0,002	Ninguna

Tabla D.1.39 Modelo Mixto UPM-2

Para poder conseguir normalidad en los residuos fue necesario transformar la variable respuesta, sin embargo, después de varias transformaciones no se consiguió normalidad. Finalmente, los resultados se calcularon con la variable transformada con la fórmula $\sqrt{\max(Efectividad) - Efectividad_{sujeto}}$ porque con ella se consiguieron los mejores valores.

La Tabla D.1.5 muestra el resultado de todas las variables utilizadas en el análisis de UPM04, si existe significancia estadística (p-valor), este valor se representa en color rojo. Por otro lado, si existe significancia práctica (tamaño del efecto) estos valores aparecen marcados en gris claro (Mediano) y gris oscuro (Grande).

Variable	p-valor	niveles	Media	Media Transf.	Dirección efecto	p-valor	Tamaño efecto
Técnica	0,000	R	6,439	58,539	R<E	0,006	0,29
		E	4,555	79,252	R<F	0,000	0,43
		F	3,667	86,553	E<F	0,444	0,10
Programa	0,001	Na	5,565	69,031	Na<Cm	1,000	0,05
		Cm	5,414	70,689	Na<Nt	0,004	0,26
		Nt	3,682	86,443	Cm<Nt	0,007	0,29
Grupo	0,966	G6	5,511	69,629	G6<G3	1,000	0,02
					G6<G4	1,000	0,03
					G6<G5	1,000	0,17
		G3	5,029	74,709	G6<G1	1,000	0,05
					G6<G2	1,000	0,01
					G3<G4	1,000	0,13
		G4	4,766	77,285	G3<G5	1,000	0,20
					G3<G1	1,000	0,13
					G3<G2	1,000	0,14
		G5	4,708	77,835	G4<G5	1,000	0,06
					G4<G1	1,000	0,18
					G4<G2	1,000	0,12
Técnica * Programa	0,731	G1	4,662	78,266	G5<G1	1,000	0,06
		G2	4,645	78,424	G5<G2	1,000	0,03
		R=>Na	7,586	42,453	G1<G2	1,000	0,13
					R=>Na<Cm	1,000	0,19
					R=>Na<Nt	0,060	0,54
		R=>Cm	6,851	53,064	R=>Cm<Nt	0,262	0,31
		R=>Nt	4,878	76,205	E=>Na<Cm	1,000	0,18
		E=>Na	5,050	74,498	E=>Na<Nt	1,000	0,41
		E=>Cm	4,482	79,912	E=>Cm<Nt	1,000	0,13
Programa * Técnica	0,731	E=>Nt	4,132	82,927	F=>Cm<Na	1,000	0,44
		F=>Cm	4,908	75,912	F=>Cm<Nt	0,061	0,57
		F=>Na	4,059	83,525	F=>Na<Nt	0,285	0,50
		F=>Nt	2,033	95,867	Na=>R<E	0,174	0,64
		Na=>R	7,586	42,453	Na=>R<F	0,010	0,68
		Na=>E	5,050	74,498	Na=>E<F	1,000	0,52
		Na=>F	4,059	83,525	Cm=>R<F	0,322	0,28
		Cm=>R	6,851	53,064	Cm=>R<E	0,190	0,26
		Cm=>F	4,908	75,912	Cm=>F<E	1,000	0,15
		Cm=>E	4,482	79,912	Nt=>R<E	1,000	0,49
		Nt=>R	4,878	76,205	Nt=>R<F	0,067	0,63
		Nt=>E	4,132	82,927	Nt=>S<F	0,297	0,38
		Nt=>F	2,033	95,867			

Tabla D.1.40 Resultados UPM-2

D.1.3 UPV05

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion),

cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

La Tabla D.1.6, se puede observar que el modelo con el menor AIC es un modelo con medidas repetidas y matriz “Diagonal”, además no ha ocurrido ninguna anomalía durante la ejecución y los residuos cumplen normalidad.

Para poder conseguir normalidad en los residuos fue necesario transformar la variable respuesta con la fórmula $Efectividad_{sujeto}^2$.

Método	AIC	Normalidad	Advertencia
Variance Components	1164,998	0,200	Ninguna
Identity	1162,999	0,200	Ninguna
Compound symmetry	1164,998	0,200	Ninguna
Diagonal	1162,140	0,200	Ninguna
Unstructured	1164,138	0,200	Ninguna
Ar(1)	1164,998	0,200	Ninguna

Tabla D.1.41 Modelo Mixto UPV05

La Tabla D.1.7 muestra el resultado de todas las variables utilizadas en el análisis de UPV05, si existe significancia estadística (p-valor), este valor se representa en color rojo. Por otro lado, si existe significancia práctica (tamaño del efecto) estos valores aparecen marcados en gris claro (Mediano) y gris oscuro (Grande).

Variable	p-valor	niveles	Media	Media Transf.	Dirección efecto	p-valor	Tamaño efecto
Técnica	0,781	F E	5283,962 5465,918	72,691 73,932	F<E	0,781	0,00
Programa	0,062	Cm	4319,625	65,724	Cm<Nt	0,323	0,09
		Nt	5611,317	74,909	Cm<Na	0,066	0,48
		Na	6193,878	78,701	Nt<Na	1,000	0,25
Técnica * Programa	0,039	F=>Cm	3098,639	55,665	F=>Cm<Nt	0,026	0,52
		F=>Nt	5732,839	75,716	F=>Cm<Na	0,001	0,69
		F=>Na	7020,408	83,788	F=>Nt<Na	0,538	0,31
		E=>Na	5367,347	73,262	E=>Na<Nt	1,000	0,01
		E=>Nt	5489,796	74,093	E=>Na<Cm	1,000	0,01
		E=>Cm	5540,610	74,435	E=>Nt<Cm	1,000	0,06
Programa * Técnica	0,039	Cm=>F	3098,639	55,665	Cm=>F<E	0,032	0,43
		Cm=>E	5540,610	74,435			
		Na=>E	5367,347	73,262	Na=>E<F	0,154	0,29
		Na=>F	7020,408	83,788			
		Nt=>E	5489,796	74,093	Nt=>E<F	0,830	0,03
		Nt=>F	5732,839	75,716			

Tabla D.1.42 Resultados UPV05

D.1.4 ORT05

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

La Tabla D.1.8, se puede observar que el modelo con el menor AIC es un modelo sin medidas repetidas y cuya estructura de covarianza es “Componentes de la varianza”, además no ha ocurrido ninguna anomalía durante la ejecución y los residuos cumplen normalidad.

Método	AIC	Normalidad	Advertencia
Variance Components	754,416	0,200	Ninguna
Identity	757,335	0,099	Ninguna
Compound symmetry	754,416	0,200	Ninguna
Diagonal	758,456	0,200	Ninguna
Unstructured	755,424	0,200	Ninguna
Ar(1)	754,416	0,200	Ninguna

Tabla D.1.43 Resultados ORT05

Para poder conseguir normalidad en los residuos fue necesario transformar la variable respuesta con la fórmula $\sqrt{\max(Efectividad) - Efectividad_{sujeto}}$

La Tabla D.1.9 muestra el resultado de todas las variables utilizadas en el análisis de ORT05, si existe significancia estadística (p-valor), este valor se representa en color rojo. Por otro lado, si existe significancia práctica (tamaño del efecto) estos valores aparecen marcados en gris claro (Mediano) y gris oscuro (Grande).

Variable	p-valor	niveles	Media	Media Transf.	Dirección efecto	p-valor	Tamaño efecto
Técnica	0,004	E F	4,969 3,821	75,309 85,400	E<F	0,004	0,16
Programa	0,000	Cm Na	5,357 3,433	71,303 88,215	Cm<Na	0,000	0,24
Grupo	0,979	G3	4,679	78,107	G3<G4	1,000	0,06
					G3<G2	1,000	0,14
					G3<G1	1,000	0,07
		G4	4,665	78,238	G4<G2	1,000	0,12
					G4<G1	1,000	0,02
					G2<G1	1,000	0,08
Periodo	0,007	S1	4,938	75,616	S1<S2	0,007	-0,14
		S2	3,852	85,162			
Técnica * Programa	0,994	E=>Cm	5,937	64,752	E=>Cm<Na	0,236	0,12
		E=>Na	4,001	83,992			
		F=>Cm	4,777	77,180	F=>Cm<Na	0,251	0,33
		F=>Na	2,864	91,798			
Programa * Técnica	0,994	Cm=>E	5,937	64,752	Cm=>E<F	0,486	0,24
		Cm=>F	4,777	77,180			
		Na=>E	4,001	83,992	Na=>E<F	0,485	0,10
		Na=>F	2,864	91,798			
Técnica * Periodo	0,423	E=>P1	5,714	67,350	E=>P1<P2	0,021	-0,28
		E=>P2	4,224	82,158			
		F=>P1	4,161	82,686	F=>P1<P2	0,283	-0,11
		F=>P2	3,480	87,890			
Periodo * Técnica	0,423	P1=>E	5,714	67,350	P1=>E<F	0,014	-0,25
		P1=>F	4,161	82,686			
		P2=>E	4,224	82,158	P2=>E<F	0,245	-0,03
		P2=>F	3,480	87,890			

Tabla D.1.44 Resultados ORT05

D.2 Experimentos de Test Driven Development (TDD)

El principal objetivo de este paso es conocer el comportamiento del enfoque de desarrollo y factores secundarios, en caso de haberlos, en cada uno de los experimentos.

D.2.1 FSecure Helsinki

De acuerdo con el diseño de este experimento, el método de análisis utilizado es ANOVA de medidas repetidas. Para que este modelo sea fiable, se tiene que comprobar la esfericidad en la matriz de varianzas-covarianzas. En caso de no poder asumir esfericidad se elegirá el estadístico de “Huyn-Feldt” porque es menos conservador.

En la Tabla D.2.1, se puede observar que TDD1 se comporta significativamente mejor que TDD2. Por otro lado, teniendo en cuenta la significancia práctica, TLD consigue una mejora mínima con respecto a TDD2; y TDD1 con respecto a TLD.

Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,009	TDD2	20,327	TDD2<TLD	0,196	-0,58
		TLD	34,082	TDD2<TDD1	0,027	-0,58
		TDD1	45,237	TLD<TDD1	0,493	-0,45

Tabla D.2.45 Resultados FSecure Helsinki

D.2.2 FSecure Kuala Lumpur

De acuerdo con el diseño de este experimento, el método de análisis utilizado es ANOVA de medidas repetidas. Para que este modelo sea fiable, se tiene que comprobar la esfericidad en la matriz de varianzas-covarianzas. En caso de no poder asumir esfericidad se elegirá el estadístico de “Huyn-Feldt” porque es menos conservador.

En la Tabla D.2.2, se puede observar que el enfoque de desarrollo no es significativo. Sin embargo, teniendo en cuenta la significancia práctica, TDD1 consigue una mejora mínima con respecto a TDD2 y a TLD.

Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,073	TDD2	21,434	TDD2<TLD	1,000	-0,19
		TLD	27,169	TDD2<TDD1	0,200	-0,42
		TDD1	43,550	TLD<TDD1	0,423	-0,32

Tabla D.2.46 Resultados FSecure Kuala Lumpur

D.2.3 FSecure Oulu

De acuerdo con el diseño de este experimento, el método de análisis utilizado es ANOVA de medidas repetidas. Para que este modelo sea fiable, se tiene que comprobar la esfericidad

en la matriz de varianzas-covarianzas. En caso de no poder asumir esfericidad se elegirá el estadístico de “Huyn-Feldt” porque es menos conservador.

