OULU 2020
A 740

ACTA UNIVERSITATIS OULUENSIS

Adrian Santos Parrilla

ANALYZING FAMILIES OF EXPERIMENTS IN SOFTWARE ENGINEERING

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 740

ADRIAN SANTOS PARRILLA

ANALYZING FAMILIES OF EXPERIMENTS IN SOFTWARE ENGINEERING

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Information Technology and Electrical Engineering of the University of Oulu for public defence in the Wetteri auditorium (IT115), Linnanmaa, on 22 April 2020, at 12 noon

UNIVERSITY OF OULU, OULU 2020

Acta Univ. Oul. A 740, 2020

Supervised by
Professor Natalia Juristo

Reviewed by
Professor Jeffrey Carver
Professor Marcela Genero

Opponent
Professor Robert Feldt

ISBN 978-952-62-2544-9 (Paperback)
ISBN 978-952-62-2545-6 (PDF)

ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2020

Santos Parrilla, Adrian, Analyzing families of experiments in software engineering.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering
Acta Univ. Oul. A 740, 2020
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Context: Experiments are commonplace in software engineering (SE). Still, two main limitations impact their suitability to assess the effectiveness of SE treatments (i.e., methods, processes, and tools): (1) their results are restricted to the configuration of the experimental settings, and (2) their results may be unreliable due to the low number of subjects typically participating. With the aim of overcoming the previous weaknesses, SE researchers are collaborating towards the construction of groups of experiments by means of replication (i.e., conducting families of experiments). Disparate aggregation techniques are being applied to aggregate experiments' results within families.

Objective: Understanding the limitations of individual experiment's results. Identifying the techniques used to aggregate experiments' results in SE families. Understanding the advantages and disadvantages of each aggregation technique in the SE context. Providing guidelines for analyzing SE families.

Method: We identified the aggregation techniques used to aggregate experiments' results in SE families. Meanwhile, we learned about the advantages and disadvantages of each aggregation technique in the literature on mature experimental disciplines such as medicine and pharmacology. Then, we applied the aggregation techniques on a representative SE family. Finally, we tailored a set of guidelines to analyze SE families based on the guidelines from medicine, but here adapted to the SE context.

Results: Families of experiments grant access to the raw data, and to the characteristics of the experiments and the participants. Families are usually comprised of a low number of experiments with small and dissimilar sample sizes and heterogeneous results. Narrative synthesis, aggregated data (AD), individual participant data (IPD), either mega-trial or stratified, and aggregation of p-values were used to analyze SE families. AD and IPD stratified, when used in tandem, seem suitable to analyze SE families.

Conclusion: The aggregation techniques used to analyze SE families should be justified in research articles to increase the reliability and transparency of the findings. Guidelines may ease such endeavour.

Keywords: aggregated data, families of experiments, individual participant data, meta-analysis, narrative synthesis

Santos Parrilla, Adrian, Ohjelmistotekniikan kokeiluperheiden analysointi.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta
Acta Univ. Oul. A 740, 2020
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Konteksti: Kokeet ovat arkipäiväisiä ohjelmistotuotannossa (SE). Kuitenkin kaksi päärajoitusta vaikuttaa niiden sopivuuteen arvioidakseen SE:n menetelmien, prosessien ja työkalujen tehokkuutta: (1) niiden tulokset rajoittuvat kokeellisten asetelmien kokoonpanoon; (2) niiden tulokset saattavat olla epäluotettavia pienestä osallistujamäärästä johtuen. SE tutkijat tekevät yhteistyötä voittaakseen edellä mainitut rajoitteet rakentamalla kokeiden ryhmiä replikoinnin kautta (eli, suorittavat kokeiden perheitä). Erilaisia koostamistekniikoita sovelletaan perheensisäisten kokeiden tulosten koostamiseen.

Tavoite:Ymmärtää yksittäisten kokeiden tuloksien rajoitukset. Tunnistaa tekniikat, joita käytetään perheen kokeiden tuloksien koostamiseen. Ymmärtääjokaisen koostamistekniikan edut ja haitatSE kontekstissa. Tarjota ohjenuoriaSE-perheiden analysointiin.

Menetelmä:Tunnistimme koostamistekniikat, joita on käytetty SE kokeiden tulosten koostamiseen. Tieteellisen kirjallisuuden avulla, koskien kokeiden tulosten koostamista mm. lääketieteen ja farmakologian aloilta, selvitimme koostamistekniikoiden hyödyt ja haitat.Seuraavaksi sovelsimme koostamistekniikoita edustavaan SE-perheeseen. Lopuksi räätälöitiin ohjenuora SE-perheiden analysointiin, joka perustuu lääketieteeseen ja on muokattu ohjelmistotuotannon kontekstiin sopivaksi.

Tulokset: Kokeiden perheet antavat pääsyn raakadataan sekä kokeiden ja osanottajien ominaispiirteisiin. Perheet käsittävät tavallisesti pienen määränkokeita,joiden näytekoot ovat pieniä ja tulokset heterogeenisiä. Kertomussynteesi, yhdistetty data (AD), yksilöllisen osallistujadatan(IDP)mega-koetta tai kerrostumaa ja p-arvon koostamista on käytetty analysoimaan SE-perheitä. AD ja IDP kerrostumaa yhdessä käytettynä näyttää sopivalta SE-perheiden analysointiin.

Johtopäätös: Koostamistekniikat, joita käytetään analysoimaan SE-perheitä, pitäisi olla perusteltuja tieteellisissä julkaisuissa, jotta havaintojen luotettavuutta ja läpinäkyvyyttä voidaan lisätä. Ohjenuorat saattavat helpottaa tällaisiapyrkimyksiä.

Asiasanat: aggregoitu tieto, koeperheet, meta-analyysi, narrative-synteesi, yksittäisten osallistujien tiedot

A la memoria de mi abuela

Preface

It would be an honor if this thesis was read by more than only those "privileged", or truly forced to read it[1]: me, my supervisor, the reviewers and my opponent. In the following I am trying to condense five years of exhausting mental effort and psychological turnabouts into a bunch of pages that are supposed to be worth a PhD. in software engineering—albeit my opponent kindly opposes to this.

All these pages would never have been possible without the constant support of my thesis supervisor, Professor Natalia Juristo[2]. I would really like to thank her kindness, understanding, and psychological support throughout the years.

I would also like to thank to all those who accompanied me along my PhD studies: Alireza, Davide, Iflaah, Ilaria, Itir, Lucy, Nebojsa, Nirnaya, Ovais, Pertti, Prabhat, Rahul, Sandun, and Woub.

I would not want to miss the opportunity of thanking the reviewers: Marcela and Jeffrey. Your comments aided in polishing the chaotic thoughts that I tried to put to paper.

Para finalizar, qusiera darle las gracias a mi familia y amigos: Mama, Papa, Noelia, María, Ricardo, Eva, Daniel, Beatriz, Sergio, Leticia, Alejandro, Rebeca, Javier, Pablo y Felix.[3] Quisiera también darle las gracias a los que ya no veo pero siguen estando ahí: mi abuela, abuelos, y tío. Vuestro apoyo ha sido vital a lo largo de estos más que intensos años. Va por vosotros.

[1] Feel free to drop me an email at adrian.santos1987@gmail.com if you read this thesis. You can also invite me for lunch.

[2] More information at http://www.grise.upm.es/miembros/natalia/

[3] Espero no olvidarme de nadie... Aun así, si me he olvidado de alguien, le recompensaré con una cena (previa demostración del olvido).

Acknowledgements

This doctoral dissertation may never have been possible without the financial support of the Experimental Software Engineering Industrial Laboratory Project (ESEIL Project[4]) and the M3S-ITEE group[5] at the University of Oulu.

[4] http://www.softwareindustryexperiments.org/
[5] http://www.oulu.fi/m3s/

List of abbreviations

AD	*Aggregated data*
ANOVA	*Analysis of variance*
CI	*Confidence interval*
IPD	*Individual participant data*
ITL	*Iterative test-last*
LMM	*Linear mixed model*
MCT	*Multicenter clinical trial*
SE	*Software engineering*
SLR	*Systematic literature review*
SMS	*Systematic mapping study*
TDD	*Test-driven development*

List of original publications

This dissertation is based on the following articles, which are referred to in the text by their Roman numerals (I–V):

I Santos, A., Spisak, J., Oivo, M., and Juristo, N. (2018). Improving Development Practices through Experimentation: an Industrial TDD Case. In *Asia-Pacific Software Engineering Conference* (pp. 465, 473). IEEE doi: 10.1109/APSEC.2018.00061

II Santos, A., Gómez, O. S., and Juristo, N., Analyzing Families of Experiments in SE: a Systematic Mapping Study, in *IEEE Transactions on Software Engineering.* doi: 10.1109/TSE.2018.2864633.

III Santos, A., and Juristo, N. 2018. Comparing Techniques for Aggregating Interrelated Replications in Software Engineering. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '18)*. ACM, New York, NY, USA, Article 8, 10 pages. DOI: https://doi.org/10.1145/3239235.3239239. *Best paper award.*

IV Santos, A., Vegas S., Oivo M. and Juristo N. A Procedure and Guidelines for Analyzing Groups of Software Engineering Replications, in *IEEE Transactions on Software Engineering* doi: 10.1109/TSE.2019.2935720

V Santos, A., Järvinen, J., Partanen, J., Oivo, M., and Juristo, N. (2018, November). Does the Performance of TDD hold across Software Companies and Premises? A Group of Industrial Experiments on TDD. In *International Conference on Product-Focused Software Process Improvement* (pp. 227-242). Springer, Cham.

Contents

1 Introduction

The motivation behind this doctoral dissertation is outlined in Section 1.1. The background is given in Section 1.2, and the objectives in Section 1.3. Finally, the structure of the doctoral dissertation is presented in Section 1.4.

1.1 Motivation

Experiments are commonplace in software engineering (SE) (B. A. Kitchenham, Budgen, & Brereton, 2015; Stol & Fitzgerald, 2018). Experiments allow for identifying cause and effect relationships (Sjoberg et al., 2002; Sjøberg et al., 2003), or facilitating the assessment of new theories or hypotheses, among others (Falessi et al., 2017; Sjoberg, Dyba, & Jorgensen, 2007).

Still, two main shortcomings usually impact the suitability of individual experiments when it comes to assessing the effectiveness of SE *treatments* (i.e., methods, processes, and tools (B. Kitchenham, 2004)): (1) the results[6] are only interpretable within the configuration of the experimental settings (Wohlin et al., 2012), and (2) results may be unreliable due to the low number of subjects (i.e., the small sample size) typically participating in SE experiments (Dybå, Kampenes, & Sjøberg, 2006).

For instance, the typical SE experiment evaluates the performance of a *binary treatment* (e.g., Treatment A and Treatment B (Dybå et al., 2006)) on a *continuous outcome of interest* (e.g., quality measured as the number of bugs per KLOC (Dybå et al., 2006)) on a certain *experimental configuration* (i.e., a combination of programming language, programming environment, experimental task, etc. (Juristo & Moreno, 2013; Wohlin et al., 2012)). This makes the generalization of the results to other settings untenable: what happens in a certain configuration may not happen in another. In experimental design parlance, experiments' configurations pose limits to the external validity of the results.

Besides, the stereotypical SE experiment involves *around 30 participants* (i.e., human-subjects (Dybå et al., 2006)) evaluating the treatments. As a consequence of such a small number of participants, a large variability in the results is expected (Cumming, 2013). Intuitively, because only a few subjects participate in SE experiments, SE experiments' results depend upon those achieved by only a small *sample* of participants

[6] Albeit both *qualitative* results (e.g., text transcripts, video recordings, etc.) and *quantitative* results (e.g., quality scores in a continuous scale) can be collected during experiments, in the following we only focus on *quantitative* results, and the techniques that have been used to aggregate them in SE.

(Cumming, 2013). This may lead to either exaggerated treatment effects (also known as small-study effects (Schwarzer, Carpenter, & Rücker, 2015)), or toward missing real differences between treatments' effects. In experimental design parlance, experiments' small sample sizes pose threats to the conclusion validity of the results.

With the aim of overcoming the shortcomings of individual SE experiments, it is increasingly common among SE researchers to collaborate towards the construction of groups of experiments by means of replication (i.e., to conduct *families of experiments*) (Abrahao, Gravino, Insfran, Scanniello, & Tortora, 2013; Canfora, García, Piattini, Ruiz, & Visaggio, 2005; Kosar, Mernik, & Carver, 2012; Krein et al., 2016; Mouchawrab, Briand, Labiche, & Di Penta, 2011; Muñoz, Mazón, & Trujillo, 2010). Collaborating with each other (e.g., by sharing instructional and experimental material and assisting each other during the design, execution and analysis phases of the experiments, etc.), researchers are able to increase the total number of participants (i.e., the sample size), and, at the same time, evaluate the effects of the treatments under different settings. This should increase the reliability of the results, and, eventually, their generalizability towards different contexts and populations (Basili, Shull, & Lanubile, 1999).

Families provide certain advantages for evaluating the effectiveness of SE treatments (Biondi-Zoccai, 2016; Cooper & Patall, 2009; Debray et al., 2015; Lyman & Kuderer, 2005; Stewart & Tierney, 2002): (1) because access to the *raw data*[7] is granted in families, researchers can apply consistent pre-processing and analysis techniques to analyze the experiments, and, in turn, increase the reliability of joint conclusions; (2) researchers conducting families may opt to reduce the amount of changes made across the experiments with the aim of increasing the internal validity of joint conclusions; (3) because families do not rely on already published results, joint conclusions are not affected by the detrimental effects of publication bias; and (4) because researchers can use identical measurement instruments across the experiments, the researchers can measure the participants' characteristics (e.g., their experience with programming, etc.) with consistent methods and scales, eventually, stratifying the results according to such characteristics.

Applying unsuitable techniques to aggregate experiments' results within families may undermine their potential to provide in-depth insights from the experiments' results. This may translate into wasted effort and resources (e.g., those involved in coordinating the researchers from different institutions, paying travelling and accommodation costs

[7] Here we consider the raw data as a spread-sheet, including the assignment of the participants to the treatments across the experiments, and the scores of the participants with each treatment on a certain *outcome of interest* (e.g., quality). The raw-data may also contain characteristics of the participants (e.g., their programming experience, etc.) or the experiments (e.g., the programming languages, or the testing tools used in the experiments.

for visiting researchers during the execution of experiments, the costs of instructional and experimental material translations, etc.).

The rest of the current dissertation focuses on how to *analyze* a *stereotypical SE family* to provide reliable joint conclusions or moderator effects (i.e., variables rather than the treatments that affect the results). Some aspects and "burning" issues in SE experimentation are left out of this thesis' scope:

- How to account for the "potentially shared bias" of identical replications (B. Kitchenham, 2008) when analyzing the data.
- How to use SE families' results to inform experimental design issues. For example, what type of replication (e.g., identical, conceptual (Gómez, Juristo, & Vegas, 2014)) should be run next in view of the family's results.

In the next section, we provide an overview of the way that SE experiments are typically analyzed, and then, discuss what has been said about SE families of experiments.

1.2 Background

Broadly speaking, SE experiments' results are typically conveyed in terms of effect sizes, *p*-values, and 95% confidence intervals (95% CIs) (Juristo & Moreno, 2013; Kampenes, Dybå, Hannay, & Sjøberg, 2007; Wohlin et al., 2012).

Effect sizes quantify the *magnitude* and *sign* of the relationship between two groups (or, more generally, between two variables: the independent and dependent variables (Borenstein, Hedges, Higgins, & Rothstein, 2011)). The larger the magnitude, the larger the relevance of the relationship and viceversa. A commonly used *standardized* effect size in SE is Cohen's d (Kampenes et al., 2007) (which quantifies the number of standard deviations that the *mean* effectiveness of one treatment is above—or below—the *mean* effectiveness of another (Cumming, 2013)). A commonly used *unstandardized* effect size in SE are *t*-test's estimates (Dybå et al., 2006) (that quantify the difference between the *mean* effectiveness of two groups in natural units (Cumming, 2013)).

p-values inform about the probability of observing the effect size obtained in the experiment—or a larger one—if the effect size in the population was equal to 0 (Cumming, 2013)). If the *p*-value is lower than a certain threshold— typically set at 0.05 (Cohen, 1994)—then it is claimed that the effect size is *statistically significant*. If the effect size is statistically significant, then it is unlikely that such effect size could have been observed if the effect size in the population was equal to 0 (i.e., if there was no effect in the population, or the true effect was equal to 0). In general, statistically

significant effect sizes tend to be associated with experiments with large sample sizes and non-statistically significant effect sizes with experiments with small sample sizes (Cumming, 2013).

