Analyzing Families of Experiments in SE: A Systematic Mapping Study

Adrian Santos, Omar Gómez and Natalia Juristo

Abstract—**Context:** Families of experiments (i.e., groups of experiments with the same goal) are on the rise in Software Engineering (SE). Selecting unsuitable aggregation techniques to analyze families may undermine their potential to provide in-depth insights from experiments' results.
Objectives: Identifying the techniques used to aggregate experiments' results within families in SE. Raising awareness on the importance of applying suitable aggregation techniques to reach reliable conclusions within families.
Method: We conduct a systematic mapping study (SMS) to identify the aggregation techniques used to analyze families in SE. We outline the advantages and disadvantages of each technique according to mature experimental disciplines such as medicine. We provide preliminary recommendations to analyze and report families in view of families' joint data analysis common limitations.
Results: Several aggregation techniques have been used to analyze SE families: Narrative synthesis, Aggregated Data (AD), Individual Participant Data (IPD) mega-trial or stratified, and Aggregation of p-values. The rationale to select aggregation techniques is rarely discussed within families. Families are commonly analyzed with unsuitable aggregation techniques according to mature experimental disciplines.
Conclusion: Data analysis' reporting practices should be improved to increase the reliability and transparency of joint results. AD and IPD stratified seem suitable to analyze SE families.

Index Terms—Family of experiments, Meta-Analysis, Narrative Synthesis, IPD, AD.

1 Introduction

In 1999, Basili et al. used the term *family of experiments* to refer to a group of experiments that pursue the same goal and whose results can be combined into joint —and potentially more mature— findings than those that can be achieved in isolated experiments [1]. In particular, families of experiments allow to increase the reliability of the findings [2], increase the statistical power and precision of results [3], and assess the impact of experimental changes (i.e., moderators) on results [4], [5], [6].

However, Basili et al.'s definition of family of experiments does not set apart two types of groups of experiments: those gathered by means of Systematic Literature Reviews (i.e., SLRs, a type of *secondary study* that aims to bring together all the available empirical evidence on a particular topic in a systematic way [7]), and those gathered by means of experimental replication —where replications are either conducted by the same researcher or a group of collaborating researchers that share experimental materials, assist to each other during the design, execution and analysis phase of the experiments, etc. [8], [9], [10]. In our opinion, such groups of experiments should be set apart as they grant access to different information and in turn, they may serve to fit different purposes.

For example, while researchers in groups of replications have access to the raw-data of all the experiments (as after all, they have conducted the experiments), this is not guaranteed in SLRs (unless the raw-data are requested to primary studies' authors and they are willing to share them). Thus, while researchers in the prior can apply consistent data-cleaning, data-processing and data-analysis techniques to ensure that differences across experiments' results are just caused by differences in the data gathered, this is unfeasible in SLRs —as potentially different data processing and analysis techniques may have been followed to analyze each individual experiment [11]. This may be detrimental to the reliability of SLRs' joint results.

Also, while researchers in groups of replications are fully aware of the settings and the characteristics of the participants across all the experiments, just the information reported in research articles is available in SLRs. If this information is scarce or incomplete due to reporting inconsistencies or length restrictions, this may limit SLRs' appropriateness to elicit moderator variables (as some moderators may pass unnoticed to the researchers aggregating results).

Another key difference between groups of experiments built by means of replication and those gathered by means of SLRs is that researchers in the prior may opt to introduce isolated changes across experiments with the aim of studying their effects on results. In contrast, experiments gathered by means of SLRs have already set conditions, and thus, differences across experiments' results may be coming from a potential "amalgamation" of effects due to the multiple elements changed across the experiments. So, researchers building groups of replications can avoid the potential impact of introducing multiple simultaneous changes across experiments and thus, minimize the amount

A. Santos is with the M3S (M-Group), ITEE University of Oulu, P.O. Box 3000, 90014, Oulu, Finland, e-mail: adrian.santos.parrilla@oulu.fi
O. Gómez is with Escuela Superior Politécnica de Chimborazo Riobamba, Chimborazo, Ecuador, e-mail: ogomez@espoch.edu.ec
N. Juristo is with the Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus Montegancedo, 28660 Boadilla del Monte, Spain, e-mail: natalia@fi.upm.es

of confounding variables impacting results. In turn, this would not just reduce the bias of joint results, but also the bias of moderator effects [12], [13], [14].

Finally, as groups of experiments gathered by means of replication do not rely on published data to provide joint results (contrary to groups of experiments in SLRs), they do not suffer from the bias introduced in results due to selective publication [11]. Thus, groups of replications may provide potentially less biased conclusions —albeit less generalizable results, as they typically involve fewer experiments— than those provided by SLRs.

To distinguish between groups of experiments gathered by means of SLRs and those gathered by means of replication, we refine Basili et al.'s definition of family of experiments and consider a family as a group of experiments where researchers have *first-hand knowledge of all experiments' settings* and have full access to their *raw-data*. Along this research we focus just on the techniques that have been used to aggregate experiments' results within families and not on those applied in SLRs —where the *de facto* aggregation technique is meta-analysis of effect sizes [7].

This investigation starts from the observation that not much research, if any, seems to have been conducted about the aggregation techniques that have been used in SE to analyze families. Selecting inappropriate aggregation techniques may lead to misleading findings, and in turn, to undermine all the effort involved in conducting a family (e.g., coordinating different research groups for conducting replications across multiple sites, having face-to-face and internet meetings, preparing and translating experimental materials to share among researchers from different countries, etc.).

In this article we conduct a Systematic Mapping Study (i.e., SMS, a type of secondary study where an overview of a specific research area is obtained [15]) with the aim of identifying the techniques that have been used to aggregate experiments' results within families in SE. In addition, we conduct a literature review in mature experimental disciplines such as medicine and pharmacology to learn about the advantages and disadvantages of each technique. Finally, we tailor a preliminary set of recommendations to *analyze* and *report* families based on the common limitations that we found with regard to joint data analysis in SE families. Along the way, we made the following **findings**:

- Families of experiments in SE are commonly constituted by three to five experiments with small and dissimilar sample sizes and multiple changes across experiments. Families usually involve different types of subjects (e.g., professionals and students), and tend to provide heterogeneous results.
- From most to least used, Narrative synthesis, Aggregated Data (AD), Individual Participant Data (IPD) mega-trial or stratified, and Aggregation of p-values have been used to analyze families in SE. Each technique seems appropriate in different circumstances and those shall be understood before aggregating results.
- SE researchers rarely justify the aggregation technique/s used within their families. Narrative synthesis and IPD mega-trial are commonly used to aggregate experiments' results within families despite their numerous shortcomings according to the literature of mature experimental disciplines.
- SE researchers rarely account for the heterogeneity of results that may have materialized as a consequence of the experimental changes introduced across experiments within families. SE researchers rarely acknowledge that differences across experiments' results may have emerged just because of natural variation of results, and not because of the changes introduced across experiments.