En la Tabla D.2.3, se puede observar que TDD1 se comporta significativamente mejor que TLD y TDD2; y que TLD se comporta mejor que TDD2.

Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,000	TLD	20,226	TLD<TDD2	0,011	-0,62
		TDD2	47,967	TLD<TDD1	0,004	-0,63
		TDD1	73,723	TDD2<TDD1	0,042	-0,59

Tabla D.2.47 Resultados FSecure Oulu

D.2.4 Electrobit

De acuerdo con el diseño de este experimento, el método de análisis utilizado es ANOVA de medidas repetidas. Para que este modelo sea fiable, se tiene que comprobar la esfericidad en la matriz de varianzas-covarianzas. En caso de no poder asumir esfericidad se elegirá el estadístico de “Huyn-Feldt” porque es menos conservador.

En la Tabla D.2.4, se puede observar que TDD1 se comporta significativamente mejor que TDD2 y TLD, por otro lado, teniendo en cuenta la significancia práctica TLD consigue una mejora mínima con respecto a TDD2.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño del efecto
Enfoque de desarrollo	0,000	TDD2	2,468	TDD2<TLD	0,209	-0,35
		TLD	15,453	TDD2<TDD1	0,000	-0,63
		TDD1	43,850	TLD<TDD1	0,040	-0,54

Tabla D.2.48 Resultados Electrobit

D.2.5 OY1

De acuerdo con el diseño de este experimento, el método de análisis a utilizar sería Modelo Lineal General Univariante, sin embargo, no se cumplen la restricción de normalidad en los residuos. Por lo tanto, el método utilizado es Prueba No Paramétrica para muestras independientes.

En la Tabla D.2.5, se puede observar que el enfoque de desarrollo no es significativo y tampoco existe significancia práctica.

Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,795	TDD TLD	39,347 43,862	TDD<TLD	0,795	-0,04

Tabla D.2.49 Resultados OY1

D.2.6 UPM

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike’s information criterion),

cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.6, se puede observar que el modelo con el menor AIC es un modelo con medidas repetidas y matriz “Sin estructura”, sin embargo, para este modelo hay una advertencia en la ejecución. El siguiente modelo con menor AIC es un modelo con medidas repetidas y matriz “AR1”, pero en este modelo no hay normalidad en los residuos. Finalmente, el modelo elegido es un modelo sin medidas repetidas y cuya estructura de covarianza es “Componentes de la varianza”, porque es el siguiente modelo con menor AIC, los residuos cumplen normalidad y no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	531,253	0,200	Ninguna
Identity	539,586	0,007	Ninguna
Compound symmetry	531,253	0,023	Ninguna
Diagonal	542,617	0,007	Ninguna
Unstructured	459,750	0,200	La convergencia no se ha conseguido
Ar(1)	531,210	0,014	Ninguna

Tabla D.2.50 Modelo Mixto UPM

La Tabla D.2.7 y la Tabla D.2.8, muestra los resultados del modelo mixto, en ella se puede observar que únicamente la tarea es significativa, y que BSK_S y Sudoku_NS se comportan mejor que MR_NS y MF_NS. Si además se tiene en cuenta la significancia práctica, BSK_S consigue una mejora mínima con respecto a Sudoku_NS; y MF_NS con respecto a MR_NS.

Aunque la interacción entre enfoque de desarrollo y tarea no es significativa, para TDD1 si que hay significación estadística, concretamente, BSK_S se comporta mejor que MR_NS y MF_NS, si además observamos la significancia práctica Sudoku_NS consigue una mejora mínima con respecto a MF_NS; y BSK_S con respecto a Sudoku_NS. Para TDD2, también hay significación estadística, concretamente, BSK_S se comporta mejor que MR_NS y MF_NS y Sudoku_NS se comporta mejor que MR_NS, si además se considera la significancia práctica, MF_NS consigue una mejora mínima con respecto a MR_NS; y BSK_S con respecto a Sudoku_NS.

Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,242	TDD1	32,127	TDD1<TLD1	1,000	-0,01
				TDD1<TDD2	1,000	-0,14
		TLD1	32,297	TDD1<TLD2	0,440	-0,11
				TLD1<TDD2	1,000	-0,16
		TDD2	35,889	TLD1<TLD2	0,432	-0,18
		TLD2	43,502	TDD2<TLD2	1,000	-0,00
Tarea	0,000	MR_NS	10,585	MR_NS<MF_NS	0,118	-0,44
				MR_NS<Sudoku_NS	0,000	-0,61
		MF_NS	25,106	MR_NS<BSK_S	0,000	-0,62
				MF_NS<Sudoku_NS	0,002	-0,54
		Sudoku_NS	49,970	MF_NS<BSK_S	0,000	-0,62
		BSK_S	58,155	Sudoku_NS<BSK_S	1,000	-0,32

Tabla D.2.51 Resultados UPM (1/2)

Para ITL1, solo se puede considerar la significancia práctica y, por lo tanto, tanto BSK_S como Sudoku_NS consiguen una mejora mínima con respecto a MF_NS y MR_NS; y MF_NS con

respecto a MR_NS. Finalmente, para TLD2, hay significancia estadística y MR_NS se comporta peor que BSK_S y Sudoku_NS, además teniendo en cuenta la significancia práctica, MF_NS consigue una mejora mínima con respecto a MR_NS; y Sudoku_NS con respecto a MF_NS.

Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Enfoque de desarrollo * Tarea	0,690	TDD1=>MR_NS	7,736	TDD1=>MR_NS<MF_NS	1,000	-0,22
				TDD1=>MR_NS<Sudoku_NS	0,069	-0,00
		TDD1=>MF_NS	19,682	TDD1=>MR_NS<BSK_S	0,001	-0,65
				TDD1=>MF_NS<Sudoku_NS	0,751	-0,67
		TDD1=>Sudoku_NS	40,291	TDD1=>MF_NS<BSK_S	0,033	-0,65
				TDD1=>Sudoku_NS<BSK_S	0,758	-0,67
		TDD1=>BSK_S	60,801	TDD2=>MR_NS<MF_NS	1,000	-0,67
				TDD2=>MR_NS<Sudoku_NS	0,011	-0,65
		TDD2=>MR_NS	9,889	TDD2=>MR_NS<BSK_S	0,000	-0,65
				TDD2=>MF_NS<Sudoku_NS	0,151	-0,22
		TDD2=>MF_NS	15,083	TDD2=>MF_NS<BSK_S	0,006	-0,67
				TDD2=>Sudoku_NS<BSK_S	,0866	-0,26
		TDD2=>Sudoku_NS	50,193	TLD1=>MR_NS<MF_NS	0,880	-0,65
				TLD1=>MR_NS<BSK_S	0,061	-0,65
		TDD2=>BSK_S	68,392	TLD1=>MR_NS<Sudoku_NS	0,059	-0,65
				TLD1=>MF_NS<BSK_S	1,000	-0,26
		TLD1=>MR_NS	10,555	TLD1=>MF_NS<Sudoku_NS	1,000	-0,44
		TLD1=>MF_NS	29,037	TLD1=>BSK_S<Sudoku_NS	1,000	-0,22
		TLD1=>BSK_S	42,918	TLD2=>MR_NS<MF_NS	0,498	-0,44
		TLD1=>Sudoku_NS	46,667	TLD2=>MR_NS<BSK_S	0,012	-0,67
Tarea * Enfoque de desarrollo	0,690	TLD2=>MR_NS	14,162	TLD2=>MR_NS<Sudoku_NS	0,023	0,00
		TLD2=>MF_NS	36,621	TLD2=>MF_NS<BSK_S	0,384	0,00
		TLD2=>BSK_S	60,507	TLD2=>MF_NS<Sudoku_NS	0,490	-0,67
		TLD2=>Sudoku_NS	62,720	TLD2=>BSK_S<Sudoku_NS	1,000	-0,22
		MR_NS=>TDD1	7,736	MR_NS=>TDD1<TDD2	1,000	-0,26
				MR_NS=>TDD1<TLD1	1,000	-0,57
		MR_NS=>TDD2	9,889	MR_NS=>TDD1<TLD2	1,000	-0,22
				MR_NS=>TDD2<TLD1	1,000	-0,16
		MR_NS=>TLD1	10,555	MR_NS=>TDD2<TLD2	1,000	0,00
				MR_NS=>TLD1<TLD2	1,000	-0,67
		MR_NS=>TLD2	14,162	MF_NS=>TDD2<TDD1	1,000	-0,50
				MF_NS=>TDD2<TLD1	1,000	-0,67
		MF_NS=>TDD2	15,083	MF_NS=>TDD2<TLD2	0,897	-0,67
				MF_NS=>TDD1<TLD1	1,000	-0,44
		MF_NS=>TDD1	19,682	MF_NS=>TDD1<TLD2	1,000	-0,44
				MF_NS=>TLD1<TLD2	1,000	-0,26
		MF_NS=>TLD1	29,037	BSK_S=>TLD1<TLD2	1,000	-0,18
				BSK_S=>TLD1<TDD1	1,000	-0,18
		MF_NS=>TLD2	26,621	BSK_S=>TLD1<TDD2	0,223	-0,57
				BSK_S=>TLD2<TDD1	1,000	0,00
		BSK_S=>TLD1	42,918	BSK_S=>TLD2<TDD2	1,000	-0,18
		BSK_S=>TLD2	60,507	BSK_S=>TDD1<TDD2	1,000	-0,65
		BSK_S=>TDD1	60,801	Sudoku_NS=>TDD1<TLD1	1,000	-0,50
		BSK_S=>TDD2	68,392	Sudoku_NS=>TDD1<TDD2	1,000	-0,18
		Sudoku_NS=>TDD1	40,291	Sudoku_NS=>TDD1<TLD2	0,870	0,00
		Sudoku_NS=>TLD1	46,667	Sudoku_NS=>TLD1<TDD2	1,000	0,00
		Sudoku_NS=>TDD2	50,193	Sudoku_NS=>TLD1<TLD2	1,000	-0,22
		Sudoku_NS=>TLD2	62,720	Sudoku_NS=>TDD2<TLD2	1,000	-0,22

Tabla D.2.52 Resultados UPM

Finalmente, considerando la significancia práctica, hay diferencia en la interacción de la tarea con el enfoque de desarrollo, para todas las tareas. Para MR_NS, tanto TDD2 como TLD1 consiguen una mejora mínima con respecto a TDD1; y TLD2 con respecto a TLD1. Para MF_NS, tanto TLD1 como TLD2 consiguen una mejora mínima con respecto a TDD1 y TDD2; y TDD1 con respecto a TDD2. Para BSK_S, TDD2 consigue una mejora mínima con respecto a TLD1 y TDD1. Por último, para Sudoku_NS, TLD1 consigue una mejora mínima con respecto a TDD1.