In loose terms (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016), 95% CIs are usually thought of a range of effect sizes plausible in the population given the data of the experiment (Cumming, 2013; Higgins & Green, 2011; Whitehead, 2002). In particular, 95% CIs are commonly used as a measure of the precision of the effect size provided (Cumming, 2013): the wider the 95% CI, the wider the array of compatible effect sizes in the population, and thus, the less precise the effect size and viceversa. Narrow 95% CIs (i.e., precise 95% CIs) tend to be associated with experiments with large sample sizes and wide 95% CIs with experiments with small sample sizes (Cumming, 2013).

Unfortunately, individual experiments' results may be unreliable. This is because SE experiments' sample sizes are typically small, so their results may be way off the real effects for the population (Dybå et al., 2006). Besides, SE experiments' contrived settings hinders the generalization of the results towards different contexts (Stol & Fitzgerald, 2018).

Back in 1999, Basili et al. (Basili et al., 1999) already devised a way of moving beyond individual experiment's limitations. They proposed conducting a series of experiments that *pursued the same goal to extract mature conclusions*. They named such series of experiments as *families of experiments* (Basili et al., 1999). Over the years, numerous benefits have been attributed to SE families, including: (1) increasing the reliability of the findings (Gómez et al., 2014); (2) increasing the precision of the results (Fernández, Dieste, Pesado, & García Martínez, 2011); (3) the possibility to assess the impact of the changes made across the experiments on the results (i.e., evaluating the effect of moderator variables, or simply identifying *moderators* (Ciolkowski, Shull, & Biffl, 2002; Juristo & Vegas, 2009)), among others.

Despite the intuitive definition provided by Basili et al. for SE families (Basili et al., 1999), such a definition does not set apart two—in our opinion—different types of groups of experiments: those gathered by means of systematic literature reviews (SLRs), and those gathered by means of experimental replication. Our rationale is that because each group of experiments provides access to *differently granular information* (e.g., summary statistics vs. raw data, reported settings vs. firsthand knowledge of the settings, etc.), they may serve different purposes.[8]

[8] Along the next lines, we try to convey the idea that SLRs and SE families are different entities. As such, each have their unique characteristics in terms of scope, data access, and control—and knowledge—over the experimental settings.

For example, although the scope of SLRs is usually wide (because all the available research on a particular topic is aimed to be brought together into a joint conclusion), the scope in groups of experiments conducted by means of replication tends to be narrower (since a small set of hypotheses on a limited set of outcomes is usually of interest).

Besides, only what is reported in research articles is known in SLRs. Thus, some relevant information (e.g., experimental configurations' characteristics, participants' characteristics, etc.) may pass unnoticed if not entirely reported. For example, due to length restrictions or reporting inconsistencies. On the contrary, in groups of experiments conducted by closely collaborating researchers, it is typical that researchers share among them laboratory packages, instructional or experimental material (Shull et al., 2004), assist each other via in-person and Internet meetings during the planification, design, execution, and analysis phases of the experiments, and so forth (Carver, Juristo, Baldassarre, & Vegas, 2014). Eventually, this close collaboration may lead to *greater knowledge about the characteristic of the experiments and the participants*, along with access to *the raw data* of the experiments. This may increase the reliability of the joint conclusions—because after all, it is possible to apply consistent pre-processing and analysis techniques to analyze the experiments—and in case the results differ across experiments, ease the identification of variables (i.e., moderators) that may be causing such differences. Such moderator variables may be either at the *experiment level* (e.g., the characteristics of the experiments, such as their programming languages), or at the *participant level* (e.g., the characteristics of the participants, such as their programming experience). Summarizing, we propose the following:

Key point
Families of experiments are groups of experiments conducted by means of replication where researchers have access to the *raw data*, and where researchers have *first-hand knowledge* of the experiments' and participants' characteristics.

Although *meta-analysis of effect sizes* has been suggested to aggregate experiments' results in SLRs (B. Kitchenham, 2004), we wonder whether its application would also be recommendable in SE families. After all, meta-analysis of effect sizes only requires summary-level statistics (such as means, standard deviations and sample sizes) to produce results (Borenstein et al., 2011). SE families grant access not only to summary statistics, but also to the raw-data. Would this not allow for applying more "flexible" aggregation techniques that leverage all the information contained within the raw-data?

The next section outlines the objectives, research questions, and contributions of this doctoral dissertation.

1.3 Objectives, research questions and contributions

The main **objectives** of the current dissertation are: (1) realizing the limitations of individual SE experiments; (2) mapping and grouping the techniques used to aggregate experiments' results within SE families; (3) learning about the advantages and disadvantages of such techniques in the SE context; and (4) providing a set of guidelines to analyze SE families. We aim to meet such objectives by answering a series of **research questions**:

- **RQ1:** What are the limitations of the results reached in individual SE experiments?
- **RQ2:** What techniques have been used to aggregate experiments' results within SE families?
- **RQ3:** Do the advantages and disadvantages of the aggregation techniques according to mature experimental disciplines materialize in the SE context?
- **RQ4:** How should SE families be analyzed?
- **RQ5:** What information can we extract from an industrial family of experiments on test-driven development (TDD) (Beck, 2003) with the aggregation techniques recommended in mature experimental disciplines?

The main **contributions** of this doctoral dissertation are as follows:

- **Contribution 1:** *an evaluation* of the shortcomings of the results of an industrial experiment on TDD.
- **Contribution 2:** a *refinement of the concept of family of experiments*, *an updated picture* of the aggregation techniques used in SE to analyze families, *a list of the advantages and disadvantages* of such techniques according to mature experimental disciplines, and preliminary advice to *analyze* and *report* families based on their common limitations regarding joint data analysis practices.
- **Contribution 3:** an *illustrative case* showing the application of the aggregation techniques on a representative SE family, and *a comparison of the findings, advantages, and disadvantages* of such techniques in the SE context.
- **Contribution 4:** a *procedure with a set of embedded guidelines* to analyze SE families, and a discussion about the particularities of SE families regarding their counter-parts in mature experimental disciplines.
- **Contribution 5:** an *analysis* of an industrial family of experiments on TDD with the techniques applied in mature experimental disciplines to learn whether TDD's effectiveness holds across companies and sites.

1.4 Thesis organization

The rest of the current doctoral dissertation is organized as follows. The research method is outlined in Chapter 2. The findings of this doctoral dissertation are discussed in Chapter 3. Chapter 4 shows further lines of research and provides the conclusions.

2 Research method

This chapter includes one section per each of the phases of research of the current dissertation: problem definition, solution proposal, and solution evaluation (Section 2.1, Section 2.2 and Section 2.3, respectively). Each section also contains an overview of the articles included within each phase of the research. This chapter ends with an extra section (Section 2.4) outlining the materials and resources used in the dissertation.

2.1 Problem definition

When starting a new piece of research, it is first necessary to identify a relevant problem in the SE community.

We realized that the individual experiments that we run on TDD in industry provided inaccurate results. In particular, in *Paper I* we report an experiment where we evaluated the effectiveness of the traditional way of coding at Paf[9] (YW), iterative-test last (i.e., ITL, the reverse approach of TDD following Erdogmus et al. (Erdogmus, Morisio, & Torchiano, 2005)), and TDD in terms of external quality.

We used descriptive statistics (i.e., means, medians, etc.), data visualizations (i.e., violin plots, box plots (Field, 2013)), linear mixed models (LMMs) (Brown & Prescott, 2014), and parametric effect sizes (i.e., Hedges' g (Borenstein et al., 2011)) for analyzing the data.

We noticed that the experiment's results where inaccurate. Thus, we could not draw definite conclusions on TDD's effectiveness for Paf. Besides, Paf's results could only be interpreted within the context of a specific set up (i.e., Java, JUnit). This limited the generalizability of the findings.

Unfortunately, increasing the experiment's sample size to obtain more accurate results was not feasible (because Paf had a limited amount of available developers). The only way to increase the reliability of results was running replications across different sites, and then aggregating their results together into joint conclusions. This point of view has been shared by the SE community since the 90s (Basili et al., 1999).

However, we soon came across a major obstacle, which became the main motivation of this thesis: How should experimental results be reliably aggregated? To the best of our knowledge, and after conducting an informal search, we realized that SE researchers followed various approaches: providing a narrative summary of results (Porter & Votta,

[9]Paf is an online gaming entertainment company (http://www.paf.com) that collaborated with the ESEIL project.

1998), aggregating the results with meta-analysis (Abrahao et al., 2013) (as typically done in SLRs (B. Kitchenham, 2004)), or analyzing the data of all the experiments as coming from a single big experiment (George & Williams, 2004). To obtain a broader picture of the aggregation techniques being applied to analyze SE families, we conducted a systematic mapping study (SMS[10]). In particular, we run such SMS (i.e., *Paper II*) with the following objectives:

- Scoping the research to narrow down the concept of *family of experiments* and identifying the characteristics that differentiate families from other types of groups of experiments.
- Building a *broad classification* of the aggregation techniques used in SE for providing joint conclusions.
- Understanding the suitability, advantages, and disadvantages of each technique according to mature experimental disciplines.

First, we searched for SE families in commonly used on-line databases (i.e., IEEE Xplore, ISI Web of Science, Science Direct and Scopus) with the terms "experiment," "family," "series," "group," or "replication". However, due to the looseness of such terms we retrieved an unmanageable list of references. To narrow down the search space, we refined the search terms, and restricted the search scope to well-known venues on SE experimental research (i.e., those used by Sjoberg et al. (Sjøberg et al., 2005)).

At the end of this new search, we identified a total of 39 valid primary studies. Each primary study reported a family of experiments. We collected certain information from each family. With this information, we obtained a picture of the typical characteristics of SE families. Also, a picture of the aggregation techniques commonly used to aggregate experiments' results within SE families.

In addition, we gained knowledge about the advantages and disadvantages of each aggregation technique from well-known references in mature experimental disciplines such as medicine and pharmacology (Higgins & Green, 2011), works on vote-counting and narrative synthesis (Hedges & Olkin, 1979, 1980), and meta-analysis and reproducibility of results (Borenstein et al., 2011; Ioannidis, Patsopoulos, & Rothstein, 2008; Petitti et al., 2000).

We used visualizations (i.e., histograms, trend-plots, etc.), and descriptive statistics (i.e., means, medians) to describe SE families' characteristics. We used archival research (i.e., literature review (Wohlin & Aurum, 2015)) on mature experimental disciplines to learn about the advantages and disadvantages of the aggregation techniques.

[10]SMS are secondary studies that provide a coarse-grained overview of a SE field of interest (Petersen, Feldt, Mujtaba, & Mattsson, 2008).

In sum, we noticed that: (1) SE researchers applied heterogeneous aggregation techniques to analyze families, and (2) most of the aggregation techniques applied in SE families were dismissed in mature experimental disciplines (such as medicine or pharmacology). Thus, the research that led to Paper II allowed us to obtain evidence that there is a problem in terms of joint data analysis practices in SE families.

2.2 Solution proposal

During this phase of research, we aimed to provide a solution to the identified problem.

With the aim of making "research-in-the-typical" (Wohlin, Höst, & Henningsson, 2003), we selected a *representative* SE family to apply the aggregation techniques we identified along the SMS. Our main objective was assessing whether the advantages and disadvantages of the aggregation techniques acknowledged in mature experimental disciplines would also materialize in the SE context. We also wanted to realize whether the selection of the aggregation technique had an influence on joint results.

To study this, in *Paper III*, we selected a family of experiments from the ESEIL project: a family comprised of four experiments on TDD (i.e., the *median* number of experiments within SE families according to the results of the SMS), with *small* and *dissimilar* sample sizes, *heterogeneous* results, *different* types of subjects (i.e., students and professionals), and *identical* experimental designs and response variable operationalizations. We analyzed such family with the aggregation techniques that we identified along the SMS (i.e., Paper II).

We used descriptive statistics (i.e., means, medians, standard deviations, and sample sizes), data visualizations (i.e., violin plots, box plots, profile plots, regression plots (Field, 2013)), linear regressions (e.g., *t*-test, ANOVAs, etc. (Field, 2013)), linear mixed models (LMMs) (Brown & Prescott, 2014), parametric effect sizes (e.g., Hedges' g (Borenstein et al., 2011)), and meta-analysis models (Borenstein et al., 2011) to analyze the data.

We compared the results obtained with the different aggregation techniques. We noticed that the aggregation technique largely affects the conclusions reached. Thus, in *Paper IV*, we aimed to provide guidance—in the form of a procedure with a set of embedded guidelines—to analyze SE families. Our main objective was removing the typical barriers faced by SE researchers when analyzing families.

With the objective of building a backbone for such guidelines, we studied the guidelines typically followed in mature experimental disciplines to analyze and report groups of experiments. We came across them using a manual search and by following the references included in the articles we read on data analysis in mature experimental

disciplines. Among others, we came across the guidelines provided by the Cochrane Association (Bero & Rennie, 1995), the framework for analyzing multicentre clinical trials (MCTs) promoted by the American Food and Drug Administration (FDA) (Anello, O'Neill, & Dubey, 2005), the guidelines for analyzing MCTs provided by the International Conference on Harmonization (ICH-9) (Lewis, 1999), the PRISMA-IPD statement of the EQUATOR Network framework (Stewart et al., 2015), and the CONSORT statement for reporting randomized controlled trials (Schulz, Altman, & Moher, 2010).

After studying them, we acknowledged certain differences between SE families and MCTs (in terms of sample size, level of adherence to prespecified protocols, etc.). Such differences had certain implications in terms of data analysis. This made the direct application of MCTs guidelines unsuitable to analyze SE families. In view of this, we built a procedure with a set of embedded guidelines tailored to analyze SE families. We illustrated the use of the procedure by applying it to analyze Paper III's family of experiments. We chose again such family of experiments as a representative of the stereotypical SE family.

We used descriptive statistics (i.e., means, medians, standard deviations, and sample sizes), data visualizations (i.e., violin plots, box plots, profile plots, regression plots (Field, 2013)), linear regressions (e.g., t-test, etc. (Field, 2013)), linear mixed models (LMMs) (Brown & Prescott, 2014), parametric effect sizes (e.g., Hedges' g (Borenstein et al., 2011)), and meta-analysis models (Borenstein et al., 2011) to analyze the data.

Once we had built the guidelines, and applied them to analyze a representative SE family, we were ready to apply them for obtaining new knowledge from SE families.

2.3 Evaluation

During this phase of research it is required to evaluate the proposed solution.

To evaluate the analysis procedure that we proposed, we applied it to analyze all the experiments on TDD from the ESEIL project.[11] We also analyzed a family of industrial experiments on TDD with the aggregation techniques recommended in mature experimental disciplines. We did so with the objective of learning whether TDD's effectiveness on quality held across different sites at two companies (i.e., F-Secure and Bittium). We report the findings in *Paper V*.

We used descriptive statistics (i.e., means, medians, standard deviations, and sample sizes) and data visualizations (i.e., violin plots, box plots, profile plots, regression plots

[11] We decided not to include such article along this dissertation as the article is still under review. Santos et al., A Family of Experiments on TDD. Submitted to *Transactions in Software Engineering (TSE)*, October 2019.

(Field, 2013)) to describe the data. We applied meta-analysis and linear mixed models (LMMs) to learn whether TDD's effectiveness held across sites and companies, and the extent to which the characteristics of the participants affected TDD's effectiveness.

2.4 Materials and resources

We accessed all the articles, monographs, theses, and books mentioned in the current doctoral dissertation with the credentials of the University of Oulu. In private Dropbox folders, we stored all the experimental material used for conducting the experiments from the ESEIL project (http://www.softwareindustryexperiments.org/). We stored the datasets and questionnaires of all the experiments in private Dropbox and Google Drive folders. We anonymized all datasets and questionnaires' responses so as to preclude the identification of the participants. Finally, we stored all the experimental solutions such as code, program stubs, and test cases within a Transporter device[12] installed at the University of Oulu, Finland. We mirrored such information in another Transporter device installed at the Universidad Politécnica de Madrid, Spain. All data providers offered recovery solutions in case of data losses.

[12]http://filetransporterstore.com/

3 Findings

Table 1 (see the next page) maps the research questions of the current doctoral dissertation, the articles where they are answered, and the corresponding findings and sub-findings that emerged during the research. In the following sections, we go over the findings of this doctoral dissertation.

Table 1. Research questions, papers, findings and sub-findings.