The main **contributions** of this research are a *map and classification* of the techniques used in SE to aggregate experiments' results within families, a *list of advantages and disadvantages* of each aggregation technique according to the literature of mature experimental disciplines, and *a set of recommendations* to analyze and report families of experiments in view of families' common limitations with regard to joint data analysis.

Along this article we argue that it is crucial to understand the advantages and disadvantages of each technique before applying them, and that also, the suitability of each technique may be influenced by the characteristics of the family undertaken. Blindly applying aggregation techniques without considering their advantages and disadvantages for the specific conditions of the family may lead to misleading conclusions, and in turn, to miss a valuable opportunity to extract in-depth insights from experiments' results. In view of this, we suggest:

Take-away messages

- The aggregation technique/s used within families should be justified to increase the transparency and reliability of joint results. Common justifications include the availability of the raw-data, the presence of changes across experimental designs or response variable operationalizations, the ability to convey heterogeneity of results in intuitive units, the availability of informative plots to summarize results, or the necessity of interpreting results in natural units.
- We discourage the use of Narrative synthesis to analyze families since it does not provide a quantitative summary of results and does not take advantage of the raw-data to provide joint results or investigate moderators.
- IPD mega-trial seems unsuitable to analyze families of experiments if different types of subjects (e.g., professionals and students) are evaluated within families, or if both missing data —due to protocol deviators or drop-outs— and experiments with dissimilar sample sizes are present within families.
- AD or IPD stratified seem suitable to analyze families of experiments. If multiple changes are introduced across experiments within families (as it is commonly the case in SE), random-effects models may be more suitable than fixed-effects models.

Paper organization. In Section 2 we outline the research method and the research questions of our study. In Section 3, 4, 5 and 6 we provide an answer to each of the research questions of our study. In Section 7 we provide a series of recommendations to analyze families. In Section 8 we provide a series of recommendations to report families. We outline the threats to validity of this study in Section 9. Finally, we outline the conclusions of our study in Section 10.

2 Research Method

We follow the guidelines proposed by Kitchenham and Charters [16] and those proposed by Petersen et al. [15] for conducting our SMS.

2.1 Objectives and Research Questions

The main *objective* of this study is to systematically identify relevant scientific literature and map the techniques that have been used to aggregate experiments' results within families from the *viewpoint* of researchers in the *context* of SE. We propose four different research questions to meet our objective:

- **RQ1.** How are families of experiments defined and characterized?
- **RQ2.** What techniques have been applied to aggregate experiments' results within families?
- **RQ3.** How do the changes introduced across experiments within families influence the aggregation technique/s used?
- **RQ4.** What limitations with regard to joint data analysis are common across families?

2.2 Search and Selection Processes

We iteratively built a search string to identify as many families of experiments as possible. The rationale behind the selection of our final search string is outlined in Appendix A[1]. Our final search string was: *(experiment*) AND (famil* OR serie* OR group*)*. A total of 1213 documents were retrieved on the 14 October 2016 from four databases: IEEE Xplore, ISI Web of Science, Science Direct and Scopus.

We needed to decide on when to consider a group of experiments as a family and exercise such decision to define our *exclusion criteria*. This exclusion criteria would serve us to separate families from other pieces of research. However, we had certain difficulties for this. In particular, we noticed that the distinction between replications and families was not clearly set in SE. This made us aware that family was an ill-defined term and so, we included an extra research question (RQ1) to tackle such issue. In Section 3 we provide an answer to RQ1 and motivate why we regard three experiments as the low boundary for considering a group of experiments as a family.

Eventually, we defined the following *exclusion criteria*:

- The article aggregates fewer than three experiments.
- The article does not compare at least two treatments (e.g., Technology A vs. Technology B) on the same response variable (e.g., quality).
- The article does not report experiments conducted with human participants.
- The article is duplicated.
- The article is not peer-reviewed (e.g., is a call for papers, keynote speech, preface, etc.)

Whenever at least one point of the exclusion criteria was met by an article, then this article was dismissed. Table 1 offers a summary of the selection process that we undertook to select families of experiments. In particular, at Stage 0 the first author went through the list of articles —by sorting them by title, year and author/s— to eliminate duplicates and non-relevant documents. The number of remaining articles at the end of this stage was 572.

In Stage 1, 2 and 3, the first two authors excluded articles by title, abstract reading or in-depth reading, respectively. In case of disagreements, the third author helped to make the decision on whether the article would make it to the next stage. Table 1 shows the number of disagreements and the number of articles making it to the next stage. At the end of Stage 3, 36 articles were considered as relevant. These 36 articles comprised our initial set of primary studies. Such primary studies were used during Stage 4 to perform a *backward snowballing process* (i.e., a procedure where relevant studies are gathered from primary studies' reference lists [17]). During Stage 4, three new articles were selected as relevant. These articles were not indexed in the previous search because: (1) they referred to a group of experiments as a "set"; (2) they contained the "meta-analysis" keyword but no term for referring to the set of experiments; (3) they used the term "replication" in order to refer to a family of experiments. After all stages, we identified a total of 39 primary studies.

2.3 Extraction Process

We designed a data extraction form (Appendix B) to extract all the relevant data from each primary study. We extracted a total of 13 fields of information. We improved each field of information after a first round of screening made by the first author. The purpose of this first screening was to establish categories to classify the aggregation techniques used in families and also the different characteristics that may have influenced the selection of such techniques.

The first and second authors gathered the information of each primary study independently using the final data extraction form. 12 out of 39 articles contained at least one field where a conflict materialized. The third author was consulted to reach to a final agreement in those cases. Conflicts materialized mostly with regard to experiments' sample sizes (some articles reported the final number of subjects after data pre-processing and not others) and on the dimensions changed across experiments (e.g., in cases where it was not clear whether response variables or protocols were changed across experiments). As an example, to solve the inconsistencies in the case of sample sizes, a decision on just using the final number of subjects after pre-processing was made. In case of doubts on whether certain changes had been introduced across experiments within families, a guess could be made by the aggregation technique/s used.

1. Find Appendixes in the supplementary material.

TABLE 1
Summary of results across stages.

Stage	Goal	Total	Excluded	Disagree	Included
Stage 0	Remove duplicates	1213	633	-	580
	Remove non-relevant documents	580	7	-	572
Stage 1	Exclude studies by title	572	383	23	189
Stage 2	Exclude studies by abstract	189	139	15	50
Stage 3	In-depth reading	50	14	7	36
Stage 4	Snowballing process	36	-	-	3

3 RQ1: Family Definition and Attributes

We started this research following Basili et al's definition of family of experiments [1]: a group of experiments that pursue the same goal in order to extract mature conclusions. However, we soon realized that this definition did not provide clear cut-off points between groups of replications and groups of experiments gathered by means of SLRs. In particular, while the scope of a SLR is likely to be wide (as all available research on a particular topic is aimed to be brought together into a joint result), in groups of replications, the goal tends to be narrower (since a small set of hypotheses on a limited set of response variables is usually aimed to be assessed). Besides, in SLRs, just what is reported in primary studies is known. Thus, some relevant information (e.g., characteristics of the participants or some characteristics of the experimental settings) may pass unnoticed if not entirely reported due to length restrictions or reporting inconsistencies —unless experimenters are contacted to share more detailed information or even to share the raw-data. Conversely, in groups of replications —either conducted by the same researcher or a group of collaborating researchers from the same or different groups and/or institutions— access to more detailed information on the experiments is guaranteed. For example, it is typical that in groups of replications researchers share among them laboratory packages, instructional material to ease the execution of experiments [18], assist each other via in-person or internet meetings during the experiment's plan or design phase, assist each other during the execution of replications, etc. [19]. In turn, this close collaboration leads to greater knowledge about all experiments' settings and guarantees full access to all experiments' raw-data. Eventually, this may increase the reliability of the joint results, and in case results differ across experiments, ease the elicitation of moderators. Thus, for us, a main difference between groups of experiments gathered by SLRs and those gathered by means of coordinated replications is the first-hand knowledge of all experiments' settings and access to the raw-data.