D.2.7 Playtech

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.9, se puede observar que hay tres modelos con el menor AIC, sin embargo, se tomarán los resultados del modelo sin medidas repetidas y cuya estructura de covarianza es "Componentes de la varianza", se puede comprobar que los residuos cumplen normalidad y que no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	347,052	0,063	Ninguna
Identity	347,731	0,134	Ninguna
Compound symmetry	347,052	0,134	Ninguna
Diagonal	349,676	0,134	Ninguna
Unstructured	348,988	0,134	Ninguna
Ar(1)	347,052	0,134	Ninguna

Tabla D.2.53 Modelo Mixto PlayTech

La Tabla D.2.10, muestra los resultados del modelo mixto, en ella se puede observar que el "Enfoque de desarrollo" y la "Tarea" son significativos.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,044	TDD ITL	43,552 64,002	TDD<ITL	0,044	-0,34
Tarea	0,030	MR_NS BSK_S	42,634 64,919	MR_NS<BSK_S	0,030	-0,32
Enfoque de desarrollo * Tarea	0,354	ITL=>MR_NS ITL=>BSK_S TDD=>MR_NS TDD=>BSK_S	59,550 68,453 25,719 61,385	ITL=>MR_NS<BSK_S TDD=>MR_NS<BSK_S	0,602 0,043	-0,25 -0,32
Tarea * Enfoque de desarrollo	0,354	MR_NS=>TDD MR_NS=>ITL BSK_S=>TDD BSK_S=>ITL	25,719 59,550 61,385 68,453	MR_NS=>TDD<ITL BSK_S=>TDD<ITL	0,054 0,678	-0,35 -0,25

Tabla D.2.54 Resultados PlayTech

Con respecto al enfoque de desarrollo se puede observar, que ITL se comporta mejor que TDD. Por otro lado, para la tarea, BSK_S se comporta mejor que MR_NS.

Se puede observar también, que, aunque la interacción entre enfoque de desarrollo y tarea no es significativa, para el caso de TDD si que hay significación estadística, concretamente, BSK_S se comporta mejor que MR_NS. Además, teniendo en cuenta la significancia práctica, para el caso de ITL, BSK_S consigue una mejora mínima con respecto a MR_NS.

Finalmente, considerando la significancia práctica, hay diferencia en la interacción de la tarea con el enfoque de desarrollo, tanto para MR_NS, como para BSK_S, ITL consigue una mejora mínima con respecto a TDD.

D.2.8 UPV

De acuerdo con el diseño de este experimento, el método de análisis utilizado es ANOVA de medidas repetidas. Para que este modelo sea fiable, se tiene que comprobar la esfericidad en la matriz de varianzas-covarianzas. En caso de no poder asumir esfericidad se elegirá el estadístico de “Huyn-Feldt” porque es menos conservador.

La Tabla D.2.11, muestra que TDD se comporta significativamente mejor que TLD.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño del efecto
Enfoque de desarrollo	0,000	TLD TDD	33,243 77,156	TLD<TDD	0,000	-0,48

Tabla D.2.55 Resultados UPV

D.2.9 Ericsson

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike’s information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.12, se puede observar que el modelo con menor AIC es un modelo con medidas repetidas y matriz “sin estructura”, además se comprueba que los residuos cumplen normalidad y no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	393,780	0,258	Ninguna
Identity	394,879	0,074	Ninguna
Compound symmetry	393,780	0,113	Ninguna
Diagonal	392,726	0,074	Ninguna
Unstructured	391,497	0,114	Ninguna
Ar(1)	393,780	0,113	Ninguna

Tabla D.2.56 Modelo Mixto Ericsson

La Tabla D.2.13, muestra los resultados del modelo mixto, en ella se puede observar que el “Enfoque de desarrollo”, la “Tarea” y la interacción de los dos son significativos. Al ser la interacción significativa, no hace falta evaluar los resultados de los factores.

Se puede observar, que existe significancia estadística en la interacción entre el enfoque de desarrollo y la tarea, concretamente cuando los participantes aplican TDD, MR_NS se

comporta peor que las demás tareas y BSK_S se comporta mejor que BSK_NS, finalmente MR_S se comporta igual que BSK_NS y BSK_NS. Sin embargo, de acuerdo con la significancia práctica, se puede observar que MR_S se comporta igual que BSK_NS, pero si que hay diferencia con BSK_S, consiguiendo este último mejora mínima.

Por otro lado, aunque para ITL no hay significancia estadística, de acuerdo con la significancia práctica se puede observar que MR_NS consigue una mejora mínima con respecto a BSK_NS.

Para la interacción entre la tarea y el enfoque de desarrollo se puede observar que a excepción de MR_S, en todas las demás tareas hay diferencia al aplicar ITL y TDD, concretamente, para MR_NS, ITL se comporta mejor que TDD, sin embargo, tanto para BSK_NS como BSK_S, TDD se comporta mejor que ITL.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,013	ITL TDD	32,671 51,690	ITL<TDD	0,013	-0,27
Tarea	0,051	MR BSK	34,886 49,475	MR<BSK	0,051	-0,17
Granularidad	0,023	NS S	32,558 51,802	NS< S	0,023	-0,22
Enfoque de desarrollo * Tarea	0,023	ITL=>BSK	27,179	ITL=>BSK<MR	0,466	-0,19
		ITL=>MR	38,162			
		TDD=>MR	31,609	TDD=>MR_S<BSK_S	0,000	-0,53
		TDD=>BSK	71,770			
Tarea * Enfoque de desarrollo	0,023	BSK=>ITL	27,179	BSK=>ITL<TDD	0,001	-0,48
		BSK=>TDD	71,770			
		MR=>TDD	31,609	MR=>TDD<ITL	0,600	-0,06
		MR=>ITL	38,162			
Enfoque de desarrollo * Granularidad	0,010	TDD=>NS	31,036	TDD=>NS<S	0,000	-0,45
		TDD=>S	72,343			
		ITL=>S	31,261	ITL=>S<NS	0,837	-0,04
		ITL=>NS	34,081			
Granularidad * Enfoque de desarrollo	0,010	NS=>TDD	31,036	NS=>TDD<ITL	0,763	-0,06
		NS=>ITL	34,081			
		S=>ITL	31,261	S=>ITL<TDD	0,001	-0,49
		S=>TDD	72,343			
Tarea * Granularidad	0,850	MR=>NS	19,009	MR=>NS<S	0,106	-0,40
		MR=>S	24,790			
		BSK=>NS	55,923	BSK=>NS<S	0,068	-0,26
		BSK=>S	89,681			
Tarea * Granularidad	0,850	NS=>MR	19,009	NS=>MR<BSK	0,178	-0,13
		NS=>BSK	55,923			
		S=>MR	24,790	S=>MR<BSK	0,143	-0,17
		S=>BSK	89,681			

Tabla D.2.57 Resultados Ericsson

D.2.10 UB1

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion),

cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.14, se puede observar que el modelo con el menor AIC es un modelo con medidas repetidas y matriz “Diagonal”, se puede comprobar que los residuos cumplen normalidad y que no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	377,172	0,059	Ninguna
Identity	376,742	0,118	Ninguna
Compound symmetry	377,172	0,118	Ninguna
Diagonal	376,076	0,118	Ninguna
Unstructured	376,272	0,118	Ninguna
Ar(1)	377,172	0,118	Ninguna

Tabla D.2.58 Modelo Mixto UB1

La Tabla D.2.15, muestra los resultados del modelo mixto, en ella se puede observar que únicamente la tarea es significativa, y que BSK_S se comportan mejor que MR_NS.

Aunque la interacción entre enfoque de desarrollo y tarea no es significativa, basándonos en la significancia práctica, se puede comprobar que, BSK_S consigue una mejora mínima con respecto a MR_NS independientemente del enfoque de desarrollo utilizado.

Finalmente, aunque la interacción entre tarea y enfoque de desarrollo no es significativa, basándonos en la significancia práctica, se puede comprobar que para BSK_S, TDD consigue una mejora mínima con respecto a TLD.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,313	TLD TDD	35,825 43,539	TLD<TDD	0,313	-0,03
Tarea	0,000	MR_S BSK_S	17,222 62,142	MR_S<BSK_S	0,000	-0,58
Enfoque de desarrollo * Tarea	0,730	TLD=>MR_S	12,052	TLD=>MR_S<BSK_S	0,000	-0,63
		TLD=>BSK_S	59,597			
		TDD=>MR_S	22,391	TDD=>MR_NS<BSK_S	0,003	-0,49
		TDD=>BSK_S	64,687			
Tarea * Enfoque de desarrollo	0,730	MR_S=>TLD	12,052	MR_NS=>TLD<TDD	0,329	-0,10
		MR_S=>TDD	22,391			
		BSK_S=>TLD	59,597	BSK_S=>TLD<TDD	0,642	-,024
		BSK_S=>TDD	64,687			

Tabla D.2.59 Resultados UB1

D.2.11 USD-UG

De acuerdo con el diseño de este experimento, el método de análisis utilizado es Modelo Lineal General Univariante. Las restricciones de este modelo son: normalidad en los residuos y homogeneidad en las varianzas de los residuos.

En la Tabla D.2.16, se puede observar que únicamente la tarea es significativa, concretamente, BSK_NS se comporta significativamente mejor que BSK_S. Por otro lado,

atendiendo a la significancia práctica se puede observar que, en la interacción entre el enfoque de desarrollo y la tarea, para el caso de TDD, BSK_S consigue una mejora mínima con respecto a BSK_NS. Además, en la interacción entre la tarea y el enfoque de desarrollo, para el caso de BSK_S, TDD consigue una mejora mínima con respecto a TLD.

Variable	p-valor	niveles	Media	Media Transf.	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,430	TLD TDD	5,662 6,066	32,058 36,796	TDD<TLD	0,430	-0,09
Tarea	0,021	BSK_NS BSK_S	5,260 6,468	27,668 41,835	BSK_S<BSK_NS	0,021	-0,25
Enfoque de desarrollo * Tarea	0,080	TDD=>BSK_NS	5,009	25,090	TDD=>BSK_NS<BSK_S	0,004	-0,52
		TDD=>BSK_S	7,123	50,737			
		TLD=>BSK_NS	5,511	30,371	TLD=>BSK_NS<BSK_S	0,678	-0,03
		TLD=>BSK_S	5,813	33,791			
Tarea * Enfoque de desarrollo	0,080	BSK_NS=>TDD	5,009	25,090	BSK_NS=>TDD<TLD	0,492	-0,19
		BSK_NS=>TLD	5,511	30,371			
		BSK_S=>TLD	5,813	33,791	BSK_S=>TLD<TDD	0,072	-0,36
		BSK_S=>TDD	7,123	50,737			

Tabla D.2.60 Resultados USD-UG

D.2.12 USD-G

De acuerdo con el diseño de este experimento, el método de análisis utilizado es Modelo Lineal General Univariante. Las restricciones de este modelo son: normalidad en los residuos y homogeneidad en las varianzas de los residuos.