RQ	Paper	Finding	Sub-finding
RQ1	I	Limitations of individual experiments on TDD	✓ Low precision of results due to small sample sizes
RQ2	II	Techniques used for analyzing SE families	✓ Definition and characteristics of SE families ✓ Aggregation techniques, advantages and disadvantages ✓ Preliminary advice for analyzing and reporting families
RQ3	III	Assessment of the techniques in the SE context	✓ Assessment of the techniques on a representative SE family ✓ Preliminary advice for analyzing families
RQ4	IV	Procedure to analyze SE families	✓ Differences between MCTs and SE families and statistical consequences ✓ Guidelines to analyze SE families and illustrative example
RQ5	V	Analysis of an industrial family of experiments on TDD	✓ TDD's effectiveness across companies ✓ TDD's effectiveness across sites

3.1 Finding 1: limitations of individual experiments on TDD

We conducted an experiment on TDD at Paf and reflected about the limitations of its results in terms of precision (see Section 3.1.1).

3.1.1 Low precision of results due to small sample sizes

Paf's managers at Helsinki, Finland, were interested in showing their software development team the advantages of weaving testing and coding. With such objective in mind, we proposed that they should try test-driven development (TDD) (Beck, 2003). Long story short, TDD is an agile software development approach that enforces the construction of software systems by means of small and continuous testing coding cycles. According to its proponents (Beck, 2003), TDD's strong emphasis on testing along the development process pays-off in the long term—in terms of quality, code maintainability, and so forth.

We proposed Paf's managers to run an experiment on TDD to obtain evidence of its effectiveness in terms of quality. If Paf's developers noticed an improvement in the quality of their software products using TDD, then, this may encourage them to adopt TDD in their daily work. Because none of Paf's developers had any previous experience with TDD, we embedded the experiment within a week-long training course on TDD. In particular, we trained Paf's developers in: (1) how to "slice" tasks specifications into smaller ones and how to develop such slices in short incremental development cycles; (2) how to embed unit testing and refactoring within such development cycles; and (3) the mechanics of ITL and TDD.

Table 2. Experimental design: Paf.

Group	YW	ITL	TDD
G1	SS	BSK	MR
G2	MR	SS	BSK
G3	BSK	MR	SS

A total of 15 subjects participated in the experiment. The subjects had to develop three different toy tasks (i.e., BSK, MR, SS), each with a different development approach (i.e., their traditional way of coding—,here being YW—,ITL, or TDD). YW was applied on the first day, ITL on the second, and TDD on the third. The participants were not trained in either ITL or TDD until the day of the experimental session. We made this decision so as to avoid the possibility that the participants could apply a mixed development approach—for example, if the participants were knowledgeable of all

the development approaches before the experimental sessions took place. This may enhance treatment adherence. We distributed the participants to three different groups to balance out the influence of the task on the development approaches. Table 2 shows the assignment of groups to the development approaches and tasks.

Table 3 provides an overview of the experiment's settings.

Table 3. Paf experiment's settings. **Reprinted by permission [Paper I] © IEEE 2018.**

Aspect	Values
Development approach	YW. vs. ITL vs. TDD
Tasks	BSK vs. MR vs. SS
Response variable	QLTY
Design	Within-subjects design
Training	TDD seminar
Training duration	3 days/6 hours
Experiment duration	2.25 hours
Technological environment	Java, Eclipse, JUnit

We measured quality as the *percentage of test cases that successfully passed* from a battery of test cases that we (i.e., the experimenters) built for measuring the participants' solutions. Specifically, we measured quality as

$$QLTY = \frac{\#Test\ Cases(Pass)}{\#Test\ Cases(All)} * 100\%.$$

Table 4 shows the descriptive statistics (i.e., mean, sd, median) of the QLTY scores achieved with each treatment (i.e., YW, ITL and TDD).

Table 4. Descriptive statistics: YW vs. ITL vs. TDD. **Reprinted by permission [Paper I] © IEEE 2018.**

Treatment	Mean	SD	Median
YW	53.65	34.12	53.18
ITL	50.43	32.77	46.37
TDD	67.64	26.24	70.78

Figure 1 shows the box plots and violin plots corresponding to the descriptive statistics.

As shown in Table 4 and Figure 1, TDD's QLTY scores seem larger and less spread than those of YW and ITL. On the contrary, YW and ITL's QLTY scores seem more similar to each other. Not large deviations from normality are expected in view of the data distributions (as all distributions in Figure 1 look bell-shaped).

We ran a LMM to analyze the data. We used pairwise contrasts to learn about the difference in effectiveness between the treatments. Table 5 shows the results of the pairwise contrasts (i.e., their estimates, standard errors (SEM), and p-values).

Table 5. Pairwise contrasts on treatments. **Reprinted by permission [Paper I] © IEEE 2018.**

Contrast	Estimate	SEM	p-value
YW vs. ITL	6.01	18.48	0.94
YW vs. TDD	-0.33	18.48	0.99
ITL vs. TDD	-6.34	20.24	0.95

As we can see in Table 5, TDD outperforms YW ($M = -0.33$) and ITL ($M = -6.34$) but to a negligible extent. Besides, relatively large SEMs materialized (i.e., $SEM >= 18.48$) and thus, approximate 95% CIs around the estimates (i.e., estimate+-1.96*SEM (Cumming, 2013)) indicate that the results are imprecise. This can be corroborated by looking at the p-values (as the p-values are close to 1). Put differently, even though TDD outperformed YW and ITL at Paf, the results may have been due to chance. Summarizing:

Key point
Paf experiment's small sample size led to imprecise results. More experiments are needed to assess the extent to which TDD performs in terms of quality.

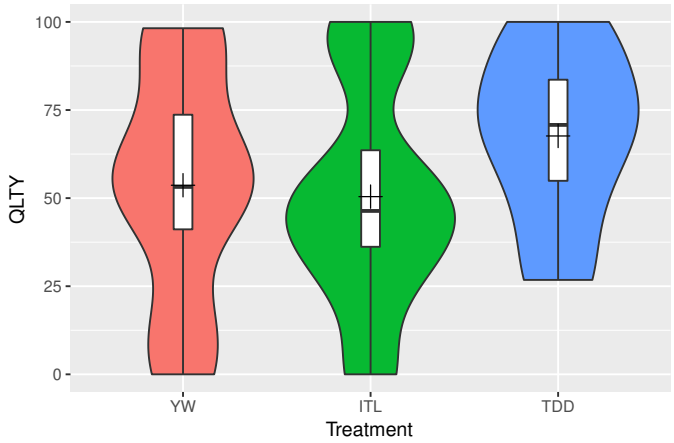

Fig. 1. Box plot and violin plot: YW vs. ITL vs. TDD. **Reprinted by permission [Paper I] © IEEE 2018.**

3.2 Finding 2: techniques used for analyzing SE families

We conducted a SMS to learn about the aggregation techniques used in SE to aggregate experiments' results within families. Eventually, the aggregated results may allow us to move beyond the common limitations with individual experiments' results. During the SMS, we made a series of findings. First, we refined the concept of SE family of experiments, and learned about their typical characteristics (see Section 3.2.1). Then, we identified the aggregation techniques that had been used to analyze SE families, and learned about the advantages and disadvantages of each aggregation technique according to the literature of mature experimental disciplines (see Section 3.2.2). Finally, we provided preliminary advice on how to *analyze* and *report* SE families in view of their common limitations regarding joint data analysis practices (see Section 3.2.3).

3.2.1 Definition and characteristics of SE families

We started the research following Basili et al.'s definition of family of experiments (Basili et al., 1999): a group of experiments that pursue the same goal and whose results can be combined into joint—and potentially more mature—conclusions than those of individual experiments. However, we soon realized that such definition did not set apart two different types of groups of experiments: those gathered by means of *replication* and those gathered by means of *SLRs*. We also realized that Basili et al.'s definition did not provide clear cut-off points for discerning between: (1) a series of planned, and coordinated replications conducted by a sole researcher—or group of collaborating researchers—and; (2) isolated replications conducted by researchers who neither interacted nor collaborated with each other. Such definition of family of experiments was also vague in other aspects:

- Could also *objects*—instead of *human subjects*—apply the treatments?
- Could the same *human subjects* participate across multiple experiments within the same family?
- Could *a sole treatment* be exercised within families (so no comparison with any other control treatment could be performed)?
- Could also *pilot* experiments (i.e., small experiments usually run to assess the suitability of the experimental materials (Wohlin et al., 2012)) be considered as experiments within families?
- *How many experiments* should a family contain (i.e., two, three, four...)?

To set the scope of the research and to narrow down the definition of family of experiments, we tuned Basili et al.'s definition with the following criteria:

– *Access to the raw data should be guaranteed* within families: so it is possible to analyze all the experiments with consistent pre-processing and statistical techniques (e.g., a *t*-test (Field, 2013)). This may aid to ensure that differences across experiments' results are due to real differences in the data gathered, not to the different procedures followed to analyze each experiment's data.
– *First-hand knowledge of the characteristics of the experiments and the participants should be guaranteed within families*: so it is possible hypothesizing on variables—at both the experiment level and participant level—behind the potential differences of results observed across the experiments.
– *Human-subjects should participate across the experiments, and different subjects should participate in each experiment*: otherwise, if the correlation between experiments' results is not correctly accounted for within the aggregation technique/s used, this may bias joint conclusions (Borenstein et al., 2011).
– *At least two treatments across at least three experiments should be assessed on the same outcome of interest*. This way, it is possible to (1) compare the effectiveness of two treatments under the same circumstances on the the same outcome of interest; (2) check the statistical assumptions of the aggregation techniques if required (e.g., the normality assumption (Field, 2013)); and (3) apply all the aggregation techniques to assess whether the characteristics of the experiments impact the results—as some techniques such as *meta-regression* (Borenstein et al., 2011) cannot be applied if fewer than three experiments are aggregated together, because a regression line with only two data-points (i.e., experiments) explains all the variability in the data.

We found that SE families share certain characteristics. In particular, families are usually comprised of *a low number of experiments* (i.e., a median number of four experiments), with *small* (i.e., smaller than 30 (B. Kitchenham et al., 2017)) and *dissimilar sample sizes* (i.e., families contained at least one experiment doubling the size of the smaller experiment). In addition, experiments commonly provide *heterogeneous results*, evaluate the effects of the treatments on *different types of subjects* (e.g., students vs. professionals), and use *identical experimental designs* and *response variable operationalizations*. Summarizing:

Key points

- We define a group of experiments as a family if access to the *raw data* is guaranteed, the researchers have first hand knowledge of the characteristics of the experimental *settings* and the participants, and if *at least three* experiments involving *different* human-subjects evaluate the effects of *at least two* treatments on the *same* outcome of interest.
- Families are usually comprised of a *low* number of experiments, with *dissimilar* and *small* sample sizes, *identical* experimental designs and response variable operationalizations and different *types of subjects*; in addition, they provide *heterogeneous* results.

3.2.2 Aggregation techniques, advantages and disadvantages

We grouped the aggregation techniques used to analyze SE families into five broad categories (from the most to least used): (1) narrative synthesis; (2) aggregated data (AD); (3) individual participant data mega-trial (IPD-MT); (4) individual participant data stratified (IPD-S); and (5) aggregation of *p*-values.

Narrative synthesis was used to analyze 46% of the families (Ali, Yue, & Rubab, 2014; Fernandez, Abrahão, & Insfran, 2013; Juristo & Vegas, 2011; Reynoso, Genero, Piattini, & Manso, 2005). Narrative synthesis relies on the synthesist's ability to condense, integrate and extrapolate the findings of multiple experiments into a joint *textual* conclusion (Borenstein et al., 2011). This synthesis process has also been termed "cognitive algebra" in other disciplines such as medicine: a set of opaque rules applied by researchers to integrate the findings from various experiments together (Cooper & Patall, 2009). SE authors applying narrative synthesis usually follow a *template* for analyzing families such as: *"...while the results are statistically significant/large/small in experiment A, they are not in experiment B and C. [Optional] This difference of results may be due to of X/Y/Z variables..."*

Despite its *intuitiveness*, and its ability to provide joint conclusions from experiments with *different designs* and *response variables*, narrative synthesis fails to provide a *quantitative summary of results* (as a joint effect size or *p*-value). Besides, narrative synthesis involves *subjective judgment* when providing joint conclusions (e.g., should all experiments be identically weighted towards the joint conclusion? Should experiments with professionals be weighted more—because after all, they may be more representative—than those with students?). In sum, narrative synthesis may hinder the *reproducibility* of results, and, eventually, the reliability of families' findings.

Aggregated data (AD) was used to analyze 38% of the families of experiments (Gonzalez-Huerta, Insfran, Abrahão, & Scanniello, 2015; Manso, Cruz-Lemus, Genero, & Piattini, 2008). AD is commonly known as *meta-analysis of effect sizes* in SE (B. Kitchenham, 2004). Generally speaking, AD is a set of statistical techniques that deliver a *weighted* effect size as a joint conclusion, where the weight of each experiment's effect size towards the joint conclusion is proportional to the experiment's sample size—if a fixed-effects model is used—or to the experiment's sample size, and the total heterogeneity of results (i.e., the variation of results that cannot be explained only by natural variation (Borenstein et al., 2011))—if a random-effects model is used.[13]

AD has some advantages: (1) it provides *intuitive visualizations* for conveying joint conclusions (i.e., forest plots (Borenstein et al., 2011)); (2) it provides *straightforward statistics* for interpreting *statistical heterogeneity* (e.g., the I^2 statistic (Borenstein et al., 2011)); and (3) it allows for investigating experiment level moderators (e.g., with sub-group meta-analysis or meta-regression (Borenstein et al., 2011)).

However, AD is no silver bullet. For example, AD cannot easily aggregate the results of *experiments with complex experimental designs*—such as factorial designs (Gillett, 2003)—into joint conclusions. Additionally, using AD also has its own perils. For example, if heterogeneity of results materializes, then, random-effects models (instead of fixed-effects models) should be used to provide joint conclusions (Borenstein et al., 2011). Besides, commonly used effect sizes such as Cohen's d rely on similar *statistical assumptions* as parametric tests, such as the t-test (e.g., the normality and homogeneity of variances assumption (Fritz, Morris, & Richler, 2012; Macbeth, Razumiejczyk, & Ledesma, 2011)). Inattention to such issues may impact the reliability of SE families' results.

Individual participant data (IPD)-MT was used to analyze 33% of the families. One of the main advantages of IPD-MT is its *intuitiveness*. This is because IPD-MT can be simply run by pooling together the raw data of all the experiments, and then, analyzing them as if the raw data were coming from the same experiment—for example, by means of a Wilcoxon test (Field, 2013).

Unfortunately, IPD-MT may provide *biased results* if the data are unbalanced across the treatments and the experiments (as it may happen if missing-data materializes) or if the participants are more similar within experiments than across experiments (as if experiments with either professionals or students are run) (Abo-Zaid et al., 2013; Kraemer, 2000; Quené & Van den Bergh, 2004). In view of this, and because SE families are typically comprised of experiments with different sample sizes, missing data, and different types of subjects, IPD MT may provide unreliable results in SE families.

[13] Assuming a common variance across experiments.

Individual participant data (IPD)-S was used to analyze 15% of the families. In IPD-S, the raw data of all experiments are pooled together and then analyzed jointly by acknowledging where the raw data come from (i.e., by including an extra "experiment" factor within the statistical model fitted—e.g., an ANOVA with "Treatment" and "Experiment" factors (Whitehead, 2002)). Among the main advantages of IPD-S are its *heightened statistical flexibility* for aggregating experiments with missing-data (e.g., with LMMs (Brown & Prescott, 2014)), its ability to provide results in natural units, and its flexibility for assessing both experiment level and participant level moderators (Fisher, Copas, Tierney, & Parmar, 2011).

However, IPD-S is no panacea: some IPD-S statistical models cannot aggregate the results of experiments with *different experimental designs* (e.g., a repeated-measures ANOVA cannot be applied to analyze data where subjects are only measured once (Field, 2013)), and all the experiments being aggregated together with IPD-S need to have *identical response variable operationalizations*. Additionally, IPD-S statistical models rely on statistical assumptions that need checking (e.g., normality and homogeneity of variances in ANOVA (Field, 2013)). Inattention to these issues may impact the reliability of SE families' results if using IPD-S.

Finally, **aggregation of *p*-values** was used to analyze 7% of the families. Aggregation of *p*-values techniques (e.g., Fisher's method (Borenstein et al., 2011)) involve the computation and posterior combination of the one-sided *p*-values obtained from each individual experiment. Among the main advantages of aggregation of *p*-values are its ability to combine the results of experiments with *different designs and response variables* into joint conclusions (because after all, only the experiments' *one-sided* *p*-values are needed to do so). However, the main shortcomings of aggregation of *p*-values' techniques are that is unable to provide a *joint effect size* (and thus, hinder the interpretation of results), and that an *identical weight* is typically assigned to each experiment, regardless of their sample size, design, or quality—unless more advanced aggregation of *p*-values techniques are used (Whitehead, 2002). As a consequence, aggregation of *p*-values may hinder the reliability of families' joint conclusions.