As a consequence of the search that we conducted, we also realized that there was no exact cut-off point to discern between: (1) families of experiments; (2) a series of planned, coordinated, or opportunistic replications conducted by a sole researcher —or group of collaborating researchers— and; (3) isolated replications conducted by researchers that neither interact nor collaborate with those who run the baseline experiments.

To start answering RQ1, we consider a group of replications as a family if:

- **At least two treatments** (e.g., Technology A vs. Technology B) are explicitly exercised and compared within all experiments on **a common response variable** (e.g., quality). Three or more treatments may be compared, but two is the minimum. Studies where a treatment is compared with the results reported in literature (e.g., reported industry averages) are excluded. This way, we ensure that the statistical assumptions required by some aggregation techniques (e.g., normality or equality of variances assumption [20]) can be thoroughly checked before providing joint results.
- **At least three experiments** are included within the family —so it is possible providing joint results and study *experiment-level* moderators (e.g., programming language, testing tool, etc.). In particular, if experiment-level moderators are aimed to be assessed within families with less than three experiments, some techniques such as meta-regression (i.e., one of the procedures for studying moderators with meta-analysis of effect sizes [21]) cannot be applied —as meta-regression requires a minimum of a least three experiments because otherwise, a regression line fitted on just two data-points would explain all the variability in the data. By using this lower bound of three experiments: (1) we avoid the problem of considering an isolated replication of a baseline experiment as a family of experiments, and (2) we ensure that all aggregation techniques can be applied within families irrespective of whether it is aimed at providing joint results or investigating moderators.
- **Full access to all experiments' raw-data** is guaranteed to ensure that homogeneous data processing, cleaning and analysis procedures can be applied to each experiment before providing a joint result. Thus, sets of replications gathered from the literature are omitted from the definition of family we propose here.
- **First-hand knowledge of the settings** is guaranteed in all experiments so it is possible minimizing the impact of potentially unknown factors on results, and in case results differ, hypothesize on which moderators may have influenced results.
- **Different subjects participate in each experiment** within the family so data are independent across experiments, as otherwise, results might be biased. In particular, if the same subjects participate across experiments, subjects' scores may be correlated across experiments [22]. If such correlation is not accounted for within the aggregation technique, this issue may invalidate joint results [22].

TABLE 2
Families of experiments' characteristics (ordered by family size).

Size	Sample Sizes	Type of subjects	Raw-data	Package	Venue	Year	ID
12	215 (16, 16, 16, 24, 22, 22, 18, 24, 16, 16, 15, 10)	Students and professionals	✓	✓	PROFES	2015	[P1]
8	455 (42, 39, 29, 35, 31, 31, 172, 76)	Undergraduates	✗	✓	ICST	2012	[P2]
6	126 (29, 44, 53, ?, ?, ?,)	Students and professionals	✗	✓	TSE	2016	[P3]
5	177 (20, 15, 29, 87, 26)	Students and professionals	✗	✗	JSS	2005	[P4]
5	232 (72, 28, 38, 23, 71)	Students, type unknown	✗	✗	MODELS	2008	[P5]
5	284 (55, 178, 13, 14, 24)	Students and professionals	✗	✓	EMSE	2009	[P6]
5	80 (6, 13, 16, 13, 32)	Graduate and undergraduates	✗	✗	AOSD	2011	[P7]
5	112 (24, 24, 28, 20, 16)	Students and professionals	Partially	✓	TSE	2013	[P8]
5	594 (48, 214, 118, 137, 77)	Undergraduates	✗	✗	EMSE	2014	[P9]
5	74 (10, 22, 16, 13, 13)	Graduate and undergraduates	✗	✗	EMSE	2014	[P10]
5	55 (7, 22, 6, 9, 11)	Graduate and undergraduates	✗	✗	TOSEM	2015	[P11]
4	72 (44, 15, 9, 4)	Students and professionals	✗	✗	JSS	2006	[P12]
4	94 (31, 25, 18, 20)	Students and professionals	✗	✗	IST	2010	[P13]
4	74 (13, 35, 18, 8)	Graduate and undergraduates	✓	✗	TSE	2010	[P14]
4	111 (48, 25, 19, 19)	Graduate and undergraduates	✗	✗	TSE	2011	[P15]
4	139 (33, 51, 24, 31)	Graduate and undergraduates	✗	✓	TOSEM	2014	[P16]
4	86 (24, 22, 22, 18)	Students and professionals	✓	✓	TOSEM	2014	[P17]
4	92 (28, 16, 36, 12)	Graduate and undergraduates	✗	✓	IST	2015	[P18]
4	88 (25, 25, 23, 15)	Students and professionals	✓	✓	TOSEM	2015	[P19]
4	81 (11, 16, 22, 32)	Graduate and undergraduates	✗	✓	EMSE	2016	[P20]
3	66 (24, 24, 18)	Students and professionals	✗	✗	EMSE	1998	[P21]
3	60 (20, 20, 20)	Professionals	✗	✗	TSE	2001	[P22]
3	24 (8,8,8)	Professionals	✗	✗	IST	2004	[P23]
3	34 (9, 12, 13)	Graduate and undergraduates	✗	✗	IST	2004	[P24]
3	115 (60, 26, 29)	Graduate and undergraduates	✗	✗	ISMS	2005	[P25]
3	34 (14, 8, 12)	Graduate and undergraduates	✗	✗	ICSE	2008	[P26]
3	15 (8, 5, 2)	Graduates	✗	✗	EMSE	2009	[P27]
3	143 (78, 29, 36)	Undergraduates	✗	✗	IST	2011	[P28]
3	172 (53, 98, 21)	Professionals	✗	✗	IST	2011	[P29]
3	84 (30, 45, 9)	Graduate and undergraduates	✗	✗	IST	2012	[P30]
3	75 (33, 18, 24)	Graduate and undergraduates	✗	✗	RESER	2012	[P31]
3	215 (70, 73, 72)	Undergraduates	✗	✗	EMSE	2012	[P32]
3	79 (19, 31, 29)	Undergraduates	✗	✗	IST	2013	[P33]
3	45 (14, 12, 19)	Graduate and undergraduates	✗	✓	SEKE	2013	[P34]
3	64 (12, 32, 20)	Graduates	✓	✗	JSS	2013	[P35]
3	75 (33, 18, 24)	Undergraduates	✗	✓	EMSE	2014	[P36]
3	92 (20, 25, 47)	Graduates	✗	✗	QRS	2014	[P37]
3	91 (35, 22, 34)	Undergraduates	✗	✗	IST	2015	[P38]
3	169 (40, 51, 78)	Graduate and undergraduates	✗	✓	IST	2015	[P39]