En la Tabla D.2.17, se puede observar que para ninguno de los factores bajo estudio hay significancia estadística. Sin embargo, atendiendo a la significancia práctica se puede observar que, en la interacción entre el enfoque de desarrollo y la tarea, para el caso de TDD, BSK_S consigue una mejora mínima con respecto a BSK_NS. Además, en la interacción entre la tarea y el enfoque de desarrollo, para el caso de BSK_NS, TLD consigue una mejora mínima con respecto a TDD.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,782	TDD TLD	72,852 77,083	TDD<TLD	0,782	-0,19
Tarea	0,915	BSK_S BSK_NS	74,154 75,781	BSK_S<BSK_NS	0,915	-0,16
Enfoque de desarrollo * Tarea	0,308	TDD=>BSK_NS	65,625	TDD=>BSK_NS<BSK_S	0,467	-0,27
		TDD=>BSK_S	80,078			
		TLD=>BSK_S	68,229	TLD=>BSK_S<BSK_NS	0,457	-0,00
		TLD=>BSK_NS	85,938			
Tarea * Enfoque de desarrollo	0,308	BSK_NS=>TDD	65,625	BSK_NS=>TDD<TLD	0,396	-0,77
		BSK_NS=>TLD	85,938			
		BSK_S=>TLD	68,229	BSK_S=>TLD<TDD	0,549	-0,07
		BSK_S=>TDD	80,078			

Tabla D.2.61 Resultados USD-G

D.2.13 USD2

De acuerdo con el diseño de este experimento, el método de análisis a utilizar sería Modelo Lineal General Univariante, sin embargo, no se cumplen la restricción de normalidad en los residuos. Por lo que fue necesario utilizar la Prueba No Paramétrica para muestras independientes Kruskal-Wallis. Para aplicar esta técnica estadística, utilizamos una única variable que está compuesta por el enfoque de desarrollo junto con la tarea, es decir, tiene 4 niveles.

En la Tabla D.2.18, se puede observar que el enfoque de desarrollo con la tarea es significativo. Concretamente, cuando los sujetos aplican ITL, BSK_NS se comporta peor que BSK_S; por otro lado, cuando los sujetos utilizan BSK_S, consiguen mejores resultados con ITL que con TDD. Finalmente, en estos dos casos también existe significancia práctica.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo con Tarea	0,008	ITL-BSK_NS	26,50	ITL-BSK_NS<TDD-BSK_NS	0,612	-0,09
		TDD-BSK_S	26,62	ITL-BSK_NS<ITL-BSK_S	0,003	-0,49
		TDD-BSK_NS	29,70	TDD-BSK_S<TDD-BSK_NS	0,648	-0,11
		ITL-BSK_S	44,11	TDD-BSK_S<ITL-BSK_S	0,006	-0,45

Tabla D.2.62 Resultados USD

D.2.14 PAF

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.19, se puede observar que el modelo con el menor AIC es un modelo con medidas repetidas y matriz "Identidad", además se comprueba que los residuos cumplen normalidad y no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	221,619	0,483	Matriz Hessiana no está definida positivamente
Identity	219,619	0,483	Ninguna
Compound symmetry	221,310	0,436	Ninguna
Diagonal	222,632	0,483	Ninguna
Unstructured	223,275	0,528	Ninguna
Ar(1)	221,079	0,450	Ninguna

Tabla D.2.63 Modelo Mixto PAF

La Tabla D.2.20, muestra los resultados del modelo mixto, en ella se puede observar que ninguno de los elementos del análisis es significativo. Sin embargo, aunque la interacción entre enfoque de desarrollo y tarea no es significativa, para el caso de YW si que hay significación estadística, concretamente, BSK_NS se comporta mejor que SS_NS. Además, existe significancia práctica para ITL, es decir, MR_NS consigue una mejora mínima con respecto a SS_NS.

Finalmente, considerando la significancia práctica, hay diferencia en la interacción de la tarea con el enfoque de desarrollo para MR_NS, concretamente, TDD consigue una mejora mínima con respecto a ITL.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,793	ITL	50,425	ITL<YW	1,000	-0,11
		YW	53,656	ITL<TDD	0,439	-0,09
		TDD	72,219	YW<TDD	0,639	-0,21
Tarea	0,128	SS_NS	38,850	SS_NS<BSK_NS	0,152	-0,09
		BSK_NS	65,625	SS_NS<MR_NS	0,098	-0,39
		MR_NS	71,825	BSK_NS<MR_NS	1,000	0,33
Enfoque de desarrollo * Tarea	0,150	YW=>SS_NS	34,545	YW=>SS_NS<BSK_NS	0,049	-0,52
		YW=>BSK_NS	72,767			
		ITL=>SS_NS	43,155	ITL=>SS_NS<MR_NS	0,436	-0,26
		ITL=>MR_NS	57,695			
		TDD=>BSK_NS	58,483	TDD=>BSK_NS<MR_NS	0,234	-0,22
Tarea * Enfoque de desarrollo	0,354	TDD=>MR_NS	85,955			
		SS_NS=>YW	34,545	SS_NS=>YW<ITL	0,643	-0,22
		SS_NS=>ITL	43,155			
		MR_NS=>ITL	57,695	MR_NS=>ITL<TDD	0,221	-0,26
		MR_NS=>TDD	85,955			
		BSK_NS=>YW	58,483	BSK_NS=>YW<TDD	0,444	-0,19
		BSK_NS=>TDD	72,767			

Tabla D.2.64 Resultados PAF

D.2.15 ETAPA

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.21, se puede observar que el modelo con el menor AIC es un modelo con medidas repetidas y matriz "AR1", sin embargo, en este modelo no hay normalidad en los residuos. Por lo tanto, el modelo elegido es un modelo sin medidas repetidas y cuya estructura de covarianza es "Componentes de la varianza", porque es el siguiente modelo con menor AIC, los residuos cumplen normalidad y no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	302,293	0,818	Ninguna
Identity	311,056	0,023	Ninguna
Compound symmetry	302,293	0,002	Ninguna
Diagonal	314,329	0,023	Ninguna
Unstructured	306,418	0,002	Ninguna
Ar(1)	302,020	0,004	Ninguna

Tabla D.2.65 Resultados ETAPA

La Tabla D.2.22, muestra los resultados del modelo mixto, en ella se puede observar que el "Enfoque de desarrollo" y la interacción con la "Tarea" son significativos. Al ser la interacción significativa, no hace falta evaluar los resultados del enfoque de desarrollo.

Con respecto a la interacción entre el enfoque de desarrollo y la tarea, se puede observar que para YW, MR_NS se comporta significativamente mejor que BSK_NS, sin embargo, teniendo en cuenta la significancia práctica SS_NS consigue una mejora mínima con respecto a BSK_NS.

Aunque no hay significancia estadística, la significancia práctica indica que, en la interacción, para el caso de TDD, tanto SS_NS como MR_NS consiguen una mejora mínima con respecto a BSK_NS. Mientras que para TLD ocurre justamente lo contrario, es decir, BSK_NS consigue una mejora mínima con respecto a SS_NS y a MR_NS.

Finalmente, con respecto a la interacción entre la tarea y el enfoque de desarrollo, se puede observar que para BSK_NS, ITL se comporta significativamente mejor que YW y TDD. Por otro lado, aunque no hay significancia estadística, la significancia práctica indica que, para el caso de MR_NS, YW consigue una mejora mínima con respecto a ITL.

Variable	p-valor	niveles	Media	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,017	TDD	9,801	TDD<YW	0,347	-0,24
		YW	16,815	TDD<ITL	0,015	-0,29
		ITL	23,210	YW<ITL	0,420	-0,49
Tarea	0,388	SS_NS	14,498	SS_NS<BSK_NS	1,000	-0,13
		BSK_NS	15,244	SS_NS<MR_NS	0,617	-0,06
		MR_NS	20,085	BSK_NS<MR_NS	0,812	0,06
Enfoque de desarrollo * Tarea	0,033	YW=>BSK_NS	1,590	YW=>BSK_NS<SS_NS	0,215	-0,65
		YW=>SS_NS	19,490	YW=>BSK_NS<MR_NS	0,043	-0,55
		YW=>MR_NS	29,365	YW=>SS_NS<MR_NS	1,000	-0,22
		TDD=>BSK_NS	7,598	TDD=>BSK_NS<SS_NS	1,000	-0,44
		TDD=>SS_NS	9,702	TDD=>BSK_NS<MR_NS	1,000	-0,50
		TDD=>MR_NS	12,103	TDD=>SS_NS<MR_NS	1,000	-0,22
		ITL=>SS_NS	14,301	ITL=>SS_NS<MR_NS	1,000	-0,22
		ITL=>MR_NS	18,786	ITL=>SS_NS<BSK_NS	0,082	-0,65
		ITL=>BSK_NS	36,544	ITL=>MR_NS<BSK_NS	0,325	-0,44
Tarea * Enfoque de desarrollo	0,033	BSK_NS=>YW	1,590	BSK_NS=>YW<TDD	1,000	0,00
		BSK_NS=>TDD	7,598	BSK_NS=>YW<ITL	0,003	-0,65
		BSK_NS=>ITL	36,544	BSK_NS=>TDD<ITL	0,033	-0,65
		SS_NS=>TDD	9,702	SS_NS=>TDD<ITL	1,000	-0,18
		SS_NS=>ITL	14,301	SS_NS=>TDD<YW	1,000	0,00
		SS_NS=>YW	19,490	SS_NS=>ITL<YW	1,000	-0,13
		MR_NS=>TDD	12,103	MR_NS=>TDD<ITL	1,000	-0,22
		MR_NS=>ITL	18,786	MR_NS=>TDD<YW	0,308	-0,22
		MR_NS=>YW	29,365	MR_NS=>ITL<YW	1,000	-0,44

Tabla D.2.66 Resultados ETAPA

D.2.16 OY2

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike's information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.23, se puede observar que el modelo con el menor AIC es un modelo sin medidas repetidas y cuya estructura de covarianza es "Componentes de la varianza", se puede

comprobar que los residuos cumplen normalidad y no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	846,605	0,200	Ninguna
Identity	855,113	0,040	Ninguna
Compound symmetry	846,605	0,026	Ninguna
Diagonal	856,179	0,040	Ninguna
Unstructured	847,436	0,024	Ninguna
Ar(1)	846,605	0,026	Ninguna

Tabla D.2.67 Modelo Mixto OY2

La Tabla D.2.24, muestra los resultados del modelo mixto, en ella se puede observar que el “Enfoque de desarrollo” y la “Granularidad” son significativos. Concretamente, TLD se comporta significativamente mejor que TLD y además existe significancia práctica. Por otro lado, grano fino (S) se comporta significativamente mejor que grano grueso (NS)

Variable	p-valor	niveles	Media	Dirección efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,000	TDD TLD	35,903 61,517	TDD<TLD	0,000	-0,43
Granularidad	0,010	NS S	45,434 54,986	NS<S	0,010	-0,20
Enfoque de desarrollo * Granularidad	0,886	TDD=>NS	29,086	TDD=>NS<S	0,129	-0,28
		TDD=>S	42,716			
		TLD=>NS	55,778	TLD=>NS<S	0,192	-0,26
		TLD=>S	67,257			
Granularidad * Enfoque de desarrollo	0,886	NS=>TDD	29,086	NS=>TDD<TLD	0,004	-0,45
		NS=>TLD	55,778			
		S=>TDD	42,716	S=>TDD<TLD	0,006	-0,30
		S=>TLD	67,257			

Tabla D.2.68 Resultados OY2

Aunque no existe significancia estadística en la interacción, grano fino (S) consigue una mejora mínima con respecto a grano grueso (NS), independientemente del enfoque de desarrollo. Por otro lado, TLD consigue una mejora mínima con respecto a TLD, independientemente de la granularidad.