Table 6 summarizes the advantages and disadvantages of each aggregation technique according to the literature of mature experimental disciplines.

3.2.3 Preliminary advice for analyzing and reporting families

We found a series of common limitations regarding joint data analysis practices in the 39 SE families of experiments that we came across:

– **Limitation 1:** we found an over-reliance on *narrative synthesis* to provide joint conclusions and assess moderators. Despite its intuitiveness, narrative synthesis may be misleading in SE due to the common small sample sizes of experiments (Dybå et al., 2006). This is because a large variability of results is expected in small experiments (i.e., also known as small study effects (Nüesch et al., 2010)), and this may translate into "conflicting" results (e.g., positive vs. negative, large vs. small) when there is no conflict in reality (e.g., when the same *population* effect size is being estimated in the experiments (Button et al., 2013)). Narrative synthesis' findings can be especially misleading if differences across experiments' results are attributed to differences across experimental configurations (e.g., professionals vs. students). Particularly, such differences may have had nothing to do with the observed conflicting results (Button et al., 2013).
– **Limitation 2:** the statistical assumptions of *parametric* effect sizes (e.g., Cohen's d (Borenstein et al., 2011)) are typically overlooked within SE families. Particularly, although SE families' authors commonly check traditional statistical tests' assumptions (e.g., the normality assumption in the independent t-test (Field, 2013)), they

Table 6. Analysis techniques' advantages and disadvantages.

Advantages	Technique	Disadvantages
✓ Fast interpretation of results ✓ Intuitive approach ✓ Independent of design, metric, or statistical test	**Narrative synthesis**	✗ No effect size nor p-value ✗ Subjective weighting ✗ Not reproducible results
✓ Independent of design, metric ✓ Straightforward visualizations ✓ Experiment moderators and heterogeneity	**AD**	✗ Statistical assumptions ✗ Simple designs
✓ Intuitive approach	**IPD-MT**	✗ Risk of biased results ✗ Statistical assumptions ✗ Depends on response variable
✓ Increased statistical flexibility ✓ Moderators and heterogeneity ✓ Interpretation in natural units	**IPD-S**	✗ Statistical assumptions ✗ Complexity ✗ Depends on response variable
✓ Independent of design, metric, or statistical test	**Aggregation of p-values**	✗ Identical weight per experiment ✗ No effect size

overlook the fact that commonly used *parametric* effect sizes also rely on similar assumptions (Fritz et al., 2012; Macbeth et al., 2011). For example, in some SE families, the non-normality of the data was acknowledged—and thus, according to the authors, it was not possible to apply parametric tests—but at the same time, they calculated Cohen's d and performed an AD meta-analysis.

– **Limitation 3:** statistical heterogeneity was rarely acknowledged with AD. Despite the appropriateness of fixed-effects models for providing joint conclusions when heterogeneity is not present (Borenstein et al., 2011), random-effects models should be used to account for the heterogeneity of results commonly present within families—at least when heterogeneity is apparent in forest plots—or in families where many changes have been made across the experiments (because such changes may translate into statistical heterogeneity (Borenstein et al., 2011; Higgins & Green, 2011; Whitehead, 2002)).
– **Limitation 4:** there is an over-reliance on IPD-MT to provide joint conclusions. Despite its intuitiveness, IPD-MT provides misleading results if subjects are more similar within experiments than across experiments (e.g., if experiments with either professionals or students are run), or when the experiments have different sample sizes and missing data materializes (as may happen due to protocol deviators or drop-outs) (Abo-Zaid et al., 2013; Kraemer, 2000). In view of this, and because different sample sizes are common in SE families, different types of subjects are commonly assessed, and missing data may materialize, IPD-MT seems unsuitable for analyzing SE families.
– **Limitation 5:** the unique characteristics of the experiments (e.g., their programming languages) or of the participants (e.g., their programming experience) may not be the only variables responsible for the differences across experiments' results. In particular, and even considering that SE families commonly contain changes across experiments, differences across experiments' results could also be due to the existence of confounding variables impacting the results (e.g., when more than one change is simultaneously made across experiments), or the existence of unknown elements impacting the experiments' results (e.g., the materialization of threats to validity in some experiments, treatment conformance, etc.). Confounding is rarely acknowledged within SE families.

In view of these limitations, we propose the following advice for *analyzing* SE families:

Advice for analyzing families

- *Avoid narrative synthesis* because of its dangers in the presence of small sample sizes.
- *Avoid aggregation of p-values* because it hinders the interpretation of results, and weighs identically each experiment towards the joint conclusion.
- *Avoid IPD MT*. Use IPD-S instead.
- *Be consistent when using statistical tests and effect sizes*: if parametric tests are unsuitable, parametric effect sizes may not be suitable either.
- *Check AD and IPD-S statistical assumptions* before interpreting joint conclusions.
- *Acknowledge heterogeneity of results* when providing joint conclusions.
- *Acknowledge that other sources of variability* (e.g., confounding variables, unacknowledged variables, etc.) *may also impact the results*.

In addition, we experienced some difficulties when assessing the suitability of the aggregation technique/s applied to analyze SE families. This was in part because families did not provide the raw data—and thus, re-analyzing the family with different aggregation technique/s was unfeasible—and in part because research articles missed some relevant information to judge the adequacy of the aggregation techniques used (e.g., were the raw data of all the experiments available at the time of the aggregation of results? Were the response variables' operationalizations identical?). In view of this, we provide a series of recommendations for *reporting* SE families:

Advice for reporting families

- *Acknowledge whether the raw data of all the experiments are available* at the time of the aggregation of results.
- *Acknowledge whether the same response variables' operationalizations* are used across all the experiments.
- *Report the sample sizes of all the experiments* to ensure that effect sizes can be calculated and, thus, experiments' results can be incorporated in prospective meta-analyses.
- *Report the relationship between the participants of the different experiments*. If they are the same participants, correlations may have been introduced across experiments' results, and this may invalidate joint conclusions if not correctly taken into account.
- *Acknowledge all the changes made across the experiments* so that it is possible evaluating whether other aggregation technique/s could have been applied instead.
- *Provide the raw data if possible* to facilitate the reproduction of results and re-analysis with potentially more appropriate aggregation techniques.

3.3 Finding 3: Assessment of the techniques in the SE context

After learning about the aggregation techniques used to analyze SE families, and the advantages and disadvantages of each technique according to mature experimental disciplines, we were interested in assessing whether the same advantages and disadvantages would materialize in the SE context. To evaluate this, we applied the aggregation techniques on a representative SE family and compared the findings obtained with each technique. We describe the family and compare the findings, advantages, and disadvantages of each technique when applied on such family in Section 3.3.1. Finally, in view of the results obtained with each aggregation technique, we provide a short list of recommendations to *analyze* SE families in Section 3.3.2.

3.3.1 Assessment of the techniques on a representative SE family

We selected a family of experiments on TDD comprised of *four experiments* (i.e., the median number of experiments in SE families), with *small* and *dissimilar sample sizes* and *heterogeneous results* (i.e., *small* and *large* effect sizes), *identical response variable operationalizations* (i.e., response variable in a percentage scale) and *experimental*

designs (i.e., an AB within-subjects design (Wohlin et al., 2012)), *missing data* (i.e., four subjects in one experiment), and *different types of subjects* (i.e., professionals in three experiments—those in F-Secure—and students in one experiment—those in UPV).

All the experiments were run by the same group of researchers. They kept the experiments' designs identical to increase the internal validity of the results (Whitehead, 2002). The experiments assessed the difference in performance between TDD and ITL in terms of functional correctness (FC). Functional correctness is a sub-characteristic of quality according to ISO 25010. It is defined as *the degree to which a system provides the correct results with the needed degree of precision* (*ISO/IEC 25010:2011*, 2011). The experimenters measured FC as the percentage of test cases that passed from a battery of test cases that the experimenters built to test the participants' solutions. Specifically, they measured functional correctness as

$$FC = \frac{\#Test\ Cases(Pass)}{\#Test\ Cases(All)} * 100.$$

A total of six, 11, seven and 33 participants took part in the experiments at F-Secure H, F-Secure K, F-Secure O, and UPV, respectively. The participants were handed a questionnaire with a series of ordinal scale (i.e., inexperienced, novice, intermediate, expert) self-assessment questions asking them to rate their experience with programming, Java, unit testing, and JUnit. I was granted access to the data for being a member of the ESEIL project.

Table 7 shows the mean experiences of the participants with programming, Java, unit testing, and JUnit across the experiments (i.e., 1-4, for inexperienced, novice, intermediate and experts, respectively).[14]

Table 7. Mean experiences across replications.

Experiment	N	Programming	Java	Unit	JUnit
F-Secure H	6	3.67	2.33	2.17	2.17
F-Secure K	11	2.91	1.82	1.64	1.27
F-Secure O	7	3.29	2.71	2.71	2
UPV	33	2.36	1.88	1.04	1

As we can see in Table 7, the most senior developers are those at F-Secure O and F-Secure H, while those at UPV are the least experienced in general. As a summary, *the family of experiments is comprised of an heterogeneous group of developers*.

[14]For simplicity's sake, we consider the ordinal variables here as continuous. This approach is commonly followed in other disciplines (Norman, 2010).

Table 8 shows the descriptive statistics (i.e., sample sizes, means, standard deviations, and medians) for ITL and TDD's FC scores across the experiments.

Table 8. Descriptive statistics: ITL vs. TDD. **Reprinted by permission [Paper III] © ACM 2018.**

Experiment	Treat.	N	Mean	SD	Median
F-Secure H	ITL	6	30.71	36.58	24.16
	TDD	6	40.23	33.43	35.34
F-Secure K	ITL	11	22.17	20.44	17.98
	TDD	11	35.42	35.40	22.41
F-Secure O	ITL	7	16.05	20.81	7.87
	TDD	7	68.97	31.53	81.03
UPV	ITL	31	33.38	39.79	6.74
	TDD	29	77.16	21.04	83.93

Figure 2 shows the profile plot of the experiments' ITL and TDD mean FC scores.

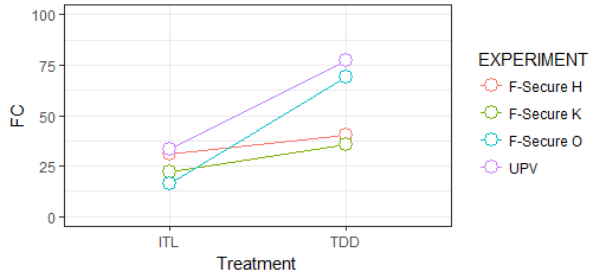

Fig. 2. Profile plot: ITL vs. TDD. **Reprinted by permission [Paper I] © ACM 2018.**

As we can see in Figure 2, TDD's mean FC scores are higher than ITL's in all the experiments. In other words, TDD outperforms ITL in all the experiments. Besides, the difference in performance between TDD and ITL looks different across the experiments (notice that the slopes of the lines are differently steep). This may point towards the presence of heterogeneity of results. As a last observation, we can notice that the students obtained greater mean FC scores than the professionals for TDD and ITL. We put this to the greater familiarity of the students with small toy programming tasks (such as those they code throughout their degrees). Also the students' greater commitment with the experiment may have led them to obtain better results than the professionals.

In the following, we go over the aggregation techniques used to analyze SE families: narrative synthesis, AD, IPD-S, and aggregation of p-values. We skip IPD-MT because of length restrictions, and because it has been disregarded in mature

experimental disciplines (Abo-Zaid et al., 2013; Kraemer, 2000). With each technique, we provide **joint conclusions** and elicit **experiment level moderators** (i.e., students vs. professionals) and **participant level moderators** (i.e., experience with programming, Java, unit testing, and JUnit), if possible.

Narrative synthesis

To apply narrative synthesis to provide **joint conclusions**, we just provide *a textual summary of the results of the individual experiments* (Borenstein et al., 2011).

Because all the experiments within the family are AB within-subjects experiments, we analyze each experiment with a *dependent t-test* (Field, 2013). Then, we provide a textual description of the results to summarize them. Table 9 shows the results of the dependent *t*-tests.

Table 9. Individual analyses: ITL vs. TDD.

Experiment	Estimate	95% CI	*p*-value
F-Secure H	9.52	(-19.58, 38.62)	0.483
F-Secure K	13.26	(-7.26, 33.77)	0.193
F-Secure O	52.91	(30.44, 75.39)	**<0.001**
UPV	42.31	(29.02, 55.62)	**<0.001**

A summary of results with narrative synthesis may go something like this: TDD outperformed ITL in all the experiments, even though to a different extent. In addition, because the difference in performance between TDD and ITL was statistically significant in two out of the four experiments, there is no conclusive evidence of the statistical significance of the results (because an identical number of experiments point in opposite directions—i.e., significant vs. non-significant). Summarizing this, *more experiments are still needed to draw definite conclusions on the magnitude, and statistical significance of the results using narrative synthesis*.

To identify **experiment level or participant level moderators** with narrative synthesis, we should look for patterns in the results (e.g., do professionals obtain greater benefits with TDD than students? Do more experienced subjects obtain more benefits than less experienced subjects? (Borenstein et al., 2011)).

Unfortunately, we cannot see clear patterns in the results: the largest benefits with TDD are obtained in one out of the three experiments with professionals, and in one experiment with students. The experiments where TDD looks more beneficial are also those with either the most senior or the most junior developers on average (i.e., F-Secure

O, and UPV, respectively). This suggests that either more experiments are needed to identify clear patterns in the results or that TDD only works for very experienced and inexperienced subjects—and not for averaged experienced subjects. In sum, *more experiments are still needed to draw definite conclusions on the presence of moderators with narrative synthesis.*

Aggregated data (AD)

To apply AD to provide **joint conclusions**, we first must calculate the experiments' effect sizes—and their corresponding variances—from the experiments' summary statistics (e.g., means, standard deviations, etc. (Borenstein et al., 2011)). Then, pool the effect sizes together by means of a meta-analysis model (Borenstein et al., 2011).

We first calculated each experiment's Hedges' g (i.e., the small samples correction for Cohen's d) following Borenstein et al. (Borenstein et al., 2011). Because four subjects at UPV had missing data, such subjects had to be removed to calculate Hedges' g. Then, we pooled the Hedges' g and their respective variances by means of a random-effects meta-analysis model (Borenstein et al., 2011). We plotted the results of such meta-analysis in a forest plot (see Figure 3).

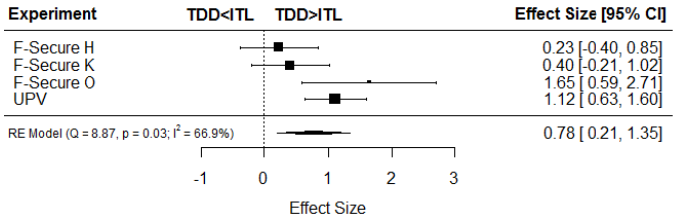

Fig. 3. Forest plot: ITL vs. TDD. **Reprinted by permission [Paper III] © ACM 2018.**

As we can see in Figure 3, a joint effect size equal to $M = 0.78$ materialized. The joint effect size favors TDD and is *almost large*—according to rules of thumb (Borenstein et al., 2011)—and *statistically significant* (because its 95% CI does not cross 0). Additionally, an observable *heterogeneity of results* materialized ($I^2 = 66.9\%$). This heterogeneity of results could be due to the different types of subjects participating across the experiments (i.e., students vs. professionals) or by the different characteristics of the participants across the experiments (i.e., the participants' different experiences with programming, Java, unit testing, and JUnit).

To assess the extent to which the **experiment level moderator** (i.e., students vs. professionals) affects the results, we ran *a sub-group meta-analysis* (Borenstein et al., 2011). A sub-group meta-analysis can be simply thought of an independent meta-

analysis per each sub-group followed by a Z-test that compares the joint conclusions achieved in both sub-groups (Borenstein et al., 2011).

Table 10 shows the results of the sub-group meta-analysis that we ran.

Table 10. Sub-group meta-analysis: experiment level moderator. **Reprinted by permission [Paper III] © ACM 2018.**

Variable	Group	N	Estimate	95% CI	I^2
Subject	Professionals	3	0.65	(-0.10, 1.41)	68.07%
	Students	1	1.12	(0.63, 1.60)	0
	Difference	-	0.47	(-0.44, 1.36)	-

As we can see in Table 10, students benefit more than professionals from using TDD ($M = 1.12$ and $M = 0.65$ for students and professionals, respectively). Even though the difference in effectiveness between students and professionals is noticeable ($M = 0.47$), such difference is not statistically significant (i.e., the 95% CI of the difference crosses 0). As a summary, and in view of the fact that relevant differences across sub-groups are not statistically significant, we conclude that *the family seems under-powered for detecting experiment level moderators with AD.*

To study the extent to which **participant level moderators** (i.e., experience of the participants with programming, Java, unit testing, and JUnit) affect the results, all we need to do is average out the experience of the participants across the experiments (see Table 7), and then, running a meta-regression where the averaged characteristics of the participants are regressed towards the experiments' effect sizes (Cooper & Patall, 2009).