Table 2 shows the 39 families of experiments that we identified. Families are ordered in Table 2 according to family size (i.e., number of experiments). Table 2's columns show respectively: the sample sizes of the experiment, the types of subjects participating in the family, whether the raw-data and laboratory package were provided, the publishing venue of the family, publication date and reference.[2]

With regard to the number of experiments within families (first column), 48% of the families include three experiments, followed by four experiments (23%) and five experiments (20%). There are just three larger families comprised by six, eight and 12 experiments, respectively. The average number of experiments included within families has increased over the years. This may suggest an improvement in the maturity of the area (since the larger the number of experiments within families, the larger the sample size, and thus, the potentially larger the reliability of joint results). However, the increase was slight: around three experiments were included within families between 1998 and 2006, while the average number of experiments in the latest years increased to four and peaked at five in 2015.

In regard to the number of subjects within families (second column), out of 39 families, 12% contained fewer than 50 subjects, whereas 48% contained between 50 and 100 subjects. The larger the number of subjects, the fewer the number of families. 12% of the families included more than 100 subjects but fewer than 150. With regard to experiments' sample sizes (second column in parentheses), most of the experiments identified within families are comprised by between 10 to 30 subjects. The number of experiments containing between 1 and 9 subjects appears to be roughly equal to the number of experiments with 30 to 39 subjects (16 experiments and 19 experiments, respectively). As recent families have on average five experiments and around 25 subjects participate in each experiment, a rise in the total number of subjects is observable over the latest years.

Regarding to the type of subjects involved within families (third column), 38% of the families involve both graduate and undergraduate students, 25% involve both students and professionals, whereas 18% involve just undergraduates. 66% of the families were carried out just with students. Only 7% of the families were conducted entirely with professionals. The number of professionals involved across families along the years seems not to follow any pattern: the number keeps constant around 20 per year.

With regard to whether the raw-data (fourth column)

2. In Appendix C we provide a series of figures to ease the visualization of the data shown in Table 2.

or the laboratory package were provided in families (fifth column), the raw-data were just provided in 15% of the families (even though only 7.5% are accessible as of March 2018) and the laboratory package in 43% of the families (even though most are not accessible as of March 2018). Experimental packages seem to be provided more often from 2012 onwards, what may in turn facilitate replication of experiments. Raw-data provision seems not to have increased over the years. Unfortunately, this prevents re-analysis by third researchers with perhaps more appropriate aggregation techniques than those applied in the original articles.

Regarding publishing venues (sixth column), IST published 25% (10 out of 39) of the families, EMSE published 20% (8 out of 39), TSE published 13% (5 out of 39) and TOSEM published 10% (4 out of 39). The rest of primary studies were published in other venues.

With regard to publication date (seventh column), a steady number of families is observed between 1998 and 2004, whereas a sharp increase is observable from 2008 onwards. This increase is possibly the result of the growing interest in experimentation and the recent calls towards replication in SE. Again, the SE experimentation area might be increasing its maturity.

Finally, we want to make a last observation. In mature experimental disciplines such as medicine, the closest representative of families of experiments (i.e., multicenter clinical trials [20]) are run with pre-stablished protocols defining the experimental settings and the set of procedures that shall be strictly adhered to during the execution and analysis of experiments [23], [24], [25]. Besides, in multicenter clinical trials aiming at assessing the efficacy of new drugs, the populations under assessment across all the centers are specifically defined to ensure consistency of results and avoiding confounding effects [20], [26], [24]. In contrast, in most families of experiments in SE, families are formed without any a-priori plan, and changes are commonly introduced across experiments opportunistically, to either increase the generalizability of results, or to assess moderators. Besides, analysis decisions within SE families seem commonly driven by statistical tests' results (e.g., normality tests' results [27]), the example from other researchers or personal preferences (as when some authors conduct similar analyses to those undertaken in previous families). This conforms to the findings reported in medicine years ago [28].

Summarizing, families of experiments in SE tend to have the following characteristics:

- Most families are comprised by **three to five experiments** with dissimilar and small sample sizes (i.e., less than 30 subjects per experiment) and **involve around 100 participants**.
- Most families are **entirely conducted with students**. However, different types of students are commonly involved. In just three out of 39 families, professionals are the only participants.
- Almost no family provided the raw-data. Less than half of the families provided a laboratory package. **Up to date, almost none is accessible**.
- **Families seem happenstance**. In other words, families do not account with a pre-specified protocol outlining the procedures for either conducting experiments, analyzing each of them, or aggregating their results.
- Most families are published in **journals (IST, EMSE, TSE, TOSEM)**.

4 RQ2: Analysis Techniques Applied

Table 3 shows the techniques that have been used to aggregate experiments' results within families (from most to least used), the amount of families that apply each technique, and the references of the families.[3]

TABLE 3
Aggregation technique by family of experiments.

Technique	N	Primary Studies
Narrative synthesis	18	[P21][P25][P4][P12][P14][P13] [P29][P2][P32][P35][P37][P10] [P16][P9][P36][P11][P19][P20]
AD	15	[P22][P24][P5][P6][P28][P30] [P35][P8][P33][P17][P39][P18] [P1][P38][P20]
IPD mega-trial	13	[P23][P26][P27][P14][P28][P15] [P7][P30][P34][P33][P10][P39] [P11]
IPD stratified	6	[P21][P26][P31][P16][P36][P3]
Aggregation of *p*-values	3	[P22][P24][P29]

Narrative synthesis was used in 46% of the families. Narrative synthesis is an aggregation technique that provides a *textual* summary of results as a joint conclusion [29]. Families applying Narrative synthesis do not explicitly mention any term to refer to this aggregation technique (even though some use the term "global analysis" [P13], [P4]). In general, it is hard to distinguish whether authors are aggregating experiments' results or just comparing them. For example, Scanniello et al. [P1] summarizes results as *"...this is true of all the experiments, the only exception being [experiment X] on the [response variable] GD, where the statistical test returned a p-value equal to 0.39..."*, and Staron et al. [P12] as *"...in general the stereotypes improve... and half of the results were statistically significant..."*. The main advantage of Narrative synthesis is that, as just a textual summary of results needs to be provided, Narrative synthesis allows to combine the findings of experiments with wildly different experimental designs, response variables or statistical tests into joint results [30], [29]. Besides, in Narrative synthesis, discordance across experiments' results are not seen as a threat to the validity of joint results, but instead, as an opportunity to study the effect of moderators on results [6]. For example, in view that results differ across experiments, Ali et al. [P37] claim *"...one plausible explanation is that participants had more experience with standard UML state machines before the experiment than Aspect state machines and hence their understandability..."*, and Ricca et al. [P26] claim *"...the use of stereotypes does not always introduce significant benefits... the provided material is different, and this can be the reason for different results..."*.