D.2.17 Novisad

De acuerdo con el diseño de este experimento, el método de análisis utilizado es un Modelo Mixto. El ajuste de este modelo se evalúa mediante AIC (Akaike’s information criterion), cuanto más pequeño es este índice, más fiable es el modelo. Por otro lado, el modelo es válido cuando los residuos siguen una distribución normal.

En la Tabla D.2.25, se puede observar que el modelo con el menor AIC es un modelo sin medidas repetidas y cuya estructura de covarianza es “Componentes de la varianza”, se puede comprobar que los residuos cumplen normalidad y no ha ocurrido ninguna anomalía durante la ejecución del modelo.

Método	AIC	Normalidad	Advertencia
Variance Components	-83,546	0,060	Ninguna
Identity	-84,324	0,002	Ninguna
Compound symmetry	-83,546	0,002	Ninguna
Diagonal	-85,268	0,002	Ninguna
Unstructured	-84,584	0,002	Ninguna
Ar(1)	-83,546	0,002	Ninguna

Tabla D.2.69 Modelo Mixto Novisad

La Tabla D.2.26, muestra los resultados del modelo mixto, en ella se puede observar que el “Enfoque de desarrollo”, la “Tarea” y la interacción de los dos son significativos. Al ser la interacción significativa, no hace falta evaluar los resultados de los factores.

Variable	p-valor	niveles	Media	Media Transf.	Dirección del efecto	p-valor	Tamaño efecto
Enfoque de desarrollo	0,084	TDD TLD	0,229 0,279	24,760 20,597	TDD<TLD	0,084	-0,08
Tarea	0,000	BSK_NS MR_NS	0,135 0,373	34,641 13,956	BSK_NS<MR_S	0,000	-0,54
Enfoque de desarrollo * Tarea	0,063	TDD=>BSK_NS	0,079	43,032	TDD=>BSK_NS<MR_NS	0,000	-0,46
		TDD=>MR_NS	0,379	13,572			
		TLD=>BSK_NS	0,191	28,350	TLD=>BSK_NS<MR_NS	0,000	-0,59
		TLD=>MR_NS	0,367	14,344			
Tarea * Enfoque de desarrollo	0,063	MR_NS=>TLD	0,367	14,344	MR_NS=>TLD<TDD	0,789	-0,07
		MR_NS=>TDD	0,379	13,572			
		BSK_NS=>TDD	0,079	43,032	BSK_NS=>TDD<TLD	0,009	-0,47
		BSK_NS=>TLD	0,191	28,350			

Tabla D.2.70 Resultados Novisad

Se puede observar, que existe significancia estadística en la interacción entre el enfoque de desarrollo y la tarea, BSK_S se comporta mejor que MR_NS, independientemente del enfoque de desarrollo utilizado.

Por otro lado, para la interacción entre la tarea y el enfoque de desarrollo, TDD es significativamente mejor que TLD, cuando se aplican sobre BSK_S.

Apéndice E Resultados de la Agregación

E.1 Experimentos de Técnicas de Prueba

E.1.1 Modelo Base (Añadir Jerarquía)

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Técnica	Identidad	1989,231	Si	
	Diagonal	1973,825	Si	
	AR(1)	1990,927	Si	
	CS	1991,232	Si	No converge
	No estructurada	1978,875	Si	No converge
Período	Identidad	1989,232	Si	
	Diagonal	1984,809	Si	
	AR(1)	1989,832	Si	
	CS	1991,232	Si	No converge
	No estructurada	1989,160	Si	No converge
Programa	Identidad	1989,232	Si	
	Diagonal	1990,310	Si	
	AR(1)	1991,591	Si	
	CS	1991,232	Si	No converge
	No estructurada	1990,508	Si	No converge

Tabla E.1.71 Resultados Modelo Base

E.1.2 Añadir Variables Identificadas a sus Respectivos Niveles

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Técnica	Identidad	1409,195	Si	La matriz no está definida positivamente
	Diagonal	1392,027	Si	La matriz no está definida positivamente
	AR(1)	1410,220	Si	La matriz no está definida positivamente
	CS	1411,195	Si	No converge
	No estructurada	1395,016	Si	No converge
Período	Identidad	1409,195	Si	La matriz no está definida positivamente
	Diagonal	1407,621	Si	La matriz no está definida positivamente
	AR(1)	1409,523	Si	La matriz no está definida positivamente
	CS	1411,195	Si	No converge
	No estructurada	1412,021	Si	No converge
Programa	Identidad	1409,195	Si	La matriz no está definida positivamente
	Diagonal	1411,950	Si	La matriz no está definida positivamente
	AR(1)	1410,905	Si	La matriz no está definida positivamente
	CS	1411,195	Si	No converge
	No estructurada	1414,801	Si	No converge

Tabla E.1.72 Resultados Modelo Añadiendo las Variables Identificadas

E.1.3 Añadiendo Variables Identificadas sin DOTSEM

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Técnica	Identidad	1415,526	Si	La matriz no está definida positivamente
	Diagonal	1397,746	Si	
	AR(1)	1416,620	Si	
	CS	1417,526	Si	No converge
	No estructurada	1401,596	Si	No converge
Período	Identidad	1415,526	Si	No converge
	Diagonal	1413,516	Si	
	AR(1)	1416,056	Si	
	CS	1417,526	Si	No converge
	No estructurada	1418,122	Si	No converge
Programa	Identidad	1415,526	Si	No converge
	Diagonal	1418,102	Si	
	AR(1)	1417,273	Si	
	CS	1417,526	Si	No converge
	No estructurada	1421,491	Si	No converge

Tabla E.1.73 Resultados Modelo Añadiendo las Variables Identificadas sin DOTSEM

E.1.4 Técnica como Efecto Aleatorio

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Técnica	Identidad	1416,096	Si	La matriz no está definida positivamente
	Diagonal	1397,746	Si	
	AR(1)	1417,313	Si	
	CS	1418,096	Si	No converge
	No estructurada	1401,609	Si	No converge
Período	Identidad	1416,096	Si	No converge
	Diagonal	1414,044	Si	
	AR(1)	1416,681	Si	
	CS	1418,096	Si	No converge
	No estructurada	1418,581	Si	No converge
Programa	Identidad	1416,096	Si	La matriz no está definida positivamente
	Diagonal	1418,466	Si	
	AR(1)	1417,869	Si	
	CS	1418,096	Si	No converge
	No estructurada	1422,136	Si	No converge

Tabla E.1.74 Resultados Modelo Añadiendo la Técnica como Efecto Aleatorio

E.1.5 Interacciones

La Tabla C.5 muestra los resultados de las interacciones teniendo como el efecto repetido el Período y como matriz de covarianza Diagonal, que es con el que menor AIC se obtuvo.

Interacciones	Variables	AIC
1	Técnica*Experiencia Relativa	1412,450
2	Técnica*Experiencia Relativa, Técnica*Experiencia Absoluta	1414,316
2	Técnica*Experiencia Relativa, Técnica*Conocimientos de Programación	1413,619

Tabla E.1.75 Resultados Modelo Añadiendo Interacciones

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Técnica	Identidad	1415,085	Si	La matriz no está definida positivamente
	Diagonal	1397,586	Si	
	AR(1)	1416,176	Si	No converge
	CS	1417,085	Si	
	No estructurada	1401,731	Si	No converge
Período	Identidad	1415,085	Si	La matriz no está definida positivamente
	Diagonal	1413,619	Si	
	AR(1)	1415,764	Si	No converge
	CS	1417,085	Si	
	No estructurada	1418,212	Si	No converge
Programa	Identidad	1415,085	Si	La matriz no está definida positivamente
	Diagonal	1416,427	Si	
	AR(1)	1416,731	Si	No converge
	CS	1417,085	Si	
	No estructurada	1419,125	Si	No converge

Tabla E.1.76 Resultados Modelo Final

E.2 Experimentos de Test Driven Development (TDD)

E.2.1 Modelo Base (Añadir Jerarquía)

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Enfoque de Desarrollo	Identidad	5965,579	Si	
	CS	5967,579	Si	No converge
	Diagonal	5961,157	Si	
	No estructurada	5963,157	Si	No converge
	AR(1)	5967,579	Si	La matriz no está definida positivamente
Programa	Identidad	5965,579	Si	
	CS	5967,579	Si	No converge
	Diagonal	5966,133	Si	
	No estructurada	5968,133	Si	No converge
	AR(1)	5967,579	Si	La matriz no está definida positivamente
Período	Identidad	5965,579	Si	
	CS	5967,579	Si	No converge
	Diagonal	5964,678	Si	
	No estructurada	5966,678	Si	No converge
	AR(1)	5967,579	Si	La matriz no está definida positivamente

Tabla E.2.77 Resultados Modelo Base

E.2.2 Añadir Variables Identificadas a sus Respectivos Niveles

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Enfoque de Desarrollo	Identidad	5621,501	Si	
	CS	5623,501	Si	No converge
	Diagonal	5620,467	Si	
	No estructurada	5622,467	Si	No converge
	AR(1)	5623,501	Si	La matriz no está definida positivamente
Programa	Identidad	5621,501	Si	
	CS	5623,501	Si	No converge
	Diagonal	5622,135	Si	
	No estructurada	5624,135	Si	No converge
	AR(1)	5623,501	Si	La matriz no está definida positivamente
Período	Identidad	5621,501	Si	
	CS	5623,501	Si	No converge
	Diagonal	5622,997	Si	
	No estructurada	5624,997	Si	No converge
	AR(1)	5623,501	Si	La matriz no está definida positivamente

Tabla E.2.78 Resultados Modelo Añadiendo las Variables Identificadas

E.2.3 Enfoque de Desarrollo como Efecto Aleatorio

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Enfoque de Desarrollo	Identidad	5606,671	Si	
	CS	5608,671	Si	No converge
	Diagonal	5606,013	Si	
	No estructurada	5608,014	Si	No converge
	AR(1)	5608,671	Si	La matriz no está definida positivamente
Programa	Identidad	5606,671	Si	
	CS	5608,671	Si	No converge
	Diagonal	5607,335	Si	
	No estructurada	5609,335	Si	No converge
	AR(1)	5608,671	Si	La matriz no está definida positivamente
Período	Identidad	5606,671	Si	
	CS	5608,671	Si	No converge
	Diagonal	5608,076	Si	
	No estructurada	5610,076	Si	No converge
	AR(1)	5608,671	Si	La matriz no está definida positivamente

Tabla E.2.79 Resultados Modelo Añadiendo Enfoque de Desarrollo como Efecto Aleatorio

E.2.4 Interacciones

La Tabla E.2.4 y la Tabla E.2.5 muestra los resultados de las interacciones del factor principal *Enfoque de Desarrollo*, teniendo como medida repetida el Enfoque de Desarrollo y como matriz de covarianza Diagonal, ya que es con el que menor AIC se obtuvo.