Table 11 shows the results of the meta-regression that we ran.

Table 11. Meta-regression: participant level moderators. **Reprinted by permission [Paper III] © ACM 2018.**

Interaction	Estimate	95% CI	p-value
Programming	-0.4	(-1.68, 0.84)	0.51
Java	0.61	(-1.38, 2.60)	0.55
Unit testing	0.12	(-1.08, 1.32)	0.84
JUnit	-0.16	(-1.60, 1.28)	0.82

As shown in Table 11, while Java and unit testing averaged experience leads to more benefits with TDD, the opposite happens with programming and JUnit. Despite this, large 95% CIs materialized for all the moderators; thus the estimates are not precise. None of the moderator effects are statistically significant either. In sum, *the family seems under-powered for detecting participant level moderators with AD.*

Individual participant data (IPD)-S

To **provide joint conclusions** with IPD-S, all we need to do is pooling the raw-data of all the experiments together, and then analyze them by using a statistical model that allows the inclusion of an extra "experiment" factor when analyzing the data (e.g., a LMM with "treatment" and "experiment" as factors (Abo-Zaid et al., 2013)).

We analyzed the family of experiments with a LMM (Brown & Prescott, 2014). We included in the analysis the participants with missing data—because LMMs can handle missing data as long as they are considered missing at random (i.e., that the missingness pattern is not due to the FC scores achieved by the participants (Hoffman & Rovine, 2007)). Table 12 shows the results of the LMM that we fitted.

Table 12. LMM family of experiments. **Reprinted by permission [Paper III] © ACM 2018.**

Factor	Estimate	95% CI	p-value
ITL	27.44	(13.08, 41.79)	<0.001
TDD	56.27	(22.12, 90.42)	<0.001
M_{Diff}	28.83	(9.72, 47.93)	0.004
sd_{Diff}	16.09		

As we can see in Table 12, *the difference in effectiveness between TDD and ITL is statistically significant* (p-value = 0.004). Besides, *TDD's mean FC scores* ($M = 56.27$) *double those of ITL* ($M = 27.44$). Unfortunately, the interpretation of heterogeneity is not as straightforward as with AD—-because no I^2 statistic is provided by the LMM. However, we can see that the standard deviation of the differences between TDD and ITL's effectiveness across the experiments ($sd_{Diff} = 16.09$) is relatively large compared with the overall difference in effectiveness between TDD and ITL ($M_{Diff} = 28.83$). This seems to point to a considerable dispersion (i.e., heterogeneity) of results across the experiments (Brown & Prescott, 2014). Thus, the characteristics of the experiments, or the participants, may be influencing the experiments' results.

To assess **experiment level moderators** with IPD-S, we need to fit statistical models (e.g., LMM) with interaction terms (Brown & Prescott, 2014). Table 13 shows the results obtained for the interaction term type of subject (i.e., professionals vs. students) in our family.

As shown in Table 13, students obtain more benefits than professionals when using TDD ($M = 16.32$). As in AD, the interaction seems large—at least considering that the treatment effect was equal to $M_{Diff} = 28.83$, and thus, the interaction term corresponds to almost 60% of the treatment effect. However, the interaction is not

statistically significant. Thus, *the family seems under-powered to identify experiment level moderators with IPD-S.*

Table 13. LMM results: experiment level moderators. **Reprinted by permission [Paper III] © ACM 2018.**

Interaction	Estimate	95% CI	*p*-value
Subject:Students	16.32	(-37.16, 69.55)	0.545

Finally, to assess **participant level moderators** with IPD-S it suffices with fitting statistical models with interaction terms, but this time, separating the variability of the moderator effects into two different terms: the intra experiment variability, and the inter experiment variability (Fisher et al., 2011). The intra experiment variability of the moderator effects should be evaluated for assessing the sign and magnitude of the participant level moderators (Fisher et al., 2011).

We assessed the effect of the participants' experiences on results with four different LMMs (one per experience variable). The parameter estimates of the interaction terms are presented in Table 14.

Table 14. LMM: participant level moderators. **Reprinted by permission [Paper III] © ACM 2018.**

Interaction	Estimate	95% CI	*p*-value
Programming	15.76	(0.49, 31.04)	0.04
Java	3.85	(-8.13, 15.83)	0.52
Unit testing	11.79	(-5.95, 29.54)	0.18
JUnit	11.07	(-6.26, 28.41)	0.20

Figure 4 shows the regression lines corresponding to the interaction terms.

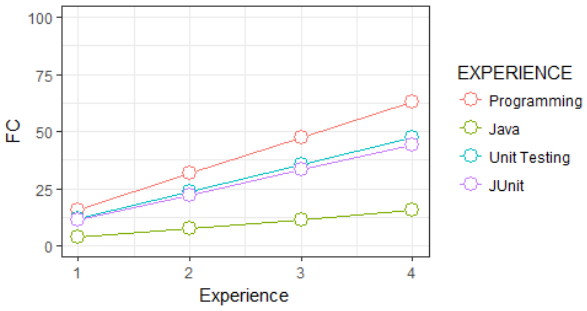

Fig. 4. LMM interactions: participant level moderators. **Reprinted by permission [Paper III] © ACM 2018.**

As we can see in Table 14, the experience of the participants with programming seems to influence TDD's effectiveness to a noticeable ($M = 15.76$) and statistically significant extent (p-value=0.04). In addition, as shown in Figure 4, all the experience variables have a positive impact on TDD's effectiveness (because all the lines have an upward trend). This is contrary to what happened with AD with the averaged experience of the participants with programming and JUnit. This ocurred in AD because the distribution of effect sizes and averaged participant level experiences were reversed across the experiments. This phenomenon, also known as ecological bias (Berlin, Santanna, Schmid, Szczech, & Feldman, 2002), can materialize with AD when the presence of an averaged effect is not representative of what happens in the participants having this type of averaged characteristic. Summing up, *the family seemed powered enough for detecting noticeable participant level moderation effects with IPD-S*.

Aggregation of p-values

To provide **joint conclusions** with aggregation of p-values, each experiment needs to be analyzed with a one-sided statistical test (e.g., a one-sided t-test), and then, the resulting p-values combined by means of an aggregation of p-values technique (e.g., Fisher's method (Borenstein et al., 2011)).

We analyzed each of the experiments within the family by means of a one-sided dependent t-test, and then pooled their p-values by means of the Fisher's method. A statistically significant difference between TDD and ITL was obtained ($\tilde{\chi}^2$=47.13; df=8; p<0.001). Thus, by means of aggregation of p-values, we conclude that in at *least one experiment*, the difference in effectiveness between TDD and ITL is statistically significant. However, this was already known before undertaking the aggregation of results. Besides, no joint effect size can be provided, which hinders the interpretation of results.

Table 15 summarizes the advantages and disadvantages that materialized with each aggregation technique in the family.

3.3.2 Preliminary advice for analyzing families of experiments

In view of the results obtained with each aggregation technique, we provide a series of recommendations to *analyze* SE families:

Advice for analyzing families

- *AD and IPD-S* seem more suitable than narrative synthesis and aggregation of p-values to analyze SE families.
- *AD and IPD-S* seem complementary for providing joint conclusions and assessing experiment level moderators.
- *There is no need to use standardized effect sizes if the experiments have identical response variable operationalizations*. Instead, AD with unstandardized effect sizes or IPD-S can be used to increase the interpretability of results.
- *IPD-S* seems more suitable than AD for identifying participant level moderators.

3.4 Finding 4: Procedure to analyze SE families

After evaluating the suitability of the aggregation techniques used for analyzing SE families, we wanted to provide guidance to facilitate the analysis of SE families. For this, we looked at the guidelines typically followed in medicine and pharmacology to

Table 15. Analysis techniques' advantages and disadvantages in our family of experiments.

Advantages	Technique	Disadvantages
✓ Intuitive approach	**Narrative synthesis**	✗ No effect size nor p-value ✗ Subjective weighting ✗ No moderators without patterns
✓ Intuitive visualizations ✓ Easy interpretation of heterogeneity	**AD**	✗ Unable to handle missing-data ✗ Interpretation standardized units ✗ Ecological bias
✓ Identifies participant level moderators ✓ Accommodates missing data ✓ Interpretation natural units	**IPD-S**	✗ Complex interpretation of heterogeneity
	Aggregation of p-values	✗ No effect size ✗ Potentially uninformative joint conclusions

analyze—and report—MCTs. We identified some differences between SE families and MCTs. Such differences have a series of statistical implications that make the application of MCTs guidelines unsuitable for analyzing SE families (Section 3.4.1). In view of such differences and their statistical implications, we propose a procedure with a set of embedded guidelines to analyze SE families (Section 3.4.2). Along the way, we also illustrate the use of the proposed procedure by applying it to analyze a representative SE family.

3.4.1 Differences between MCTs and SE families and statistical consequences

After reading the guidelines typically followed in medicine and pharmacology for analyzing and reporting MCTs (Anello et al., 2005; Bero & Rennie, 1995; Lewis, 1999; Schulz et al., 2010; Stewart et al., 2015), we noticed a series of differences between MCTs and SE families.

For example, MCTs tend to specify in detailed protocols the experimental settings under which all experiments are to be run and the set of procedures that shall be strictly adhered to during the execution of the experiments (Bero & Rennie, 1995; Lewis, 1999; Whitehead, 2002). Not to mention, in MCTs aiming at assessing the efficacy of new drugs, specific requirements need to be met by all the participants in the experiments (e.g., certain blood parameters, lack of "co-morbid" conditions, etc. (Bero & Rennie, 1995; Lewis, 1999)). All these measures ensure consistency of results across experiments and thus, aid to minimize the risk of heterogeneity. This facilitates the application of fixed-effects models—the default statistical models used in medicine (Whitehead, 2002)—to analyze the data.

On the contrary, SE families are usually built "ad-hoc." That is, experiments with different configurations are typically run (e.g., experiments with different types of subjects, session lengths, or experimental tasks), and then, their results are opportunistically aggregated. Such heterogeneity of participants and configurations increases the chance of heterogeneous results, thus making unsuitable fixed-effects models to provide joint conclusions (Borenstein et al., 2011; Whitehead, 2002).

Besides, MCTs commonly undertake a planning phase where sample sizes at both the participant level (how many subjects are needed?) and at the experiment level (how many centers are needed if it is only plausible allocating X subjects to each experiment?) are calculated (Anello et al., 2005; Bero & Rennie, 1995). This ensures enough statistical power for detecting true population effect sizes. Such calculation also favours the presence of experiments with similar (i.e., balanced) sample sizes. Again,

this makes the application of fixed-effects models tenable (Chu et al., 2011; Localio, Berlin, Ten Have, & Kimmel, 2001).

On the contrary, SE families rarely undertake any sample size calculations and instead recruit participants in an opportunistic manner (i.e., use convenience sampling). Besides, a low number of experiments are typically run (Santos, Gómez, & Juristo, 2018). This makes the presence of a few, small, and unbalanced sample sizes a common issue in SE families. This translates into a hindered ability to detect moderators (Kraemer, 2000). The sample size being unbalanced across experiments makes fixed-effects models under-powered to detect true population effects as well (Chu et al., 2011; Localio et al., 2001).

Table 16 summarizes the differences between MCTs and SE families and the statistical consequences of such differences.

Table 16. Differences between MCTs and SE families and statistical consequences. **Reprinted by permission [Paper IV] © IEEE 2019.**

MCTs	SE families	Consequence
✓ Identical configurations	✗ Opportunistic changes	- Heterogeneity
✓ Rigid selection criteria	✗ Convenience sampling	- Heterogeneity
✓ Balanced adequate sample sizes	✗ Unbalanced small sample sizes	- Fixed effects low precision and power
✓ Adequate overall sample size	✗ Small overall sample size	- Inability to detect moderators

3.4.2 Guidelines to analyze SE families and illustrative example

We propose to follow a four-step procedure to analyze SE families. Each step serves its own purpose:

- **Step 1: Describe participants**. *Objectives.* Informing about the population under assessment and uncovering possible sources of heterogeneity at the participant level.
- **Step 2. Analyze individual experiments**. *Objectives.* Easing the incorporation of results in prospective studies, identifying patterns across experiments' results, and avoiding heterogeneity of results due to different analysis procedures.
- **Step 3. Aggregate results**. *Objectives.* Maximizing the informativeness of joint conclusions.
- **Step 4. Conduct exploratory analyses**. *Objectives.* Identifying moderators at the experiment level (i.e., characteristics of the experiments that may be impacting the

results) and at the participant level (i.e.., characteristics of the participants that may be affecting the results).

In the following, we go over the steps of the analysis procedure, and apply each to analyze a representative SE family (i.e., the family described in Section 3.3.1).[15]

Step 1: Describe participants

First, we recommend providing *descriptive statistics* and *visualizations* (i.e., profile plots) of the participants' characteristics. This has two main objectives: (1) describing the population to which inferences are aimed at and (2) revealing potential sources of heterogeneity at the participant level.

Example. In the family of experiments on TDD that we analyze, we collected the experience of the participants with programming, Java, unit testing, and JUnit. Table 17 shows the average and standard deviation of the participants' experiences with programming, Java, unit testing and JUnit across the experiments. Figure 5 complements the descriptive statistics by plotting the average experience of the participants across the experiments.

Table 17. Descriptive statistics for participants' characteristics. **Reprinted by permission [Paper IV] © IEEE 2019.**

Experiment	Prog.	Java	Unit	JUnit
F-Secure H	3.67 (0.52)	2.33 (1.21)	2.17 (0.98)	2.17 (1.17)
F-Secure K	2.91 (0.70)	1.82 (0.87)	1.64 (0.5)	1.27 (0.47)
F-Secure O	3.29 (0.76)	2.71 (1.11)	2.71 (0.76)	2 (0.82)
UPV	2.36 (0.57)	1.88 (0.60)	1.04 (0.20)	1 (0)

As we can see in Table 17 and Figure 5, F-Secure H and F-Secure O's participants are the most experienced, while those of UPV are the least experienced in general. F-Secure K's participants stay in the middle. As a summary, a heterogeneous population of developers participates in the experiments. This heterogeneity of participants may lead to heterogeneous results in later stages of the analysis.

[15] Because we analyze the same family of experiments that we analyzed in Section 3.3, some figures and tables are repeated. We chose this approach for illustrative purposes and to follow the same structure as in Article IV.

Step 2: Analyze individual experiments

As a second step, we recommend providing *descriptive statistics* and *data visualizations* (i.e., box plots and profile plots) and then *analyzing the experiments* with consistent statistical models. This step has three main objectives: (1) easing the incorporation of results in prospective studies; (2) identifying patterns across experiments' results; and (3) avoiding heterogeneity of results due to the use of different analysis procedures.

Example. In the family of experiments, we compared TDD's and ITL's effectiveness in terms of QLTY (i.e., quality on a 0-100 scale). Table 18 shows the descriptive statistics for the QLTY scores reached with ITL and TDD across the experiments.

Table 18. Descriptive statistics for QLTY: ITL vs. TDD. **Reprinted by permission [Paper IV] © IEEE 2019.**

Experiment	Treatment	N	Mean	Corr	SD	Median
F-Secure H	ITL	6	30.71	0.59	36.58	24.16
	TDD	6	40.23		33.43	35.34
F-Secure K	ITL	11	22.17	0.42	20.44	17.98
	TDD	11	35.42		35.40	22.41
F-Secure O	ITL	7	16.05	0.52	20.81	7.87
	TDD	7	68.97		31.53	81.03
UPV	ITL	31	33.38	0.47	39.79	6.74
	TDD	29	77.16		21.04	83.93

Figure 6 complements Table 18, showing the distribution of the QLTY scores across the experiments.

Finally, the profile plot depicted in Figure 7 provides an overview of the average QLTY scores achieved across the experiments with ITL and TDD.

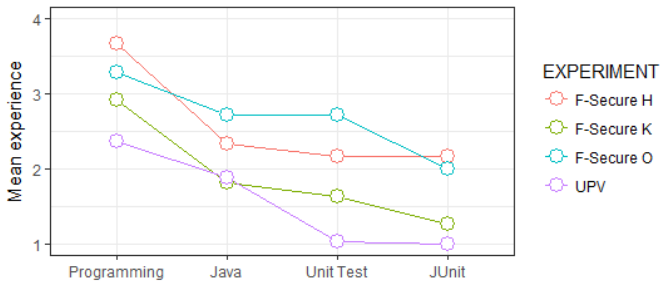

Fig. 5. profile plot for the participants' characteristics. **Reprinted by permission [Paper IV] © IEEE 2019.**

As shown in the descriptive statistics and plots, ITL's QLTY averaged scores are lower than those reached with TDD in all the experiments. Also, as shown in Figure 7, TDD tends to outperform ITL in all the experiments to a different extent (see that the lines are differently steep; the steeper lines indicating the greater the difference between TDD and ITL's performance). This may point towards the presence of heterogeneous results.