3. We map primary studies to analysis technique/s in Section 5.

Even though using Narrative synthesis is straightforward, it has some relevant shortcomings. In particular, Narrative synthesis does not provide a joint effect size or p-value. In turn, this hinders the incorporation of results in prospective studies. Another of its main shortcomings is that a subjective weight (i.e., dependent upon the analyst) is commonly assigned to each experiment towards the overall conclusion [21] (e.g., should all experiments be weighted identically regardless of their sample size?; should experiments with professionals be weighted more than those with students —as after all, professionals are more representative of what may happen in reality?). In turn, the reliability and reproducibility of results is hindered with Narrative synthesis [21].

Aggregated Data meta-analysis (i.e., AD) was used in 38% of the families. AD is commonly known in SE as *meta-analysis of effect sizes*, and is the *de facto* aggregation technique to aggregate experiments' results in SLRs [7]. Families' authors refer to AD simply as *meta-analysis*: *"the set of statistical techniques used to combine the different effect sizes of the experiments"* [P8]. Effect sizes quantify the relationship between two groups (or more generally, between two variables: the dependent and the independent variable [21]). Effect sizes can be computed from experiment's summary statistics (e.g., means, standard deviations and sample sizes [31]) or from statistical tests' results (e.g., t-test's t-value and degrees of freedom [27]).

Effect sizes are commonly divided into two big families [32]: the r family and the d family. The r family's effect sizes quantify the strength of relationship between two variables. The strength of this relationship is usually provided as a Pearson correlation [27]. For example, Gonzalez et al. [P18] calculated the Pearson correlation between two variables in each experiment and then aggregated them using AD. Manso et al. [P5] followed an identical procedure but this time, they used the Spearman correlation instead (a non-parametric correlation coefficient [27]). The d family's effect sizes quantify the difference between the means of two groups. The size of the difference is usually conveyed in *standardized* terms as to rule out differences across experiments' response variable scales. Cohen's d or Hedge's g [33], [34] are common representatives of the d family in SE [35]. For example, Fernandez et al. [P35] analyzed each experiment with the U-Mann Whitney (the non-parametric counterpart of the independent t-test) and aggregated all experiments' Hedge's gs into a joint result by means of AD.

AD is a statistical technique that delivers a *weighted* average of all experiments' effect sizes as a joint effect size [21]. Generally speaking, the *weight* given to each experiment towards the joint result with AD is directly proportional to the sample size of the experiment —if a fixed-effects model is used [21]— or to the sample size of the experiment and the total *heterogeneity* of results (i.e., the variability of results that cannot be explained by natural variation) —if a random-effects model is used [21].[4]

One of the main advantages of AD is that it can be used to combine the results of experiments with different designs and response variable scales into a joint conclusion —as long as a suitable standardized effect size can be calculated [21], [36]. In addition, AD can easily handle heterogeneity of results (by simply fitting a random-effects model instead of a fixed-effects model [21]) or elicit experiment-level moderators (e.g., by using meta-regression or sub-group meta-analysis [21]). Other advantages of AD are its intuitive visualizations (i.e., forest-plots [21]) and its straightforward statistics to quantify heterogeneity (e.g., I^2 [21]).

Figure 1 shows an example of forest-plot. As it can be seen in Figure 1, the effect size of each experiment is represented by a square. The size of each square represents the weight of the effect size in the overall result (see the black diamond at the bottom) and the width of the line crossing each square represents the uncertainty of the effect size in each experiment (i.e., its 95% confidence interval [36]). Furthermore, the assessment of the heterogeneity is straightforward by means of the I^2 statistic (e.g., the heterogeneity can be small, medium or large if the I^2 statistic is larger than 25%, 50% or 75% respectively [21]).

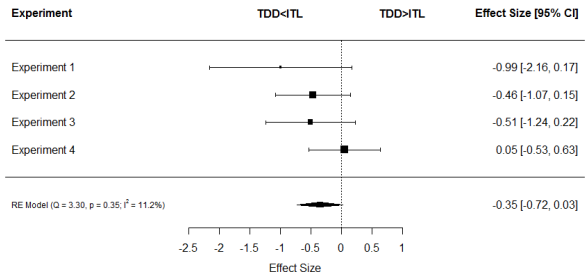

Fig. 1. Forest-plot: AD example.

Even though appealing, AD has among its main limitations that it cannot assess at the same time the effect of multiple factors on results (e.g., the effects of the treatments, experimental tasks and their interactions) and instead, AD is commonly applied to just aggregate experiments with relatively simple designs [21]. Another shortcoming of AD is that effect sizes' statistical assumptions need to be checked before providing joint results (e.g., normality or homogeneity of variances for Cohen's d [37][38][39]).

Individual Participant Data mega-trial[5] (i.e., IPD mega-trial) was used in 33% of the families. In IPD mega-trial, the raw-data of all experiments are *pooled* together into a joint data-set and then analyzed as if the raw-data were coming from a single "big" experiment [P27]. Families' authors in SE name IPD mega-trial after the statistical model being applied (e.g., *ANOVA*, *GLM*, etc. [P11]). As an example of its application, Cruz et al. [P28] analyzed individually each experiment by means of the Kruskal-Wallis test (i.e., the non-parametric counterpart of the one-way ANOVA [27]) and then pooled together all experiments' raw-data into a joint data-set to analyze them jointly with the same test. Ricca et al. [P26] followed an identical procedure, but this time with the Wilcoxon test (i.e., the non-parametric counterpart of the dependent t-test [27]). Table 4 shows the IPD mega-trial

4. Assuming a common variance across experiments.

5. The term "mega-trial" is not to be mistaken with the term "multi-centre" trial. While mega-trial refers to an analysis approach [28], "multi-centre" trial refers in medicine to the multiple experiments conducted at different sites with a common underlying protocol [20], [40].

statistical models that were used within families. As Table 4 shows, non-parametric tests are dominant.

TABLE 4
IPD mega-trial statistical model by family.

Statistical Test	Primary studies
Non-parametric	[P26][P27][P28][P15][P33][P39][P11][P10]
ANOVA	[P14][P7][P30]
Others	[P23][P34]

Despite its intuitiveness, IPD mega-trial may provide biased results if experiments are unbalanced across treatments (e.g., due to missing data because of protocol deviators or drop-outs) and have different sample sizes, or if subjects resemble more to each other within the same experiment than across experiments [41], [42], [43] (e.g., what may happen when experiments with either professionals or students are run within families). As a result, the use of IPD mega-trial is commonly discouraged [41], [42], [43].