Interacciones	Variables	AIC
1	Factor*Experiencia en Programación	5605,060
2	Factor*Exp. en Prog., Factor*Exp. con el Lenguaje de Prog.	5606,936
2	Factor*Exp. en Prog., Factor*Exp. en Pruebas Unitarias	5606,597
2	Factor*Exp. en Prog., Factor*Exp. con el framework	5606,643
2	Factor*Exp. en Prog., Factor*Tipo de Sujeto	5603,930
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Exp. con el Lenguaje de Prog.	5605,672
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Exp. en Pruebas Unitarias	5605,689
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Exp. con el framework	5605,533
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Lenguaje	5605,918
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Aspectos de transmisión del tratamiento	5605,734
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Entrenador	5605,912
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Orden entre entrenamiento y sesiones	5605,893
3	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión	5604,057
4	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Exp. con el Lenguaje de Prog.	5605,922
4	Factor* Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Exp. en Pruebas Unitarias	5605,590

Tabla E.2.80 Resultados Modelo Añadiendo Interacciones (1/2)

Interacciones	Variables	AIC
4	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Exp. con el framework	5605,221
4	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Lenguaje	5606,054
4	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Aspectos de transmisión del tratamiento	5606,055
4	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador	5604,166
4	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Orden entre entrenamiento y sesiones	5605,977
5	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. con el Lenguaje de Prog.	5606,060
5	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. en Pruebas Unitarias	5605,767
5	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. con el framework	5605,310
5	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Lenguaje	5606,143
5	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Aspectos de transmisión del tratamiento	5606,051
5	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Orden entre entrenamiento y sesiones	5606,083
6	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. framework, Factor*Exp. Lenguaje de Prog.	5607,306
6	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. framework, Factor*Exp. Pruebas Unitarias	5607,294
6	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. con el framework, Factor*Lenguaje	5607,284
6	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. con el framework, Factor*Aspectos de transmisión del tratamiento	5607,121
6	Factor*Exp. en Prog., Factor*Tipo de Sujeto, Factor*Duración de la Sesión, Factor*Entrenador, Factor*Exp. con el framework, Factor*Orden entre entrenamiento y sesiones	5607,276

Tabla E.2.81 Resultados Modelo Añadiendo Interacciones (2/2)

Efecto Repetido	Matriz de covarianza	AIC	Normalidad	Comentarios
Enfoque de Desarrollo	Identidad	5606,332	Si	
	CS	5608,332	Si	No converge
	Diagonal	5605,310	Si	
	No estructurada	5607,362	Si	No converge
	AR(1)	5608,332	Si	La matriz no está definida positivamente
Programa	Identidad	5606,332	Si	
	CS	5608,332	Si	No converge
	Diagonal	5606,833	Si	
	No estructurada	5608,833	Si	No converge
	AR(1)	5608,332	Si	La matriz no está definida positivamente
Período	Identidad	5606,332	Si	
	CS	5608,332	Si	No converge
	Diagonal	5607,316	Si	
	No estructurada	5609,316	Si	No converge
	AR(1)	5608,332	Si	La matriz no está definida positivamente

Tabla E.2.82 Resultados Modelo Final

BIBLIOGRAFÍA

- Basili, V., 1996. The Role of Experimentation in Software Engineering: Past, Current, and Future. *Proceedings Int'l Conf on Software Engineering*, pp. 442- 449.
- Basili, V. R., Selby, R. & Hutchens, D., 1986. Experimentation in Software Engineering. *IEEE Transactions on Software Engineering*, 12(7), pp. 733-743.
- Basili, V., Shull, F. & Lanubile, F., 1999. Building Knowledge through Families of Experiments. *IEEE Transactions on Software Engineering*, 25(4), pp. 456-473.
- Beck, K., 2003. *Test Driven Development: By Example*. s.l.:Addison Wesley.
- Beizer, B., 1990. *Software Testing Techniques*. 2th edition ed. s.l.:International Thomson Computer Press.
- Berander, P., 2004. Using students as subjects in requirements prioritization.. *Proceedings of the International Symposium on Empirical Software Engineering (ISESE '04)*, p. 167–176.
- Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R., 2011. *Introduction to Meta-Analysis*. s.l.:John Wiley & Sons.
- Carver, J., Jaccheri, L., Morasca, S. & Shull, F., 2003. Issues in Using Students in Empirical Studies in Software Engineering Education. *Proceedings of the International Software Metrics Symposium (METRICS'03)*, pp. 239-249.
- Carver, J., 2010. Towards Reporting Guidelines for Experimental Replications: A Proposal. *Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER) [Held during ICSE 2010]*.
- Ciolkowski, M., 2009. What Do We Know About Perspective-Based Reading? An Approach for Quantitative Aggregation in Software Engineering. *Third International Symposium on Empirical Software Engineering and Measurement*, pp. 133-144.
- Ciolkowski, M., 2012. An approach for quantitative aggregation of evidence from controlled experiments in software engineering. En: University of Kaiserslautern, Fraunhofer-Verlag: s.n., pp. 1-233.
- Ciolkowski, M., Shull, F. & Biffl, S., 2002. A Family of Experiments to Investigate the Influence of Context on the Effect of Inspection Techniques. *6th International Conference on Empirical Assessment in Software Engineering*, pp. 48-60.
- Dieste, O., Grimán, A. & Juristo, N., 2009. Developing search strategies for detecting relevant experiments. *Empirical Software Engineering*, 14(5), p. 513–539.
- Dyba, T., Kitchenham, B. & Jorgensen, M., 2005. Evidence-based software engineering for practitioners. *IEEE Software*, 22(1), pp. 58-65.
- Feldt, R. y otros, 2018. Four commentaries on the use of students and professionals in empirical software engineering experiments. *Empirical Software Engineering*, 23(6), pp. 3801-3820.
- Fenton, N., 1993. How Effective Are Software Engineering Methods?. *Journal of Systems and Software*, 22(2), pp. 141-146.
- Field, A., 2013. *Discovering Statistics Using IBM SPSS Statistics*. 4th edition ed. s.l.:Sage Publications Inc.
- Fucci, D. y otros, 2016. An External Replication on the Effects of Test-drivenDevelopment Using a Multi-site Blind Analysis Approach. *IEEE International Symposium on Empirical Software Engineering and Measurement*.
- Fucci, D. & Turhan, B., 2013. A Replicated Experiment on the Effectiveness of Test-First Development. *IEEE International Symposium on Empirical Software Engineering and Measurement*, pp. 103-112.
- Goldstein, H., 1995. *Multilevel statistical models*. 4th Edition ed. New York, NY.: Halsted.
- Gonzalez-Huerta, J., Insfran, E., Abrahão, S. & Scanniello, G., 2015. Validating a model-driven software architecture evaluation and improvement method: A family of experiments. *Information and Software Technology*, Volumen 57, p. 405–429.
- Gómez, O., Juristo, N. & Vegas, S., 2014. Understanding replication of experiments in software engineering: A classification. *Information and Software Technology*, Volumen 56, pp. 1033-1048.
- Gómez, O. S. G., 2012. *Tipología de Replicaciones para la Síntesis de Experimentos en Ingeniería del Software*. s.l.:s.n.
- Höfer, A. & Tichy, W. F., 2007. Status of Empirical Research in Software Engineering. *Empirical Software Engineering Issues. Critical Assessment and Future Directions*, Volumen 4336, pp. 10-19.
- Höst, M., Regnell, B. & Wohlin, C., 2000. Using Students as Subjects: A Comparative Study of Students and Professionals in Lead-Time Impact Assessment. *Empirical Software Engineering*, 5(3), p. 201–214.

- Heck, R., Scott, T. & Lynn, T., 2014. *Multilevel and Longitudinal Modeling with IBM SPSS*. s.l.:Taylor & Francis.
- IBM SPSS Statistics for Macintosh, Version 25 (2017).
- Jedlitschka, A., Ciolkowski, M. & Pfahl, D., 2008. Reporting experiments in software engineering. *Guide to Advanced Empirical Software Engineering*, p. 201–228.
- Juristo, N., 2016. Once is not Enough: Why we Need Replication. En: T. Menzies, L. Williams & T. Zimmermann, eds. *Perspectives on Data Science for Software Engineering*. s.l.:Elsevier Inc, pp. 299-302.
- Juristo, N. & Moreno, A., 2001. *Basics of Software Engineering Experimentation*. s.l.:Kluwer Academic Publishers.
- Juristo, N. y otros, 2012. Comparing the Effectiveness of Equivalence Partitioning, Branch Testing and Code Reading by Stepwise Abstraction Applied by Subjects. *IEEE Fifth International Conference on Software Testing, Verification and Validation*.
- Kitchenham, B., 2004. *Procedures for Performing Systematic Reviews*. s.l.:Keele University.
- Kitchenham, B., 2008. The Role of Replications in Empirical Software Engineering—A Word of Warning. *Empirical Software Engineering*, 13(2), pp. 219-221.
- Kitchenham, B. A. y otros, 2002. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8), pp. 721-734.
- Krein, J. L. y otros, 2016. A Multi-Site Joint Replication of a Design Patterns Experiment Using Moderator Variables to Generalize across Contexts. *IEEE Transactions on Software Engineering*, 43(4), pp. 302-321.
- Linger, R. C., Mills, H. D. & Witt, B. I., 1979. *Structured Programming: Theory and Practice*. s.l.:Addison-Wesley.
- McMeekin, D., Von Konsky, B., Robey, M. & Cooper, D. J. A., 2009. The Significance of Participant Experience when Evaluating Software Inspection Techniques. *Proceedings of the Australian Software Engineering Conference (ASWEC'09)*, p. 200–209.
- Myers, G., Badgett, T., Thomas, T. & Sandler, C., 2004. *The Art of Software Testing*. 2th edition ed. s.l.:Wiley-Interscience.
- Petersen, K., Feldt, R., Mujtaba, S. & Mattsson, M., 2008. Systematic mapping studies in software engineering.. *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE'08)*, p. 71–80.
- Pfleeger, S. L., 1994. *Software engineering : theory and practice / Shari Lawrence Pfleeger*. s.l.:Prentice Hall.
- Porter, A. & Votta, L., 1998. Comparing Detection Methods for Software Requirements Inspections: A Replication Using Professional Subjects. *Empirical Software Engineering*, 3(4), p. 355–379.
- Remus, W., 1989. Using Students as Subjects in Experiments on Decision Support Systems.. *Proceedings of the Hawaii International Conference on System Sciences (HICSS'89)*, pp. 176-180.
- Runeson, P., 2003. Using Students as Experiment Subjects— An Analysis on Graduate and Freshmen PSP Student Data. *Proceedings of the International Conference on Empirical Assessment in Software Engineering (EASE'03)*, p. 95–102.
- Santos, A., Gomez, O. & Juristo, N., 2018. Analyzing Families of Experiments in SE: a Systematic Mapping Study. *IEEE Transactions on Software Engineering*.
- Santos, A., Vegas, S., Oivo, M. & Juristo, N., 2019. A Procedure and Guidelines for Analyzing Groups of Software Engineering Replications. *IEEE Transactions on Software Engineering*.
- Shadish, W., Cook, T. & Campbell, D., 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. s.l.:Cengage Learning.
- Shepperd, M., Ajenkab, N. & Couns, S., 2018. The role and value of replication in empirical software engineering results. *Information and Software Technology*, Volumen 99, pp. 120-132.
- Shull, F., Carver, J., Vegas, S. & Juristo, N., 2008. The Role of Replications. *Empirical Software Engineering*, 13(2), pp. 211-218.
- Sjøberg, D., Hannay, J. E., Hansen, O. & Kam, V. B., 2005. A survey of controlled experiments in software engineering.. *IEEE Transactions on Software Engineering*., 31(9), p. 733–753.
- Svahnberg, M., Aurum, A. & Wohlin, C., 2008. Using Students as Subjects - An Empirical Evaluation.. *Proceedings of the Second ACMIEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'08)*, p. 288–290.
- Tosun, A. y otros, 2017. An Industry Experiment on the Effects of Test-Driven Development on External Quality and Productivity. *Empirical Software Engineering*, 22(6), pp. 2763-2805.
- Uyaguari, F., 2017. s.l.:s.n.