Finally, as all the experiments have an identical experimental design (i.e., an AB within-subjects design), we analyze all the replications with an identical statistical test: a dependent t-test (Field, 2013). The results of the dependent t-tests are shown in Table 19.

Table 19. Individual analyses: ITL vs TDD. Reprinted by permission [Paper IV] © IEEE 2019.

Experiment	Estimate	95% CI	p-value
F-Secure H	9.52	(-19.58, 38.62)	0.483
F-Secure K	13.26	(-7.26, 33.77)	0.193
F-Secure O	52.91	(30.44, 75.39)	<**0.001**
UPV	42.31	(29.02, 55.62)	<**0.001**

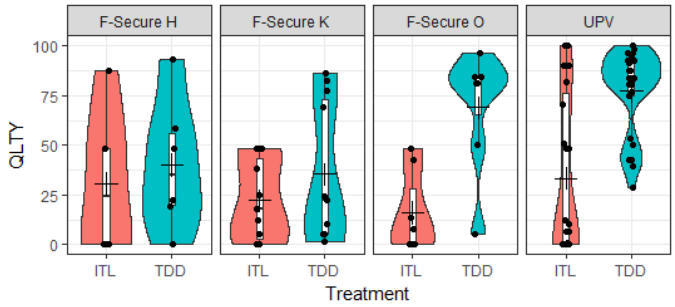

Fig. 6. Box plot and violin plot: ITL vs. TDD. Reprinted by permission [Paper IV] © IEEE 2019.

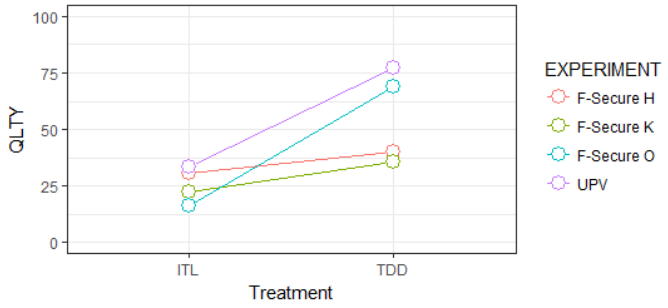

Fig. 7. Profile plot: ITL vs. TDD. Reprinted by permission [Paper IV] © IEEE 2019.

As shown in Table 19, TDD outperforms ITL to a different extent across the replications, being the outperformance statistically significant in F-Secure O and UPV but not in F-Secure H and F-Secure K.

Step 3: Aggregate results

Step 3 involves the aggregation of the experiments' results. For this, suitable aggregation techniques should be selected. The main objective of Step 3 is maximizing the informativeness of joint conclusions. We break Step 3 down into three different sub-steps: (1) avoid narrative synthesis and aggregation of p-values; (2) avoid IPD-MT; (3) use AD and IPD-S in tandem. In the following, we go over the sub-steps.[16]

Step 3.1. Avoid Narrative synthesis and Aggregation of p-values

According to statistical reformers (Cumming, 2013) and associations (Wasserstein & Lazar, 2016), the results should be assessed at least under two different points of view: the statistical significance and the practical significance points of view. The statistical significance point of view is typically judged in terms of *p*-values (Cumming, 2013). The practical significance point of view is typically judged in terms of effect sizes and their respective 95% CIs.

Having this in mind, neither narrative synthesis nor aggregation of *p*-values are suitable for providing joint conclusions. This is because narrative synthesis can provide neither a joint *p*-value nor effect size (but instead a textual description of results), and aggregation of *p*-values only produces a joint *p*-value (and thus hinders the assessment of the practical significance of results).

Besides, narrative synthesis and aggregation of *p*-values have other limitations. For example, narrative synthesis assigns a subjective weight to each replication towards the joint conclusion. This may be undesirable in terms of reproducibility of results (as different synthesists may assign different weights to each replication). Aggregation of *p*-values techniques (e.g., Fisher's or Stouffer's methods) assign an identical weight to each replication (Borenstein et al., 2011). This may be undesirable if replications have different sample sizes (as larger replications may provide more precise results than smaller replications (Cumming, 2013)).

[16] Albeit we do not encourage using narrative synthesis, aggregation of *p*-values and IPD-MT to analyze families, in the following sections, we apply them for illustrative purposes. In this way we follow the same structure followed in Paper IV.

We illustrate the use of narrative synthesis and aggregation of *p*-values in the family on TDD.

Example. A summary of results with *narrative synthesis* may go something like this: "TDD outperforms ITL across all the replications. However, to a different extent: while TDD outperforms ITL to a large extent in F-Secure O and UPV, the extent of such outperformance is smaller in F-Secure H and F-Secure K. Besides, the difference in performance between TDD and ITL is statistically significant in only two out of four experiments. Thus, because an identical number of experiments point in opposite directions in terms of statistical significance (significant vs. non-significant), we cannot draw definite conclusions on the statistical significance of results."

A summary of results with *aggregation of p-values* may go something like this: "According to Fisher's method, the joint one sided *p*-value is statistically significant. Thus, according to aggregation of *p*-values, in at least one experiment, the difference in performance between TDD and ITL is statistically significant. However, this was already known before the aggregation of results (i.e., at F-Secure O and UPV)."

As a take-away:

Guideline 1. Avoid narrative synthesis and aggregation of *p*-values when providing joint conclusions.

Step 3.2. Avoid IPD-MT

IPD-MT has been discouraged in mature experimental disciplines due to its potential to provide unreliable and/or under-powered results (Abo-Zaid et al., 2013; Kraemer, 2000). IPD-MT tends to be more misleading the larger the sample size unbalance across groups within experiments and the larger the dissimilarities among participants' scores across experiments (Abo-Zaid et al., 2013; Kraemer, 2000).

However, IPD-MT may provide reliable results when the data are perfectly balanced within experiments (e.g., as in repeated-measures designs (Wohlin et al., 2012)) and when the variability of the data are similar across them. However, because this cannot be guaranteed in SE families (as SE families may contain missing data, experiments may have different sample sizes, and heterogeneous participants may obtain heterogeneous scores (Santos et al., 2018)), we propose avoiding IPD-MT by default and using IPD-S instead. Similar advice has already been given in medicine (Abo-Zaid et al., 2013; Feaster, Mikulich-Gilbertson, & Brincks, 2011; Kraemer, 2000).

Example. According to *IPD-MT*, the difference in performance between TDD and ITL in the family on TDD is equal to $M = 33.67$ with a p-value<0.001. Thus, the difference in performance between TDD and ITL is large—at least when looking at the fact that QLTY spans between 0 and 100—and statistically significant. Applying IPD-MT seems safe because: (1) all the experiments are perfectly balanced (because they are AB within-subjects experiments); and (2) because the variability of the data is similar across the replications. However, this cannot be guaranteed in other families.

As a take-away:

Guideline 2. Avoid IPD-MT for analyzing SE families due to its potential to provide unreliable and/or under-powered results.

Step 3.3. Use AD and IPD-S in tandem

In general terms, both AD and IPD-S can be seen as procedures that deliver a weighted average of the results of the experiments as a joint conclusion (Borenstein et al., 2011; Cooper & Patall, 2009; Whitehead, 2002). The weight assigned to each experiment is either proportional to the experiments' sample sizes—if a fixed effects model is fitted—or to the experiments' sample sizes and the heterogeneity of results—if a random-effects model is fitted.

Because IPD-S and AD tend to provide similar results (Smith et al., 2016), we propose looking at AD and IPD-S as complementary strategies to analyze SE families. This is because each has its own advantages. Among the advantages of AD over IPD-S, we find the following:

- *AD can aggregate the results of experiments with different response variable scales*. For example, this can be done using standardized effect sizes such as Cohen's d (Borenstein et al., 2011). This is unfeasible with IPD-S because it requires identical response variable scales across the experiments (Whitehead, 2002).
- *AD provides intuitive visual summaries of results* such as forest plots that allow visualizing individual and joint results at a glance (Borenstein et al., 2011). On the contrary, less standardized visualizations are available for IPD-S (such as error bars, 95% CI plots, etc. (Cumming, 2013)). Thus, the intuitiveness of forest plots for representing the results is a plus for AD over IPD-S.
- *AD allows for interpreting the heterogeneity of results with straightforward tests and statistics* (e.g., the Q-test and the I^2 statistic (Borenstein et al., 2011)). On the contrary, the interpretation of heterogeneity with IPD-S is less intuitive. It either involves

contextualizing the joint result with the variability of results across experiments, or fitting a statistical model with an interaction term between the treatment and the experiments, and judging the statistical significance of the interaction (Whitehead, 2002).

Among the advantages of IPD-S over AD, we find the following:

- IPD-S allows for interpreting results in natural units and comparing the effectiveness of the treatments in relative terms. For example, if the joint effectiveness for Treatment A equals 40 and for Treatment B equals 80, Treatment B doubles Treatment A's effectiveness (Whitehead, 2002). On the contrary, AD's results are typically conveyed in standardized units (e.g., in terms of Cohen's d). This may hinder the interpretability of results (e.g., how relevant is a medium Cohen's d of say, 0.3? (Borenstein et al., 2011)).
- *IPD-S allows for interpreting the effect of multiple factors and their interaction on the results*. For example, interpreting the effect of the task, the treatment, and their interaction on the results. On the contrary, AD is typically used to perform *pairwise comparisons* between treatments (e.g., Treatment A vs. Treatment B). Thus, IPD-S provides greater flexibility than AD for analyzing SE families where the results depend on multiple factors (e.g., when the effects of the treatments reverse depending on the task).
- *Some IPD-S models such as LMMs allow for analyzing families of experiments with missing data*—assuming data missing at random (Twisk, de Boer, de Vente, & Heymans, 2013). On the contrary, AD rests on the assumption of complete data (because otherwise the variance of some repeated measures effect sizes cannot be computed (Borenstein et al., 2011)) or the use of imputation techniques for handling missing data (Schafer & Graham, 2002).

As a final point, in view of the fact that heterogeneous results are common in SE experiments (Sjoberg et al., 2007), that multiple factors may impact SE experiments' results (Basili et al., 1999), and that SE families typically contain experiments with multiple changes and heterogeneous participants (Santos et al., 2018), we recommend relying by default on random-effects models. Specifically, we recommend relying on LMMs with IPD-S, and on random-effects meta-analysis models with AD. Similar advice has already been provided in other disciplines in similar circumstances (Borenstein et al., 2011; Whitehead, 2002).

Example. To analyze the family on TDD with *AD*, we first compute Hedges' g (i.e., the small sample size correction for Cohen's d (Borenstein et al., 2011)) from each experiment's descriptive statistics. As UPV has four subjects with missing data,

we remove their data from the computations. Then, we pool the effect sizes of all the experiments together with a random-effects model. We finally provide a forest plot with the results of the meta-analysis. Figure 8 shows such forest plot.

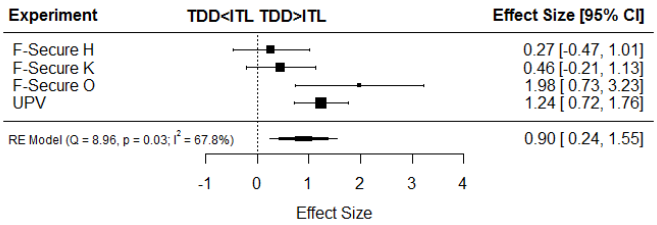

Fig. 8. Forest plot: ITL vs. TDD. **Reprinted by permission [Paper IV] © IEEE 2019.**

As we can see in Figure 8, TDD outperforms ITL to a *large* extent—according to rules of thumb (Borenstein et al., 2011). Additionally, a *medium* heterogeneity of results is present—according to rules of thumb (Borenstein et al., 2011). This may point towards the presence of moderators in the family of experiments.

To analyze the family with *IPD-S*, we run a LMM with "treatment" as a random effect and "experiment" as a random factor (Brown & Prescott, 2014). Table 20 shows the results of the LMM that we fitted.

Table 20. LMM results. **Reprinted by permission [Paper IV] © IEEE 2019.**

Factor	Estimate	95% CI	p-value
ITL	27.44	(13.08, 41.79)	$<$0.001
TDD	56.27	(22.12, 90.42)	$<$0.001
M_{Diff}	28.83	(9.72, 47.93)	0.004
sd_{Diff}	16.09		

As shown in Table 20, TDD's mean effectiveness ($M = 56.27$) doubles that of ITL ($M = 27.44$). In other words, participants applying TDD are expected to increase their performance up to a 100% over those applying ITL. Finally, the variance of results seems considerable compared to the joint result (i.e., 28.83/16.09). Again, this may point toward the presence of moderators.

As a take-away:

Guideline 3. Use AD and IPD-S in tandem to provide joint conclusions. Use AD because of its ability to combine experiments with different scales, its intuitive visualizations, and its straightforward heterogeneity statistics. Use IPD-S because of its ability to convey results in natural units, its greater statistical flexibility, and its ability to handle missing data.

Step 4: Conduct exploratory analyses

Finally, once the aggregation of the results is done, we recommend conducting exploratory analyses. Exploratory analyses are useful for identifying variables—either at the experiment level, or at the participant level—that may be impacting the results of the experiments (i.e., moderator variables, or simply moderators). We break down Step 4 into 3 sub-steps: (1) identify experiment level moderators with AD and IPD-S; (2) identify participant level moderators with IPD-S; and (3) acknowledge exploratory analyses' limitations.

Step 4.1. Identify experiment level moderators with AD and IPD-S.

Experiment level moderators are variables at the experiment level that may be impacting the results of the experiments. For example, the IDE or the type of subjects participating in the experiments (e.g., professionals vs. students) are potential experiment level moderators.

AD and IPD-S provide similar results when eliciting experiment level moderators (Simmonds & Higgins, 2007). Thus, we recommend using both. This way, the advantages of both techniques will materialize and more informative conclusions can potentially be drawn.

To elicit experiment level moderators with AD it suffices with running a *meta-regression* (Borenstein et al., 2011)—if the moderator is continuous (e.g., the mean age of the participants in each experiment)—or a *sub-group meta-analysis* (Borenstein et al., 2011)—if the moderator is categorical (e.g., students vs. professionals). To identify experiment level moderators with IPD-S, we can fit LMMs with interaction terms (Brown & Prescott, 2014; Whitehead, 2002).

Example. The type of subject participating in the experiments (i.e., students vs. professionals) may be impacting the results of the family of experiments on TDD. To assess the role of the type of subject on the results with *AD*, we run a sub-group

meta-analysis. Figure 9 shows the corresponding forest plot. Table 21 summarizes the results of the sub-group meta-analysis.

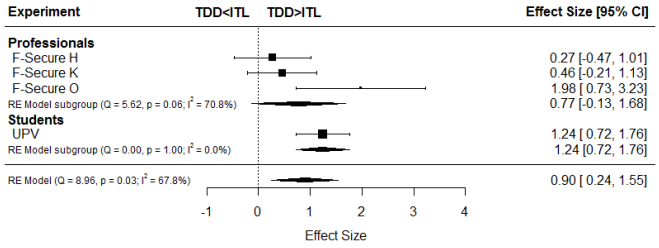

Fig. 9. Forest plot: Professionals vs. Students. **Reprinted by permission [Paper IV] © IEEE 2019.**

Table 21. Sub-group meta-analysis: Professionals vs. Students. **Reprinted by permission [Paper IV] © IEEE 2019.**

Group	N	Estimate	95% CI	I^2
Professionals	3	0.77	(-0.13, 1.68)	70.8%
Students	1	1.24	(0.72, 1.76)	0
Difference	-	0.47	(-0.58, 1.52)	-

As we can see in Figure 9 and Table 21, students seem to obtain more benefits with TDD than professionals. In other words, compared with the performance achieved with ITL, students increase their performance with TDD to a greater extent than professionals. The difference between students and professionals is medium according to rules of thumb (Borenstein et al., 2011). Despite the relevance of such a difference, it does not reach statisical significance. This might be due to the low number of experiments in the family. To sum up, a medium heterogeneity of results materialized in the sub-group of professionals. This may be pointing to the presence of participant level moderators impacting the results.

Finally, we assess the role of the type of subject on the results with *IPD-S*. To do so, we run a LMM with an interaction term. Table 22 shows the result of the LMM fitted.