Another disadvantage of IPD mega-trial is that some statistical tests cannot analyze jointly experiments with different designs. For example, if within-subjects experiments (i.e., repeated-measures experiments) and between-subjects experiments need to be analyzed together by means of a repeated-measures ANOVA [27], the repeated-measures ANOVA "throws away" all the data coming from between-subjects experiments (because data in such experiments are not repeated within subjects). Another shortcomings of IPD mega-trial are that it needs all experiments to use identical response variable scales —as otherwise, the joint effect provided may be biased depending upon their operationalizations [11]— and that IPD mega-trial statistical models are built on top of some statistical assumptions (e.g., normality or homogeneity of variances assumption [27]) that need to be checked before interpreting results. Finally, heterogeneity of results across experiments cannot be included within IPD mega-trial statistical models (as the experiment where the raw-data come from is not accounted for in IPD mega-trial). However, IPD mega-trial's statistical flexibility is a plus: some statistical models (e.g., ANOVA [27]) allow to include as many factors as desired to model the relationship between the data and the characteristics of the family. For example, Ricca et al. [P14] fitted an IPD mega-trial ANOVA model with experience (i.e., undergraduate, graduate, research assistant), and separately another ANOVA with ability (high and low) with the aim of checking the effect of experience and ability on results, respectively.

Individual Participant Data stratified (i.e., IPD stratified) was applied in 15% of the families. As in IPD mega-trial, IPD stratified involves the central collection and processing of all experiments' raw-data into a joint data-set. However, instead of analyzing the raw-data jointly as coming from a single "big" experiment, in IPD stratified, a factor representing the experiment where the raw-data come from is included within the statistical test [28]. This relationship is considered within statistical models by including an extra factor within statistical models: "Experiment". As an example, commonly applied IPD stratified statistical tests are ANOVAs accounting for two factors: "Treatment" and "Experiment". Again, families' authors using IPD stratified refer to it after the name of the technique applied (e.g., *ANOVA* [P21], *linear regression* [P3], etc.) -even though some have called it *comprehensive analysis* [P16]. As an example of its application, Runeson et al. [P31] analyzed individually each experiment with a Wilcoxon test and then the family as a whole with an ANOVA model including "Treatment" and "Experiment" as factors. Table 5 shows the IPD stratified statistical models applied within families. Table 5 shows that ANOVAs are the most used in IPD stratified [27].

TABLE 5
IPD stratified statistical model by family.

Statistical Test	Primary studies
ANOVA	[P21][P26][P31][P36]
Linear regression	[P3]
Permutation test	[P16]

Contrary to IPD mega-trial, IPD stratified allows to include heterogeneity of results across experiments. In addition, IPD stratified allows to include flexible statistical assumptions within statistical models (e.g., different variances across experiments vs. identical variance across experiments, etc.) so as to increase the reliability of joint results [20]. IPD stratified is considered as the *gold-standard* in medicine to analyze groups of interrelated experiments when the raw-data are accessible [44], [45], [28].

The main shortcomings of IPD stratified are its complexity for assessing heterogeneity of results (as no straightforward statistic such as I^2 is yielded by IPD stratified models), its reliance on identical response variables across experiments for providing joint results, and the difficulty of fitting and checking the statistical assumptions of some relatively complicated statistical models (e.g., Linear Mixed Models [20]).

Finally, **Aggregation of p-values** was used in just 7% of the families. In Aggregation of p-values, all *one-sided* experiment's p-values are pooled together by means of a statistical model such as Fisher's or Stouffer's method [21].[6] For example, Laitenberger et al. [P22] analyzed each experiment with a *one-sided* dependent t-test and then pooled their p-values into a joint p-value by means of the Fisher method. Aggregation of p-values main advantage is that it can aggregate the p-values of experiments with whatever design, response variables or statistical tests into joint results [46], [47], [20]. Unfortunately, among Aggregation of p-values' main advantages are that it cannot provide a joint effect size -but just a joint p-value. In turn, as p-values confound effect size and sample size (i.e., a small p-value may emerge because of a large and relevant effect size, or due to a huge sample size and an almost negligible effect size [36]), the interpretability of results is not straightforward [21], [20]. Another of the main disadvantages of Aggregation of p-values is that in its basic procedures (e.g., Fisher or Stouffer's methods [21]) an identical weight is assigned to each experiment regardless of its quality or sample size. This may affect the reliability joint results [20].

6. Even though more advanced Aggregation of p-values techniques exist also [20], just the Fisher's and Stouffer's methods [21] have been used in SE.

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**	**✗ Provides no effect size nor p-value** ✗ Subjective weighting ✗ Not reproducible results
✓ **Independent of design and metric** ✓ Straightforward visualizations ✓ Moderators and heterogeneity	**AD**	✗ Statistical assumptions ✗ Simple designs
✓ Intuitive approach ✓ Heightened statistical flexibility ✓ Moderators	**IPD mega-trial**	**✗ Biased results may be provided** ✗ Statistical assumptions ✗ Dependent on design, response variable
✓ **Heightened statistical flexibility** ✓ Moderators and heterogeneity	**IPD stratified**	✗ Statistical assumptions ✗ Dependent on design, response variable ✗ Complexity
✓ Independent of design, metric or statistical test	**Aggregation of p-values**	**✗ Provides no effect size** ✗ p-value dependent on sample size and effect size

Table 6 shows a summary of the advantages and disadvantages of each technique to analyze families of experiments.

5 RQ3: Technique Selection Within Families

The changes introduced across experiments within families may impact the suitability of the aggregation technique/s applied. For example, Runeson et al. [P36] claimed that as many changes were made in the third experiment, its results could not be aggregated [P36]. This argument was also used previously in SE to prevent aggregation of experiments' results [48]. Thus, we thought it would be sensible to investigate whether certain changes across experiments within families hindered or benefited the application of certain aggregation techniques.

According to Gomez et al. [2], different dimensions can be changed across experiments:

- **Operationalization**: refers to the operationalization of the treatments, metrics and measurement procedures used within the experiments (e.g., response variable scales, the use of test cases or experts to score participants' solutions, etc.)
- **Population**: refers to the characteristics of the participants within the experiments (e.g., students vs. professionals, different skills and backrounds, etc.)
- **Protocol**: refers to the "apparatus, materials, experimental objects, forms and procedures" used within the experiments (e.g., experimental tasks, experimental session length, training duration, etc.)
- **Experimenters**: refers to the personnel involved within the experiments (e.g., the trainer, the measurer, the analyst, etc.)

Table 7 shows a map between the families that we identified (first column), the aggregation techniques that they used (second column, where the aggregation techniques may be [N]arrative synthesis, Aggregation of [P]-values, [A]D, IPD [M]ega-trial or IPD [S]tratified), the dimensions that changed across their experiments (third column, where the dimensions changed may be the [O]perationalization, [P]opulation, [Pr]otocol or [E]xperimenters), and other information that we will discuss later.

The population dimension is the one that varies the most within families. Populations are commonly changed across experiments to increase the external validity of results [P21], [P13], [P1]. For example, Porter and Votta [P21] claimed that one of the goals of their family was to *"...extend the external credibility of our results by studying professionals developers..."*. Unfortunately, introducing population changes across experiments may potentially affect the effect size being estimated and thus, impact conclusion validity [2]. For example, a treatment could be effective for students but not for professionals (or viceversa). In such circumstance, as each experiment is estimating a potentially different effect size, and as experiments' sample sizes are usually small in SE, an "amalgamation" of potentially unreliable effect size estimates is being offered as a joint result within families —if a fixed-effects model such as ANOVAs, linear regressions or fixed-effects meta-analysis models are used [21], [20], [49] (as it is commonly the case within families). This issue becomes more relevant the larger the number of changes introduced across experiments [2] and the smaller the sample sizes [50].