Vegas, S., Apa, C. & Juristo, N., 2016. Crossover Designs in Software Engineering Experiments: Benefits and Perils. *IEEE Transactions on Software Engineering*, Volumen 42, pp. 120-135.

Wohlin, C., Runeson, P. & Höst, M., 1999. *Experimentation in Software Engineering: An Introduction*. s.l.:Springer.

Lista de Estudios Primarios (Familias de Experimentos)

- [EP1] Abrahão, S., Gravino, C., Insfran, E., Scanniello, G., Tortora, G. Assessing the Effectiveness of Sequence Diagrams in the Comprehension of Functional Requirements: Results from a Family of Five Experiments. *IEEE Transactions on Software Engineering*, 39(3): 327-342, 2013.
- [EP2] Acuña, S. T., Gómez, M. N., Hannay, J. E., Juristo, N., Pfahl, D. Are team personality and climate related to satisfaction and software quality? aggregating results from a twice replicated experiment. *Information and Software Technology*, 57, 141-156, 2015.
- [EP3] Ali, S., Yue, T., Rubab, I. Assessing the modeling of aspect state machines for testing from the perspective of modelers. *14th International Conference on Quality Software*, pp. 234-239, 2014.
- [EP4] Canfora, G., García, F., Piattini, M., Ruiz, F., Visaggio, C. A. A family of experiments to validate metrics for software process models. *Journal of Systems and Software*, 77(2), 113-129, 2005.
- [EP5] Ceccato, M., Penta, M., Falcarin, P., Ricca, F., Torchiano, M., Tonella, P. A family of experiments to assess the effectiveness and efficiency of source code obfuscation techniques. *Empirical Software Engineering*, 19, 1040-1074, 2014.
- [EP6] Ceccato, M., Marchetto, A., Mariani, L., Nguyen, C. D., Tonella, P. Do automatically generated test cases make debugging easier? an experimental assessment of debugging effectiveness and efficiency. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 25(1), 1-38, 2015.
- [EP7] Cruz-Lemus, J. A., Genero, M., Caivano, D., Abrahão, S., Insfran, E., Carsí, J. A. Assessing the influence of stereotypes on the comprehension of UML sequence diagrams: A family of experiments. *Information and Software Technology*, 53(12), 1391-1403, 2011.
- [EP8] Cruz-Lemus, J. A., Genero, M., Manso, M. E., Morasca, S., Piattini, M. Assessing the understandability of UML statechart diagrams with composite states--A family of empirical studies. *Empirical Software Engineering*, 14(6), 685-719, 2009.
- [EP9] Fernandez, A., Abrahão, S., Insfran, E. Empirical validation of a usability inspection method for model-driven web development. *Journal of Systems and Software*, 86(1), 161-186, 2013.
- [EP10] Fernández-Sáez, A. M., Genero, M., Caivano, D., Chaudron, M. R. V. Does the level of detail of UML diagrams affect the maintainability of source code?: A family of experiments. *Empirical Software Engineering*, 21, 212-259, 2016. 2013.
- [EP11] Fernández-Sáez, A. M., Genero, M., Chaudron, M. R. V., Caivano, D., Ramos, I. Are forward designed or reverse-engineered UML diagrams more helpful for code maintenance?: A family of experiments. *Information and Software Technology*, 57, 644-663, 2015.
- [EP12] Figueiredo, E., Garcia, A., Maia, M., Ferreira, G., Nunes, C., Whittle, J. On the impact of crosscutting concern projection on code measurement. *AOSD Aspect-Oriented Software Development*. 81-92, 2011.
- [EP13] George, B., & Williams, L. A structured experiment of test-driven development. *Information and Software Technology*, 46(5), 337-342, 2004.
- [EP14] Gonzalez-Huerta, J., Insfran, E., Abrahão, S., & Scanniello, G. Validating a model-driven software architecture evaluation and improvement method: A family of experiments. *Information and Software Technology*, 57, 405-429, 2015.
- [EP15] Hadar, I., Reinhartz-Berger, I., Kuflik, T., Perini, A., Ricca, F., Susi, A. Comparing the comprehensibility of requirements models expressed in use case and tropos: Results from a family of experiments. *Information and Software Technology*, 55(10), 1823-1843, 2013.
- [EP16] Jørgensen, M. Contrasting ideal and realistic conditions as a means to improve judgment-based software development effort estimation. *Information and Software Technology*, 53(12), 1382-1390, 2011.
- [EP17] Juristo, N., Vegas, S., Solari, M., Abrahao, S., Ramos, I. Comparing the effectiveness of equivalence partitioning, branch testing and code reading by stepwise abstraction applied by subjects. *IEEE Fifth International Conference on Software Testing, Verification and Validation*, 330-339, 2012.

- [EP18] Kosar, T., Mernik, M., Carver, J. C. Program comprehension of domain-specific and general-purpose languages: Comparison using a family of experiments. *Empirical Software Engineering*, 17, 276-304, 2012.
- [EP19] Krein, J. L., Prechelt, L., Juristo, N., Nanthaamornphong, A., Carver, J. C., Vegas, S., Knutson, C. D., Seppi, K. D., Eggett, D. L. A multi-site joint replication of a design patterns experiment using moderator variables to generalize across contexts. *IEEE Transactions on Software Engineering*, 42(4), 302-321, 2016.
- [EP20] Lima, C. R., Do Carmo Machado, I., Cardoso, V., Santana De Almeida, E. Analyzing the effectiveness of a system testing tool for software product line engineering. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, 584-588, 2013.
- [EP21] Laitenberger, O., El Emam, K., Harbich, T. G. An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code documents. *IEEE Transactions on Software Engineering*, 27(5), 387-421, 2001.
- [EP22] Manso, M. E., Cruz-Lemus, J. A., Genero, M., Piattini, M. Empirical validation of measures for UML class diagrams: A meta-analysis study. *International Conference on Model Driven Engineering Languages and Systems (MODELS)*, LNCS, 5421, 303-313, 2009.
- [EP23] Maćkowiak, M., Nawrocki, J., Ochodek, M. On some end-user programming constructs and their understandability. *Journal of Systems and Software*, 142, 206-222, 2018.
- [EP24] Morales, J. M., Navarro, E.; Sánchez P., Alonso, D. A family of experiments to evaluate the understandability of TriStar and i* for modeling teleo-reactive systems. *The Journal of Systems and Software* 114: 82-100, 2016.
- [EP25] Mouchawrab, S., Briand, L. C., Labiche, Y., Di Penta, M. Assessing, comparing, and combining state machine-based testing and structural testing: A series of experiments. *IEEE Transactions on Software Engineering*, 37(2), 161-187, 2011.
- [EP26] Muñoz, L., Mazón, J., Trujillo, J. A family of experiments to validate measures for UML activity diagrams of ETL processes in data warehouses. *Information and Software Technology*, 52(11), 1188-1203, 2010.
- [EP27] Panach, J. I., Dieste, O., Marín, B., España, S., Vegas, S., Pastor, O., Juristo, N. Evaluating Model-Driven Development Claims with respect to Quality: A Family of Experiments. *IEEE Transactions on Software Engineering*, 2018.
- [EP28] Pfahl, D., Laitenberger, O., Ruhe, G., Dorsch, J., Krivobokova, T. Evaluating the learning effectiveness of using simulations in software project management education: Results from a twice replicated experiment. *Information and Software Technology*, 46(2), 127-147, 2004.
- [EP29] Porter, A., Votta, L. Comparing detection methods for software requirements inspections: A replication using professional subjects. *Empirical Software Engineering*, 3(4), 355-379, 1998.
- [EP30] Reynoso, L., Genero, M., Piattini, M., Manso, E. Assessing the impact of coupling on the understandability and modifiability of OCL expressions within UML/OCL combined models. *11th IEEE International Software Metrics Symposium (METRICS'05)*, 10-14, 2005.
- [EP31] Riaz, M., King, J., Slankas, J., Williams, L., Massacci, F., Quesada-López, C., Jenkins, M. Identifying the implied: Findings from three differentiated replications on the use of security requirements templates. *Empirical Software Engineering*, 22, 2127-2178, 2017.
- [EP32] Ricca, F., Di Penta, M., Torchiano, M., Tonella, P., Ceccato, M. How developers' experience and ability influence web application comprehension tasks supported by UML stereotypes: A series of four experiments. *IEEE Transactions on Software Engineering*, 36(1), 96-118, 2010.
- [EP33] Ricca, F., Di Penta, M., Torchiano, M., Tonella, P., Ceccato, M., Visaggio, C. A. Are fit tables really talking? A series of experiments to understand whether fit tables are useful during evolution tasks. *ACM/IEEE 30th International Conference on Software Engineering*, 361-370, 2008.
- [EP34] Ricca, F., Scanniello, G., Torchiano, M., Reggio, G., Astesiano, E. Assessing the effect of screen mockups on the comprehension of functional requirements. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 24(1), 1-38, 2014.
- [EP35] Ricca, F., Torchiano, M., Leotta, M., Tiso, A., Guerrini, G., Reggio, G. On the impact of state-based model-driven development on maintainability: a family of experiments using UniMod. *Empirical Software Engineering*, 23, 1743-1790, 2018.
- [EP36] Runeson, P., Stefik, A., Andrews, A. Variation factors in the design and analysis of replicated controlled experiments Three (dis)similar studies on inspections versus unit testing. *Empirical Software Engineering*, 19, 1781-1808, 2014.

- [EP37] Salleh, N., Mendes, E., Grundy, J. Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments. *Empirical Software Engineering*, 19, 714-752, 2014.
- [EP38] Scanniello, G., Gravino, C., Genero, M., Cruz-Lemus, J. A., Tortora, G. On the impact of UML analysis models on source-code comprehensibility and modifiability. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 23(2), 1-26, 2014.
- [EP39] Scanniello, G., Gravino, C., Tortora, G., Dodero, G. Documenting design-pattern instances: A family of experiments on source-code comprehensibility. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 24(3), 1-35, 2015.
- [EP40] Scanniello, G., Gravino, C., Tortora, G., Genero, M., Risi, M., Cruz-Lemus, J. A., Dodero, G. Studying the effect of UML-based models on source-code comprehensibility: Results from a long-term investigation. *International Conference on Product-Focused Software Process Improvement (PROFES), LNCS*, 9459, 311-327, 2015.
- [EP41] Scanniello, G., Risi, M., Porfirio, T., Romano, S. Fixing Faults in C and Java Source Code: Abbreviated vs. Full-Word Identifier Names. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 26(2), 43 pages, 2017.
- [EP42] Souza, E., Moreira, A., Araújo, J., Abrahão, S., Insfran, E., Silva da Silveira, D. Comparing business value modeling methods: A family of experiments. *Information and Software Technology*, 104, 179-193, 2018.
- [EP43] Staron, M., Kuzniarz, L., Wohlin, C. Empirical assessment of using stereotypes to improve comprehension of UML models: A set of experiments. *Journal of Systems and Software*, 79(5), 727-742, 2006.
- [EP44] Teruel, M. A., Navarro, E., López-Jaquero, V., Montero, F., Jaen, J., González, P. Analyzing the understandability of requirements engineering languages for CSCW systems: A family of experiments. *Information and Software Technology*, 54(11), 1215-1228, 2012.
- [EP45] Tofan, D., Galster, M., Lytra, I., Avgeriou, P., Zdun, U., Fouche, M. A., Boer, R., Solms, F. Empirical evaluation of a process to increase consensus in group architectural decision making. *Information and Software Technology*, 72, 31-47, 2016.
- [EP46] Von Wangenheim, C. G., Thiry, M., Kochanski, D. Empirical evaluation of an educational game on software measurement. *Empirical Software Engineering*, 14, 418-452, 2009.