Table 22. LMM experiment level moderators: professionals vs. students. **Reprinted by permission [Paper IV] © IEEE 2019.**

Interaction	Estimate	95% CI	p-value
Type:Students	16.32	(-37.16, 69.55)	0.545

As we can see in Table 22, students using TDD drop their performance to a lesser extent than the professionals. Besides, the size of the interaction is considerable (i.e.,

M =16.7), at least when keeping in mind that the joint result equals to $M = 28.83$. In sum, professionals obtain less benefits with TDD than students.

As a take-away:

Guideline 4. Use AD and IPD-S in tandem to elicit experiment level moderators. Use meta-regressions and sub-group meta-analyses with AD. Use LMMs with interaction terms with IPD-S.

Step 4.2. Identify participant level moderators with IPD-S

participant level moderators are variables at the participant level that may be impacting the results. For example, the experience of the participants with programming or Java are potential participant level moderators.

IPD-S has been found to be superior to AD for identifying participant level moderators (Fisher et al., 2011; Lambert, Sutton, Abrams, & Jones, 2002). Thus, we recommend relying solely on IPD-S to identify them. Specifically, we recommend relying on LMMs with interaction terms by following Fisher et al.'s approach (i.e., separating the variability of the moderators across and within the experiments (Fisher et al., 2011)).

Example. The experience of the participants with programming, Java, unit testing, and JUnit may be impacting the results of the experiments. To assess their influence on the results we run four LMMs (i.e., one per experience variable) following Fisher et al.'s approach (Fisher et al., 2011). Table 23 shows the results of the LMMs that we fitted. Figure 10 shows the corresponding regression plots.

Table 23. LMM participant level moderators: participants' experiences.

Interaction	Estimate	95% CI	p-value
Programming	15.76	(0.49, 31.04)	0.04
Java	3.85	(-8.13, 15.83)	0.52
Unit testing	11.79	(-5.95, 29.54)	0.18
JUnit	11.07	(-6.26, 28.41)	0.20

As we can see in Table 23 and Figure 10, the larger the experience with programming, Java, unit testing or JUnit, the larger the performance with TDD compared with ITL. In other words, the larger the experience of the participants, the larger the benefits achieved with TDD. However, the extent of such benefit is not statistically significant in three out

of four cases (but in programming experience). This suggests that the results may be due to chance; thus, more experiments are needed to draw definite conclusions on the effect of the participant level moderators on the results.

In sum:

> **Guideline 5**. Use IPD-S to elicit participant level moderators. Use Fisher et al.'s approach (Fisher et al., 2011) and split the variance of the moderators within and across the experiments.

Step 4.3. Acknowledge exploratory analyses limitations

As a final step, we recommend acknowledging the limitations of exploratory analyses:

– **Impossibility of providing cause and effect relationships due to risk of confounding. (Lau, Ioannidis, & Schmid, 1998)** This limitation arises because individual experiments are designed exclusively for providing cause and effect relationships between the treatments and the response variables. Thus, it is impossible to establish cause and effect relationships between other variables (i.e., moderators) and the response variables. For example, if the results of two experiments with different types of subjects vary, it is risky claiming that such difference of results is due to the type of subject. This is because other variables (e.g., the age of the participants across the experiments, the treatment conformance of the participants, etc.) may be also behind the difference of results observed.

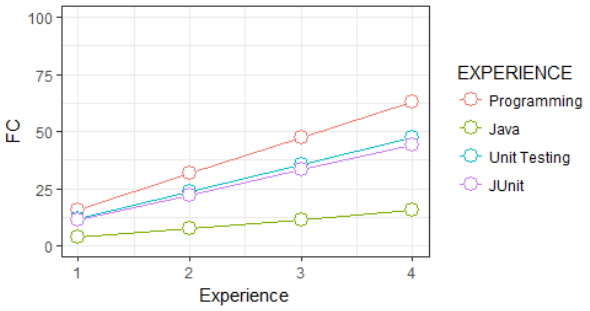

Fig. 10. LMM interactions: participant level moderators. **Reprinted by permission [Paper IV] © IEEE 2019.**

– **Increased risk of statistical errors due to multiple testing** (Quinn & Keough, 2002). This limitation arises because of the multiple statistical tests typically run to identify moderators (e.g., one per moderator). To overcome such limitation some authors recommend relying on multiple testing correction procedures, such as the Bonferroni correction (Quinn & Keough, 2002).

Example. We assessed the effect of one moderator at the experiment level (i.e., type of subject) and four moderators at the participant level (i.e., the experience of the participants with programming, Java, unit testing, and JUnit). We acknowledge that none of the above moderators may be the real cause behind the difference of results observed across the experiments, and that other variables may be the real cause of such differences (e.g., the age, or the treatment conformance of the participants). Besides, we acknowledge an inflated Type I error rate due to the multiple tests that we ran (a total of five tests, one per moderator). Thus, exploratory analyses' results should be taken with caution and only serve as a starting point for guiding future research.

As a take-away:

Guideline 6. Acknowledge exploratory analyses' limitations: impossibility of providing cause and effect relationships due to the risk of confounding, and increased risk of statistical errors due to multiple testing.

3.5 Finding 5: Analysis of an industrial family of experiments on TDD

After learning about the advantages and disadvantages of the aggregation techniques in the SE context, and providing guidelines to analyze SE families, we wanted to apply the most promising aggregation techniques (i.e., IPD-S and AD) to extract knowledge from an industrial family of experiments on TDD.[17] In particular, we wanted to learn: (1) whether the effectiveness of TDD in terms of external quality held across two companies (Section 3.5.2); and (2) whether the effectiveness of TDD in terms of external quality held across the premises of such companies (Section 3.5.2).

[17] Albeit in the previous sections we have analyzed a family with one academic experiment, in this section we purposely change such experiment by an industrial experiment. We do this to study how TDD works across different companies.

3.5.1 The effectiveness of TDD across companies

We ran a total of four experiments (three at F-Secure[18], and one at Bittium[19]) to evaluate the effectiveness of TDD in terms of external quality. We embedded the experiments within seminars on TDD to increase their appeal to practitioners.

Table 24 summarizes the settings of the experiments that we ran. The only difference between the experiments that we ran at F-Secure and Bittium are the programming languages and the testing tools used.

Table 24. Experiments' settings: F-Secure and Bittium.

Aspect	Values
Factors	Development Approach
Treatments	TDD vs ITL
Response variables	QLTY
Design	AB within-subjects
Training	TDD seminar
Training duration	3 days/6 hours
Experiment duration	2.25 hours
Programming language	F-Secure: *Java*; Bittium: *C++*
Unit testing tool	F-Secure: *JUnit*; Bittium: *GTest*

We measured external quality as the *percentage* of test cases that successfully passed from a battery of tests that we (i.e., the researchers) specifically built for testing participants' solutions. Specifically, we measured external quality as

$$QLTY = \frac{\#Test\ Cases(Pass)}{\#Test\ Cases(All)} * 100.$$

A total of six, 11, seven and nine subjects participated at F-Secure H, F-Secure K, F-Secure O, and Bittium, respectively. The subjects were handed a questionnaire some days before the experiment. The questionnaire contained a series of ordinal-scale (i.e., inexperienced, novice, intermediate, and expert) self-assessment questions asking the participants about their experience with programming, unit testing, and the programming language and testing tool used during the experiment.[20]

Table 25 shows the mean—and standard deviations—of the participants' experiences with programming, the programming language, unit testing, and the testing tool used

[18] F-Secure is a multinational cyber-security products company: https://www.f-secure.com/en/welcome
[19] Bittium is a multinational telecommunications company: https://www.bittium.com/home
[20] The questionnaire and its results were published elsewhere (Dieste et al., 2017).

across the experiments (1-4, for inexperienced, novice, intermediate, and experts, respectively).[21]

Table 25. Mean and standard deviation of experiences across experiments. **Reprinted by permission [Paper V] © Springer 2018.**

Experiment	N	Programming	Language	Unit Testing	Testing Tool
F-Secure H	6	3.67 (0.52)	2.33 (1.21)	2.17 (0.98)	2.17 (1.17)
F-Secure K	11	2.91 (0.7)	1.82 (0.87)	1.64 (0.5)	1.27 (0.47)
F-Secure O	7	3.29 (0.76)	2.71 (1.11)	2.71 (0.76)	2 (0.82)
Bittium	9	3 (0.87)	2.89 (0.93)	1.67 (0.87)	1 (0)

We used an **IPD two stages approach** (Burke, Ensor, & Riley, 2017) to learn about the joint effectiveness of TDD at F-Secure and Bittium. We followed and IPD two-stage approach because (Burke et al., 2017): (1) it allows for conveying results in natural units; (2) it allows for portraying the results in a visual manner (i.e., with forest plots (Borenstein et al., 2011)); and (3) it allows for assessing the heterogeneity of results with straightforward statistics such as I^2.

In an IPD two-stage approach, each experiment is first analyzed individually with an identical statistical model (in our case, a dependent t-test because all the experiments are AB within-subjects experiments), and then, the results are pooled together by means of a meta-analysis model (Burke et al., 2017).

After calculating the joint effectiveness of TDD at F-Secure and Bittium, we compared them by means of a sub-group meta-analysis (Borenstein et al., 2011). Figure 11 shows the forest plot corresponding to the sub-group meta-analysis that we ran.

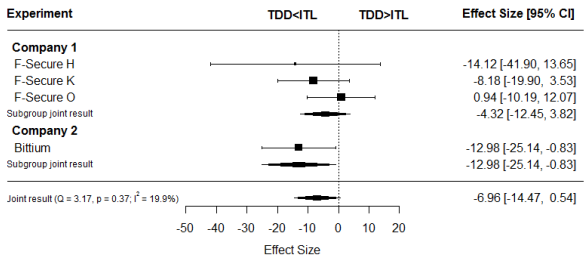

Fig. 11. IPD two-stage approach: F-Secure vs. Bittium. **Reprinted by permission [Paper V] © Springer 2018.**

[21]For simplicity's sake, we consider the variables measured along the survey as continuous. This approach is commonly followed in other disciplines (Norman, 2010).

As we can see in Figure 11, TDD's joint effectiveness at F-Secure ($M = -4.32$) and Bittium ($M = -12.98$) are negative and small—at least considering that QLTY scores can span between 0 and 100. Besides, TDD's joint effectiveness at Bittium overlaps with that at F-Secure (see that Bittium's black-diamond overlaps with that of F-Secure). Thus, in view of the results achieved in each sub-group, *TDD's effectiveness seems similar in F-Secure and Bittium* despite the different programming environments used in each company.

Although the results seem consistent across the companies, heterogeneity of results emerged across the experiments ($I^2 = 19.9\%$). This may be due to the different characteristics of the participants across the experiments (i.e., their different experiences with programming, programming languages, unit testing, and testing tools). In the next section, we analyze the influence of such variables on the results.

3.5.2 The effectiveness of TDD across sites

As we saw, TDD's effectiveness scores ranged between $M = -14.12$ at F-Secure H to $M = 0.94$ at F-Secure O—despite the fact that all F-Secure's experiments had an identical experimental configuration. Such difference of results may have been caused either by unacknowledged characteristics of the experiments, or by the different characteristics of the participants across the experiments.

Because we had access to the responses provided by the participants to the self-assessment questionnaires, we ran a series of IPD-S models (i.e., repeated measures ANOVAs (Field, 2013)), one per variable (i.e., experience with programming, programming languages, unit testing, and testing tools) to check the influence of the characteristics of the participants on the results. For this, we followed Fisher et al.'s advice (Fisher et al., 2011). In particular, we separated the variability of the moderator effects across and within experiments (i.e., by including in the statistical model two interaction terms with the treatment: one with the mean moderator effect in each experiment and another with the difference between the moderator effect of each participant and the mean moderator effect in the corresponding experiment (Fisher et al., 2011)).

Table 26 shows the interaction terms for the participant level moderators in the IPD-S models that we ran.

Table 26 can be read as the impact of a one unit increase in experience on the effectiveness achieved with TDD *beyond that achieved with ITL*. As we can see in Table 26, the experience of the participants with programming and the programming language seems to have an almost negligible positive effect on TDD's effectiveness. However, *the*

Table 26. Participant level characteristics impact on TDD's effectiveness. **Reprinted by permission [Paper V] © Springer 2018.**

Factor	Estimate	SEM	t-value	p-value	Significance
Programming	0.74	5.32	0.13	0.89	✗
Language	0.78	3.94	0.19	0.84	✗
Unit Testing	-3.54	5.09	-0.69	0.49	✗
Testing Tool	-6.80	5.66	-1.19	0.24	✗

larger the experience of the participants with unit testing or testing tools, the larger the benefits with ITL over those achieved with TDD. This may be because the participants who already knew testing (i.e., typically following a test-last approach), may find it difficult "changing gears" and adapting to a reverse testing approach (i.e., test-first). The larger the experience of the participants with unit testing and testing tools, the more beneficial ITL was compared with TDD. It is as if the previous testing experience of the participants made them more resistant to change.

4 Conclusions and future work

Isolated SE experiments commonly face limitations in both the reliability and generalizability of their results (Dybå et al., 2006; Wohlin et al., 2012). This is because SE experiments are typically small, and thus, their results may not be representative of what happens in the population. Also, this is the case because isolated experiments' results cannot be generalized beyond the set up of the experiment.

Families of experiments (i.e., groups of experiments with the same goal run by means of replication) allow for moving beyond the restrictions of isolated experiments' results. They also allow for investigating the effects of the changes made across the experiments on results (i.e., investigating experiment level moderators), or identifying the characteristics of the participants that affect the results (i.e., investigating participant level moderators). However, families of experiments are no panacea: applying unsuitable aggregation technique/s to analyze them may undermine their potential to provide in-depth insights from the experiments' results.

According to the results of a systematic mapping study that we ran, SE families are usually comprised of a low number of experiments, with small and dissimilar sample sizes and opportunistic experimental changes—albeit having identical experimental designs and response variable operationalizations. Families commonly assess the effect of two treatments on different types of subjects. Experiments within families typically provide heterogeneous results.

Five aggregation techniques have been used to analyze SE families:

- *Narrative synthesis*: an aggregation technique that provides a textual description of the experiments' results (in either *p*-value or effect size terms) as a joint conclusion.
- *Aggregated data (AD)*: a statistical procedure that delivers a weighted average of the experiments' effect sizes as a joint conclusion (Borenstein et al., 2011).
- *Individual participant data (IPD) mega-trial*: a statistical test that analyzes the raw data of all the experiments as coming from a single "big" experiment (e.g., by means of the Wilcoxon test (Field, 2013)).
- *Individual participant data (IPD) stratified*: a statistical test such as a linear regression (e.g., ANOVA (Field, 2013)), that analyzes the raw data of all the experiments acknowledging where the raw data came from (i.e., by including an extra parameter accounting for "experiment" in the statistical model).
- *Aggregation of p-values*: a statistical procedure that combines the one-sided *p*-values of all the experiments into a joint *p*-value (Borenstein et al., 2011).

According to the literature on mature experimental disciplines such as medicine and pharmacology, some aggregation techniques are more suitable than others for certain purposes (Higgins & Green, 2011). Not to mention, some aggregation techniques may provide biased results in certain circumstances (e.g., IPD mega-trial (Kraemer, 2000)). The aggregation techniques used to analyze SE families are rarely motivated in research articles. To what extent may the application of different aggregation techniques affect the conclusions reached in a stereotypical SE family?

To answer such question, we applied all the aggregation techniques to analyze a representative SE family: a family of experiments on TDD (Beck, 2003) comprised of four experiments with small and dissimilar sample sizes, identical response variable operationalizations and experimental designs, different types of subjects (i.e., professionals vs. students), and heterogeneous results. We observed that neither narrative synthesis nor aggregation of *p*-values took advantage of the information contained within the raw data to provide joint conclusions. We also noticed that AD and IPD stratified seemed complementary: while the prior provided intuitive visualizations (i.e., forest plots (Borenstein et al., 2011)) and heterogeneity statistics (i.e., I^2 (Borenstein et al., 2011)), the latter provided greater statistical flexibility for conveying joint conclusions, including missing data, or identifying participant level moderators (Brown & Prescott, 2014; Fisher et al., 2011; Whitehead, 2002). We concluded that the aggregation technique/s applied to analyze SE families may largely affect their conclusions.

In view of this, we tailored a procedure with a set of embedded guidelines to analyze SE families of experiments. We tailored a similar procedure to those followed in medicine, but taking into account the particularities of SE families: heterogeneous results, opportunistic recruiting of participants, and small and dissimilar sample sizes. We broke down the procedure into four different steps: (1) describe participants; (2) analyze individual experiments; (3) provide joint conclusions; and (4) perform exploratory analyses. We ended up recommending the use of AD and IPD stratified in tandem to provide joint conclusions and identify experiment level moderators. We recommend using IPD stratified to identify participant level moderators. In addition, we suggest favoring random-effects models over fixed-effects models due to the common heterogeneity of results present in SE families, the multiple changes typically made across SE families' experiments, and the many factors that may influence SE experiments' results.