Judging by the number of dimensions changed within families and the frequencies of use of each aggregation technique, we cannot observe any relationship between the number of changes introduced and the aggregation technique/s used. For example, it would have been expected to find the more strict techniques (i.e., IPD mega-trial or stratified) being applied less often than others such as Narrative synthesis in families with many dimension changes. Even though the number of IPD mega-trial analyses conducted in such cases seems small (only five families with three or four dimension changes used IPD mega-trial), almost all IPD stratified analyses have been conducted in families introducing three to four dimension changes across the experiments. We read this as preliminary evidence suggesting that researchers seem not to follow a procedure for selecting aggregation techniques driven by the characteristics of the family. This agrees with the findings made in medicine years ago [28].

We identified several other elements that may have also influenced the selection of the aggregation technique/s. The

rest of the columns of Table 7 go over such elements. Let us examine them one by one.

Response variable changes appear in the fourth column of Table 7. As IPD cannot accommodate different response variables across experiments, changing response variables' operationalizations may hinder the application of IPD and at the same time, increase the appeal of other techniques (e.g., AD with standardized effect sizes such as Cohen's d). According to Table 7, response variables' operationalizations rarely change across experiments (only in 20% of the families, including families where we lack information). As expected, whenever response variables' operationalizations change, either AD, Narrative synthesis or Aggregation of p-values were used to aggregate experiments' results. Even though Narrative synthesis and Aggregation of p-values seem appealing in this circumstance, we recommend to use AD instead (e.g., as Scanniello et al. did [P1]), as AD weights *transparently* each experiment towards the joint result and thus, increases the reliability and transparency of results, and at the same time, provides an effect size and a p-value, and thus, allows to assess both the relevance and the significance of results [35] —contrary to Aggregation of p-values that just provides a p-value and Narrative synthesis that provides no quantitative summary of results.

Experimental design changes appear in the fifth column of Table 7. As some IPD models cannot accomodate groups of experiments with different experimental designs (e.g., repeated-measures ANOVA can only analyze experiments with within-subjects designs [27]), introducing changes across experiments' designs may hinder the application of IPD and favour the application of other techniques such as Narrative synthesis, AD or Aggregation of p-values. As shown in Table 7, design changes were rarely introduced across experiments within families (in only 15% of the families). In 66% of those families, Narrative synthesis was used. For example, Juristo et al. [P2] analyzed independently each experiment with an ANOVA model with four factors (i.e., Technique, Program, Version and Fault) and then categorized each factor into three categories (i.e., non-significant, significant or "doubtful"), depending on the number of times each factor was significant across experiments. In spite of its appeal, Narrative synthesis may be specially dangerous given the small sample sizes common in SE experiments [51] and thus, the large variability of results expected because of natural variation of results [52]. For example, if two exact replications, each with a statistical power of 30% are conducted (and thus, there is a 30% probability of obtaining statistically significant results in each), there is a 0.09 (i.e., 0.3*0.3) probability of achieving two statistically significant results and a 0.49 (i.e., 0.7*0.7) probability of obtaining two non-significant results. Thus, there is a 0.42 (i.e., 1-0.49-0.09) probability of obtaining one significant and one non-significant result and thus claiming conflicting results -when in reality both experiments' estimate exactly the same population effect size [36], [14]. In addition, finding two non-significant results across two different experiments does not imply that the joint result is not statistically significant [21]: simply that larger sample sizes would have been required to achieve statistical significance in each individual experiment [36].

The analysis technique/s used to analyze each experiment individually within families are shown in the last column of Table 7 (i.e., [A]NOVA, [N]on-parametric, [T]-test, [C]orrelation, [R]egression, [O]thers or None (-)). We assessed the techniques used for analyzing each experiment individually as it may be "tempting" to use the same technique for analyzing the family as a whole if all experiments are identical (e.g., by pooling the raw-data of all experiments together and then analyzing them as coming from a "big" experiment).

As it can be seen in Table 7, non-parametric statistical tests (e.g., U-Mann Whitney, Wilcoxon, etc. [27]) have been largely used to analyze individual experiments within families (in 51% of the families). The most common reason for using non-parametric tests seems the lack of normality of the data. For example, Fernandez et al. [P20] claimed that they used the Wilcoxon test to analyze each experiment individually as *"...in most cases the data were not normal..."*, while Cruz-Lemus et al. [P28] claimed that they used the Kruskall-Wallis test as *"... Kruskal-Wallis is the most appropriate test... when there is non-normal distribution of the data..."*. Among those families relying on non-parametric tests to analyze individual experiments, 62% rely on the same test to analyze the family as a whole (and thus, use IPD mega-trial), regardless of the overall sample size achieved at the family level. For example, Hadar et al. [P33] used the Mann-Whitney test to analyze each individual experiment and then the family as a whole —despite having achieved a sample size of around 80 whenever pooling the raw-data together.

Even though such procedure seems consistent, this may not be optimal, as traditional statistical tests such as ANOVA are robust to departures from normality [53] (specially when sample sizes get larger, as what happens when pooling the raw-data of all experiments together [54]), and they should be preferred over their non-parametric counterparts when sample sizes get larger [55]. We suggest that this large reliance on IPD mega-trial in SE may be caused by the lack of normality of the data, the common use of non-parametric tests to analyze individual experiments, and the impossibility of accommodating more factors apart from "treatment" within traditionally used non-parametric tests (e.g., the U-Mann Whitney, the Wilcoxon, etc. [27]).

6 RQ4: Families Limitations

We found a series of common limitations with regard to joint data analysis in the families that we identified.

For example, there is an over-reliance on Narrative synthesis to provide joint results. Narrative synthesis has been discouraged in mature fields such as medicine or pharmacology due to its inability to provide a quantitative summary of results and its subjectivity when providing joint results [11], [21]. Apart from the above, and as we saw previously, Narrative synthesis fails to take into account the natural variation of results [21]. This issue may be specially relevant in SE, where small sample sizes are the norm rather than the exception [51], and thus, a large variability of results is expected [56]. As an example, and perhaps unknowingly, Narrative synthesis may have tricked Ceccato et al. [P10] into thinking that conflicting results materialized across two small experiments (with sample sizes of 13 each)

TABLE 7
Changes introduced within families (ordered by total number of changes).