Lista de Estudios Primarios (Impacto de los Estudiantes)

- [S1] S. Abrahão, C. Gravino, E. Insfran, G. Scanniello, G. Tortora. Assessing the Effectiveness of Sequence Diagrams in the Comprehension of Functional Requirements: Results from a Family of Five Experiments. *IEEE Transactions on Software Engineering*. 39(3):327-342, 2013.
- [S2] W. Afzal, A.N. Ghazi, J. Itkonen, R. Torkar, A. Andrews, K. Bhatti. An Experiment on the Effectiveness and Efficiency of Exploratory Testing. *Empirical Software Engineering*, 20:844-878, 2015.
- [S3] S. Alimadadi, S. Sequeira, A. Mesbah, K. Pattabiraman. Understanding JavaScript event-based interactions. In *Proceedings of the 36th International Conference on Software Engineering (ICSE 2014)*. Pp. 367-377, 2014. Also appears as S. Alimadadi, S. Sequeira, A. Mesbah, K. Pattabiraman. Understanding JavaScript Event-Based Interactions with Clematis. In *Proceedings of the 36th International Conference on Software Engineering (ICSE 2014)*. Pp. 367-377, 2014.
- [S4] V. Arnaoudova, M. Di Penta, G. Antoniol. Linguistic Antipatterns: What they Are and How Developers Perceive them. *Empirical Software Engineering*, 21:104-158, 2016.
- [S5] G. Bavota, R. Oliveto, M. Gethers, D. Poshyanyk, A. De Lucia. Methodbook: Recommending Move Method Refactorings via Relational Topic Models. *IEEE Transactions on Software Engineering*. 40(7): 671- 694, 2014.
- [S6] G. Bavota, A. Qusef, R. Oliveto, A. De Lucia, D. Binkley. Are Test Smells Really Harmful? An Empirical Study. *Empirical Software Engineering*, 20:1052-1094, 2015.
- [S7] D. Binkley, M. Davis, D. Lawrie, J.I. Maletic, C. Morrell, B. Sharif. The impact of identifier style on effort and comprehension. *Empirical Software Engineering*. 18(2):1382-3256, 2013.
- [S8] B. Cornelissen, A. Zaidman, A. van Deursen. A Controlled Experiment for Program Comprehension through Trace Visualization. *IEEE Transactions on Software Engineering*. 37(3): 341-355, 2011.
- [S9] G. Fraser, M. Staats, P. McMinn, A. Arcuri, F. Padberg. Does Automated Unit Test Generation Really Help Software Testers? A Controlled Empirical Study. *ACM Transactions on Software Engineering and Methodology*. 24(4), Article 23, 2015.
- [S10] L. Guerrouj, M. Di Penta, Y.G. Guéhéneuc, G. Antoniol. An experimental investigation on the effects of context on source code identifiers splitting and expansion. *Empirical Software Engineering*. 19(6):1706-1753, 2014.

- [S11] E. Guzman, O.A. Valeo, B.Bruegge. Retrieving Diverse Opinions from App Reviews. In *Proceeding of the International Symposium on Empirical Software Engineering and Measurement (ESEM'15)*, 2015.
- [S12] W. Heijstek, T. Kühne, M.R.V. Chaudron. Experimental Analysis of Textual and Graphical Representations for Software Architecture Design. In *Proceeding of the International Symposium on Empirical Software Engineering and Measurement (ESEM'11)*. Pp 167-176, 2011.
- [S13] A. Heydarnoori, K. Czarnecki, W. Binder, T.T. Bartolomei. Two Studies of Framework-Usage Templates Extracted from Dynamic Traces. *IEEE Transactions on Software Engineering*. 38(6):1464-1487, 2012.
- [S14] R. Holmes, R.J. Walker. Systematizing pragmatic software reuse. *ACM Transactions on Software Engineering and Methodology*. 21(4), Article 20, 44 pages, 2013.
- [S15] J. Jung, K. Hoefig, D. Domis, A. Jedlitschka, M. Hiller. Experimental Comparison of Two Safety Analysis Methods and Its Replication. In *Proceedings of the International Symposium on Empirical Software Engineering and Measurement (ESEM'13)*. Pp. 223-232, 2013.
- [S16] J. Jurkiewicz, J. Nawrocki, M. Ochodek, T. Glowacki. HAZOP-Based Identification of Events in Use Cases. An Empirical Study. *Empirical Software Engineering*, 20:82-109, 2015.
- [S17] S. Kaleeswaran, V. Tulsian, A. Kanade, A. Orso. MintHint: automated synthesis of repair hints. In *Proceedings of the 36th International Conference on Software Engineering (ICSE 2014)*. Pp. 266-276, 2014.
- [S18] D. Kim, J. Nam, J. Song, S. Kim. Automatic patch generation learned from human-written patches. In *Proceedings of the 2013 International Conference on Software Engineering (ICSE '13)*. Pp. 802-811, 2013.
- [S19] J. Lee, K.H. Hsu. GEA: A Goal-Driven Approach to Discovering Early Aspects. *IEEE Transactions on Software Engineering*. 40(6):584-602, 2014.
- [S20] Y.Y. Lee, D. Marinov, R.E. Johnson. Tempura: Temporal Dimension for IDEs. In *Proceedings of the International Conference on Software Engineering (ICSE '15)*. Pp. 212-222, 2015.
- [S21] N. Machado, D. Quinta, B. Lucia, L. Rodrigues. Concurrency Debugging with Differential Schedule Projections. *ACM Transactions on Software Engineering and Methodology*. 25(2), Article 14, 2016.
- [S22] P. Mäder, A. Egyed. Do Developers Benefit from Requirements Traceability when Evolving and Maintaining a Software System? *Empirical Software Engineering*, 20:413-441, 2015.
- [S23] P.W. McBurney, C. McMillan. An Empirical Study of the Textual Similarity between Source Code Summaries. *Empirical Software Engineering*, 21:17-42, 2016.
- [S24] P.W. McBurney, C. McMillan. Automatic Source Code Summarization of Context for Java Methods. *IEEE Transactions on Software Engineering*. 42(2):103-119, 2016.
- [S25] C. McMillan, M. Grechanik, D. Poshyvanyk, C. Fu, Q. Xie. Exemplar: A Source Code Search Engine for Finding Highly Relevant Applications. *IEEE Transactions on Software Engineering*. 38(5):1069-1087, 2012.
- [S26] C. McMillan, M. Grechanik, D. Poshyvanyk, Q. Xie, C. Fu. Portfolio: finding relevant functions and their usage. In *Proceedings of the 33rd International Conference on Software Engineering (ICSE'11)*. Pp. 111-120, 2011.
- [S27] J. Melo, C. Brabrand, A. Wasowski. How Does the Degree of Variability Affect Bug Finding? In *Proceedings of the International Conference on Software Engineering (ICSE '16)*. Pp. 679-690, 2016.
- [S28] S. Panichella, A. Panichella, M. Beller, A. Zaidman, H.C. Gall. The Impact of Test Case Summaries on Bug Fixing Performance: An Empirical Investigation. In *Proceedings of the International Conference on Software Engineering (ICSE'16)*. Pp 547-558, 2016.
- [S29] M. Perepletchikov, C. Ryan. A Controlled Experiment for Evaluating the Impact of Coupling on the Maintainability of Service-Oriented Software. *IEEE Transactions on Software Engineering*. 37(4): 449-465, 2011.
- [S30] E.M. Rodrigues, F.M. de Oliveira, L.T Costa, M. Bernardino, A.F. Zorzo, S. R.S. Souza, R. Saad. An Empirical Comparison of Model-Based and Capture and Replay Approaches for Performance Testing. *Empirical Software Engineering*, 20:1831-1860, 2015.
- [S31] I. Salman, A. Tosun Misirli, N. Juristo. Are Students Representatives of Professionals in Software Engineering Experiments? In *Proceedings of the International Conference on Software Engineering (ICSE'15)*. Pp 666-676, 2015.
- [S32] A. Santhiar, O. Pandita, A. Kanade. Mining Unit Tests for Discovery and Migration of Math APIs. *ACM Transactions on Software Engineering and Methodology*. 24 (1), Article 4, 33 pages, 2014.
- [S33] G. Scanniello, C. Gravino, M. Genero, J.A. Cruz-Lemus, G. Tortora. On the impact of UML analysis models on source-code comprehensibility and modifiability. *ACM Transactions on Software Engineering and Methodology*. 23(2), Article 13, 26 pages, 2014.
- [S34] G. Scanniello, C. Gravino, M. Risi, G. Tortora, G. Doderio. Documenting Design-Pattern Instances: A Family of Experiments on Source-Code Comprehensibility. *ACM Transactions on Software Engineering and Methodology*. 24(3), Article 14, 2015.
- [S35] S.E. Sim, M. Umarji, S. Ratanotayanon, C.V. Lopes. How Well Do Search Engines Support Code Retrieval on the Web? *ACM Transactions on Software Engineering and Methodology*. 21(1), Article 4, 2011.
- [S36] Z. Soh, T. Drioul, P.A. Rappe, F. Khomh, Y.G. Guéhéneuc, N. Habra. Noises in Interaction Traces Data and their Impact on Previous Research Studies. In *Proceeding of the International Symposium on Empirical Software Engineering and Measurement (ESEM'15)*, 2015.

- [S37] Y. Tao, J. Kim, S. Kim, C. Xu. Automatically generated patches as debugging aids: a human study. In *Proceedings of the 22nd International Symposium on Foundations of Software Engineering (FSE'14)*. Pp. 64-74, 2014.
- [S38] Y. Tu, E. Tempero, C. Thomborson. An Experiment on the Impact of Transparency on the Effectiveness of Requirements Documents. *Empirical Software Engineering*, 21:1035-1066, 2016.
- [S39] P.M. Uesbeck, A. Stefik, S. Hanenberg, J. Pedersen, P. Daleiden. An Empirical Study on the Impact of C++ Lambdas and Programmer Experience. In *Proceedings of the International Conference on Software Engineering (ICSE'16)*. Pp 760-771, 2016.
- [S40] R. Wettel, M. Lanza, R. robbes. Software Systems as Cities: A Controlled Experiment. In *Proceedings of the International Conference on Software Engineering (ICSE '11)*. Pp. 551-560, 2011.
- [S41] M. Würsch, E. Giger, H.C. Gall. Evaluating a query framework for software evolution data. *ACM Transactions on Software Engineering and Methodology*. 22(4), Article 38, 38 pages, 2013.