Finally, we ran and analyzed an industrial family of experiments to learn whether TDD's effectiveness—in terms of external quality—held across two companies and four different sites. For analyzing the data, we relied on the aggregation techniques commonly recommended in mature experimental disciplines (i.e., an IPD two-stage

approach to provide joint conclusions (Burke et al., 2017) and IPD stratified with interaction terms and AD sub-group meta-analysis to elicit moderators (Fisher et al., 2011; Riley, Lambert, & Abo-Zaid, 2010)). Overall, TDD's effectiveness seemed to hold across the companies—despite their different programming environments. However, the participants' characteristics affected the results: the larger the participants' experience with unit testing or testing tools, the lower TDD's effectiveness compared with ITL's. In sum, AD and IPD seem suitable to extract knowledge from SE families.

These findings are only the tip of the iceberg. There still remain other research questions regarding joint data analysis practices in SE families. For example, (1) may families' joint data analyses' results serve to inform about the most "suitable" type of replication to be run next (e.g., identical, or conceptual replication)? (2) How should we account in the joint analysis phase for the potentially "shared bias" of identical replications (B. Kitchenham, 2008)? (3) To what extent would it be possible aggregating experiments' results if the same subjects participated across the experiments? (4) Would it be feasible combining IPD and AD results into a joint conclusion? (5) Could experiments evaluating different treatments be combined together to offer a joint conclusion?. These and many other potential research questions still remain open for further research in SE.

To conclude, we are witnessing an increasing number of SE families of experiments. This may be a consequence of the calls made across the sciences—and SE in particular (Shepperd, Ajienka, & Counsell, 2018)—to move beyond the results of individual studies to make evidence-based and informed decisions (Gurevitch, Koricheva, Nakagawa, & Stewart, 2018; Ioannidis, 2018; Shokraneh et al., 2018; Zwaan, Etz, Lucas, & Donnellan, 2018). Given the momentum of SE families, we think it is timely to reflect on the reliability and transparency of the aggregation techniques used. Otherwise, we may miss the opportunity of obtaining the most from SE families. We may miss the train of replication (Ioannidis, 2018; Zwaan et al., 2018).

References

Abo-Zaid, G., Guo, B., Deeks, J. J., Debray, T. P., Steyerberg, E. W., Moons, K. G., & Riley, R. D. (2013). Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology*, *66*(8), 865–873.

Abrahao, S., Gravino, C., Insfran, E., Scanniello, G., & Tortora, G. (2013). 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.

Ali, S., Yue, T., & Rubab, I. (2014). Assessing the modeling of aspect state machines for testing from the perspective of modelers. In *2014 14th international conference on quality software* (pp. 234–239).

Anello, C., O'Neill, R. T., & Dubey, S. (2005). Multicentre trials: a us regulatory perspective. *Statistical Methods in Medical Research*, *14*(3), 303–318.

Basili, V. R., Shull, F., & Lanubile, F. (1999). Building knowledge through families of experiments. *IEEE Transactions on Software Engineering*, *25*(4), 456–473.

Beck, K. (2003). *Test-driven development: by example*. Addison-Wesley Professional.

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Feldman, H. I. (2002). Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in medicine*, *21*(3), 371–387.

Bero, L., & Rennie, D. (1995). The cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA*, *274*(24), 1935–1938.

Biondi-Zoccai, G. (2016). *Umbrella reviews: Evidence synthesis with overviews of reviews and meta-epidemiologic studies*. Springer.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Brown, H., & Prescott, R. (2014). *Applied mixed models in medicine*. John Wiley & Sons.

Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in medicine*, *36*(5), 855–875.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365.

Canfora, G., García, F., Piattini, M., Ruiz, F., & Visaggio, C. A. (2005). A family of experiments to validate metrics for software process models. *Journal of Systems and Software*, *77*(2), 113–129.

Carver, J. C., Juristo, N., Baldassarre, M. T., & Vegas, S. (2014). Replications of software engineering experiments. *Empirical Software Engineering*, *19*(2), 267–276.

Chu, R., Thabane, L., Ma, J., Holbrook, A., Pullenayegum, E., & Devereaux, P. J. (2011). Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. *BMC Medical Research Methodology*, *11*(1), 21.

Ciolkowski, M., Shull, F., & Biffl, S. (2002). A family of experiments to investigate the influence of context on the effect of inspection techniques. *Proceedings of the Empirical Assessment in Software Engineering, IEEE*.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 997–1003.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*(2), 165.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Debray, T., Moons, K. G., Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H., & Reitsma, J. B. (2015). Get real in individual participant data (ipd) meta-analysis: a review of the methodology. *Research synthesis methods*, *6*(4), 293–309.

Dieste, O., Aranda, A. M., Uyaguari, F., Turhan, B., Tosun, A., Fucci, D., … Juristo, N. (2017). Empirical evaluation of the effects of experience on code quality and programmer productivity: an exploratory study. *Empirical Software Engineering*, *22*(5), 2457–2542.

Dybå, T., Kampenes, V. B., & Sjøberg, D. I. (2006). A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, *48*(8), 745–755.

Erdogmus, H., Morisio, M., & Torchiano, M. (2005). On the effectiveness of the test-first approach to programming. *IEEE Transactions on software Engineering*, *31*(3), 226–237.

Falessi, D., Juristo, N., Wohlin, C., Turhan, B., Münch, J., Jedlitschka, A., & Oivo, M. (2017). Empirical software engineering experts on the use of students and professionals in experiments. *Empirical Software Engineering*, 1–38.

Feaster, D. J., Mikulich-Gilbertson, S., & Brincks, A. M. (2011). Modeling site effects

in the design and analysis of multi-site trials. *The American journal of drug and alcohol abuse*, *37*(5), 383–391.

Fernandez, A., Abrahão, S., & Insfran, E. (2013). Empirical validation of a usability inspection method for model-driven web development. *Journal of Systems and Software*, *86*(1), 161–186.

Fernández, E., Dieste, O., Pesado, P. M., & García Martínez, R. (2011). The importance of using empirical evidence in software engineering. *Computer Science & Technology Series*, 181.

Field, A. (2013). *Discovering statistics using ibm spss statistics*. Sage.

Fisher, D., Copas, A., Tierney, J., & Parmar, M. (2011). A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *Journal of Clinical Epidemiology*, *64*(9), 949–967.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2.

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

Gillett, R. (2003). The metric comparability of meta-analytic effect-size estimators from factorial designs. *Psychological Methods*, *8*(4), 419.

Gómez, O. S., Juristo, N., & Vegas, S. (2014). Understanding replication of experiments in software engineering: A classification. *Information and Software Technology*, *56*(8), 1033–1048.

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*, *57*, 405–429.

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, *555*(7695), 175.

Hedges, L. V., & Olkin, I. (1979). Three vote-counting methods for the estimation of effect size and statistical significance of combined results. In *annual meeting of the american research association, san francisco, california.*

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*(2), 359.

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: foundations and illustrative examples. *Behavior Research Methods*,

39(1), 101–117.

Ioannidis, J. P. (2018). Why replication has more scientific value than original discovery. *Behavioral and Brain Sciences*, *41*.

Ioannidis, J. P., Patsopoulos, N. A., & Rothstein, H. R. (2008). Research methodology: reasons or excuses for avoiding meta-analysis in forest plots. *BMJ: British Medical Journal*, *336*(7658), 1413.

ISO/IEC 25010:2011. (2011). Retrieved 2017-04-10, from `https://www.iso.org/obp/ui/iso:std:iso-iec:25010:ed-1:v1:en`

Juristo, N., & Moreno, A. M. (2013). *Basics of software engineering experimentation*. Springer Science & Business Media.

Juristo, N., & Vegas, S. (2009). Using differences among replications of software engineering experiments to gain knowledge. In *Proceedings of the 2009 3rd international symposium on empirical software engineering and measurement* (pp. 356–366).

Juristo, N., & Vegas, S. (2011). The role of non-exact replications in software engineering experiments. *Empirical Software Engineering*, *16*(3), 295–324.

Kampenes, V. B., Dybå, T., Hannay, J. E., & Sjøberg, D. I. (2007). A systematic review of effect size in software engineering experiments. *Information and Software Technology*, *49*(11), 1073–1086.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1–26.

Kitchenham, B. (2008). The role of replications in empirical software engineering, a word of warning. *Empirical Software Engineering*, *13*(2), 219–221.

Kitchenham, B., Madeyski, L., Budgen, D., Keung, J., Brereton, P., Charters, S., … Pohthong, A. (2017). Robust statistical methods for empirical software engineering. *Empirical Software Engineering*, *22*(2), 579–630.

Kitchenham, B. A., Budgen, D., & Brereton, P. (2015). *Evidence-based software engineering and systematic reviews* (Vol. 4). CRC Press.

Kosar, T., Mernik, M., & Carver, J. C. (2012). Program comprehension of domain-specific and general-purpose languages: comparison using a family of experiments. *Empirical Software Engineering*, *17*(3), 276–304.

Kraemer, H. C. (2000). Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin*, *26*(3), 533–541.

Krein, J. L., Prechelt, L., Juristo, N., Nanthaamornphong, A., Carver, J. C., Vegas, S., … Eggett, D. L. (2016). 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.

Lambert, P. C., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, *55*(1), 86–94.

Lau, J., Ioannidis, J. P., & Schmid, C. H. (1998). Summing up evidence: one answer is not always enough. *The lancet*, *351*(9096), 123–127.

Lewis, J. A. (1999). Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in medicine*, *18*(15), 1903–1942.

Localio, A. R., Berlin, J. A., Ten Have, T. R., & Kimmel, S. E. (2001). Adjustments for center in multicenter studies: an overview. *Annals of Internal Medicine*, *135*(2), 112–123.

Lyman, G. H., & Kuderer, N. M. (2005). The strengths and limitations of meta-analyses based on aggregate data. *BMC Medical Research Methodology*, *5*(1), 14.

Macbeth, G., Razumiejczyk, E., & Ledesma, R. D. (2011). Cliff's delta calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica*, *10*(2), 545–555.

Manso, M. E., Cruz-Lemus, J. A., Genero, M., & Piattini, M. (2008). Empirical validation of measures for uml class diagrams: A meta-analysis study. In *International conference on model driven engineering languages and systems* (pp. 303–313).

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Mouchawrab, S., Briand, L. C., Labiche, Y., & Di Penta, M. (2011). Assessing, comparing, and combining state machine-based testing and structural testing: a series of experiments. *IEEE Transactions on Software Engineering*, *37*(2), 161–187.

Muñoz, L., Mazón, J.-N., & Trujillo, J. (2010). 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.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632.

Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W., Tschannen, B., Altman, D. G., … Jüni, P. (2010). Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *Bmj*, *341*, c3515.

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. In *12th international conference on evaluation and assessment in software engineering* (Vol. 17, pp. 1–10).

Petitti, D. B., et al. (2000). *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine* (No. 31). OUP USA.

Porter, A., & Votta, L. (1998). Comparing detection methods for software requirements inspections: a replication using professional subjects. *Empirical Software Engineering*, *3*(4), 355–379.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication*, *43*(1-2), 103–121.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

Reynoso, L., Genero, M., Piattini, M., & Manso, E. (2005). Assessing the impact of coupling on the understandability and modifiability of ocl expressions within uml/ocl combined models. In *11th ieee international software metrics symposium (metrics'05)* (pp. 10–pp).

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ: British Medical Journal*, *340*, c221.

Santos, A., Gómez, O., & Juristo, N. (2018). Analyzing families of experiments in se: a systematic mapping study. *IEEE Transactions on Software Engineering*, *?*(?), ?–? doi: 10.1109/TSE.2018.2864633

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, *8*(1), 18.

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). Small-study effects in meta-analysis. In *Meta-analysis with r* (pp. 107–141). Springer.

Shepperd, M., Ajienka, N., & Counsell, S. (2018). The role and value of replication in empirical software engineering results. *Information and Software Technology*, *99*, 120–132.

Shokraneh, F., Adams, C., Clarke, M., Amato, L., Bastian, H., Beller, E., ... others (2018). Why cochrane should prioritise sharing data. *British Medical Journal (Clinical research ed.)*, *362*, k3229.

Shull, F., Mendoncça, M. G., Basili, V., Carver, J., Maldonado, J. C., Fabbri, S., ... Ferreira, M. C. (2004). Knowledge-sharing issues in experimental software engineering. *Empirical Software Engineering*, *9*(1-2), 111–137.

Simmonds, M., & Higgins, J. (2007). Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. *Statistics in*

medicine, *26*(15), 2982–2999.

Sjoberg, D. I., Anda, B., Arisholm, E., Dyba, T., Jorgensen, M., Karahasanovic, A., ... Vokác, M. (2002). Conducting realistic experiments in software engineering. In *Empirical software engineering, 2002. proceedings. 2002 international symposium n* (pp. 17–26).

Sjøberg, D. I., Anda, B., Arisholm, E., Dybå, T., Jørgensen, M., Karahasanovic, A., & Vokác, M. (2003). Challenges and recommendations when increasing the realism of controlled software engineering experiments. *Lecture notes in computer science*, 24–38.

Sjoberg, D. I., Dyba, T., & Jorgensen, M. (2007). The future of empirical methods in software engineering research. In *Future of software engineering, 2007. fose'07* (pp. 358–378).

Sjøberg, D. I., Hannay, J. E., Hansen, O., Kampenes, V. B., Karahasanovic, A., Liborg, N.-K., & Rekdal, A. C. (2005). A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, *31*(9), 733–753.

Smith, C. T., Marcucci, M., Nolan, S. J., Iorio, A., Sudell, M., Riley, R., ... Williamson, P. R. (2016). Individual participant data meta-analyses compared with meta-analyses based on aggregate data. *Cochrane Database of Systematic Reviews*(9).

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., & Tierney, J. F. (2015). Preferred reporting items for a systematic review and meta-analysis of individual participant data: the prisma-ipd statement. *JAMA*, *313*(16), 1657–1665.

Stewart, L. A., & Tierney, J. F. (2002). To ipd or not to ipd? advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the health professions*, *25*(1), 76–97.

Stol, K.-J., & Fitzgerald, B. (2018, September). The abc of software engineering research. *ACM Transactions Software Engineering Methodology*, *27*(3), 11:1–11:51. Retrieved from `http://doi.acm.org/10.1145/3241743` doi: 10.1145/3241743

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of Clinical Epidemiology*, *66*(9), 1022–1028.

Wasserstein, R. L., & Lazar, N. A. (2016). *The asa's statement on p-values: context, process, and purpose.* Taylor & Francis.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials* (Vol. 7). John Wiley & Sons.

Wohlin, C., & Aurum, A. (2015). Towards a decision-making structure for selecting a research design in empirical software engineering. *Empirical Software Engineering*, *20*(6), 1427–1455.

Wohlin, C., Höst, M., & Henningsson, K. (2003). Empirical research methods in software engineering. In *Empirical methods and studies in software engineering* (pp. 7–23). Springer.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.

Original publications

I Santos, A., Spisak, J., Oivo, M., and Juristo, N. (2018). Improving Development Practices through Experimentation: an Industrial TDD Case. In *Asia-Pacific Software Engineering Conference* (pp. 465, 473). IEEE doi: 10.1109/APSEC.2018.00061

II Santos, A., Gómez, O. S., and Juristo, N., Analyzing Families of Experiments in SE: a Systematic Mapping Study, in *IEEE Transactions on Software Engineering.* doi: 10.1109/TSE.2018.2864633.

III Santos, A., and Juristo, N. 2018. Comparing Techniques for Aggregating Interrelated Replications in Software Engineering. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '18).* ACM, New York, NY, USA, Article 8, 10 pages. DOI: https://doi.org/10.1145/3239235.3239239. *Best paper award.*

IV Santos, A., Vegas S., Oivo M. and Juristo N. A Procedure and Guidelines for Analyzing Groups of Software Engineering Replications, in *IEEE Transactions on Software Engineering* doi: 10.1109/TSE.2019.2935720

V Santos, A., Järvinen, J., Partanen, J., Oivo, M., and Juristo, N. (2018, November). Does the Performance of TDD hold across Software Companies and Premises? A Group of Industrial Experiments on TDD. In *International Conference on Product-Focused Software Process Improvement* (pp. 227-242). Springer, Cham.

Book orders:
Virtual book store
http://verkkokauppa.juvenesprint.fi

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

UNIVERSITY of OULU
OULUN YLIOPISTO

ISBN 978-952-62-2544-9 (Paperback)
ISBN 978-952-62-2545-6 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)