ID	Techniques	Changes	Response	Design	Individual
[P2]	N,-,-,-,-	O,P,PR,E	?	✓	A
[P4]	N,-,-,-,-	O,P,PR,E	✗	✗	C
[P5]	-,-,A,-,-	O,P,PR,E	✗	✗	C
[P6]	-,-,A,-,-	O,P,PR,E	?	✗	A
[P31]	-,-,-,-,S	O,P,PR,E	✗	✗	N
[P36]	N,-,-,-,S	O,P,PR,E	✗	✗	N
[P1]	-,-,A,-,-	-,P,PR,E	✓	✓	-
[P3]	-,-,-,-,S	-,P,PR,E	✗	✗	R
[P7]	-,-,-,M,-	-,P,PR,E	✗	✗	-
[P8]	-,-,A,-,-	-,P,PR,E	?	✗	N
[P9]	N,-,-,-,-	O,P,PR,-	✓	✓	A
[P11]	N,-,-,M,-	O,P,-,E	✗	✗	N,O
[P15]	-,-,-,M,-	-,P,PR,E	✗	✗	T
[P16]	N,-,-,-,S	-,P,PR,E	✓	✗	N
[P17]	-,-,A,-,-	-,P,PR,E	✗	✗	N,T
[P18]	-,-,A,-,-	-,P,PR,E	✗	✗	N
[P20]	N,-,A,-,-	-,P,PR,E	?	✗	N
[P24]	-,P,A,-,-	-,P,PR,E	✗	✗	T
[P26]	-,-,-,M,S	-,P,PR,E	✗	✗	N
[P28]	-,-,A,M,-	-,P,PR,E	✗	✗	N
[P29]	N,P,-,-,-	O,P,PR,-	✓	✓	N
[P35]	N,-,A,-,-	O,P,PR,-	✗	✗	T
[P37]	N,-,-,-,-	-,P,PR,E	✗	✓	N
[P38]	-,-,A,-,-	O,P,PR,-	✓	✗	C
[P10]	N,-,-,M,-	-,P,-,E	✗	✗	N,O
[P12]	N,-,-,-,-	-,P,PR,-	✗	✗	N,T
[P13]	N,-,-,-,-	-,P,PR,-	✗	✗	C
[P14]	N,-,-,M,	-,P,-,E	✗	✗	N,T
[P19]	N,-,-,-,-	-,P,PR,-	✗	✗	N
[P22]	-,P,A,-,-	-,P,PR,-	✗	✓	T
[P25]	N,-,-,-,-	-,P,-,E	✗	✗	C
[P27]	-,-,-,M,-	-,P,-,E	✗	✗	-
[P30]	-,-,A,M,-	-,P,-,E	✗	✗	A
[P32]	N,-,-,-,-	-,P,PR,-	✗	✗	N
[P33]	-,-,A,M,-	-,P,PR,-	✗	✗	N
[P34]	-,-,-,M,-	-,P,-,E	✗	✗	N,T
[P21]	N,-,-,-,S	-,P,-,-	✗	✗	A
[P23]	-,-,-,M,-	-,-,PR,-	✗	✗	-
[P39]	-,-,A,M,-	-,P,-,-	✗	✗	N,A

when they claimed that *"...strangely enough, the trend observed... in Exp IV and V has alternating directions"* or Runeson et al. [P36] when they claimed that *"...the first replication was designed to be as exact as possible... Despite an attempt at an exact replication, the outcomes were not the same..."* in two experiments with sample sizes 33 and 18, respectively. In particular, conflicting results (in either *p*-value terms or effect size terms) may materialize just because of the presence of small sample sizes, and thus, the large variation of results expected, and not because experiments observe different realities [50], [14].

Narrative synthesis has also been used to assess moderator effects in 15% of the families. Again, and despite its appeal, Narrative synthesis may be unreliable to detect moderators, specially when experiments are small. In particular, as a large variability of results is expected in small sample sizes [50], [14], there is a high risk of claiming that differences across experiments' results are due to moderator effects —when in reality, differences across experiments' results could have just emerged by natural variation of results. For example, in view that one experiment provided different results than the rest, Ali et al. [P37] claimed that *"...One plausible explanation is that participants had more experience with standard UML state machines..."*. On its side, after noticing that different results were achieved across two experiments within the family, Juristo et al. [P2] claimed that *"subjects might swap information about the programs and their faults at the end of each session. As a result of copying, the techniques applied... could be more effective at UdS and UPV than at other sites"*. Even though such claims may add to the discussion, they may also be misleading. Unfortunately, Narrative synthesis does not allow to distinguish how much variability in the results comes from natural variation and how much does not [50], [14]. This may impact the reliability of the conclusions reached within families using Narrative synthesis to elicit moderators.

With regard to the use of AD to provide joint results, we have noticed some inconsistencies in its use. For example, in some families non-parametric statistical tests (e.g., U-Mann Whitney [27]) were used to analyze individual experiments as data did not follow normality. Then, parametric effect sizes (such as Cohen's d or Hedge's g) were calculated for each experiment to integrate them with AD. As an example, Fernandez et al. [P39] analyzed each individual experiment by means of the U-Mann-Whitney (as data did not follow normality), and then computed the Hedge's g of each experiment to provide a joint result by means of AD. Unfortunately, the use of parametric effect sizes (such as

Cohen's d or Hedge's g) also comes at the cost of checking the statistical assumptions on which they are built onto (e.g., normality and homogeneity of variances [39], [38], [57]). This becomes more relevant the smaller the sample sizes [38] as otherwise, there is a risk of providing biased results due to the a-priori unknown shape of the sampling distribution of each experiment's effect size [39]. Thus, we suggest that if non-parametric tests are used to analyze individual experiments, at least for consistency's sake, non-parametric effect sizes such as Cliff's delta should be used to provide joint results with AD [39], [38], [57].

Also with regard to the use of AD, we have noticed that despite the multiple changes usually introduced across experiments within families (and thus, the potential heterogeneity of results inadvertently introduced within families), fixed-effects models are commonly preferred over random-effects models —perhaps because of their generally smaller p-values [21], [20], and in turn, their more "significant" results. For example, Scanniello et al. [P1] analyzed a group of five different replications with a fixed-effects meta-analysis model (despite obtaining a relatively large and statistically significant heterogeneity of results) and Cruz et al. [P6] analyzed a group of nine effect sizes with fixed-effects models despite the observable heterogeneity in the forest-plot. Unfortunately, not acknowledging heterogeneity of results during the statistical analysis may limit the reliability of joint results [21]. This may be worrying, specially if a large number of changes have been introduced across experiments.

With regard to the use of IPD mega-trial to analyze families, and as previously discussed in medicine [43], [42], IPD mega-trial may provide biased results if subjects resemble more to each other within experiments than across experiments (e.g., when some experiments are run with professionals and others with students), or if data are unbalanced across treatments and experiments (e.g., when some experiments are larger than others, and missing data materializes in some experiments but not in others). Given that missing-data is common in SE due to protocol deviators or drop-outs [58], that families commonly conduct experiments with professionals and students, and that experiments run with professionals tend to be smaller than those run with students, we are skeptical about the suitability of IPD mega-trial to analyze SE families.

IPD mega-trial has been also used to elicit moderators in 18% of the families [P10], [P11], [P28], [P26], [P33], [P30], [P15]. It is common in such families running experiments with different types of subjects, and then, using a certain "tag" to represent each type of subject as if it was the "Experiment" factor in an IPD stratified analysis. For example, experience (i.e., PhD, master, undergraduate) was used by Ricca et al. and Ceccato et al. [P26], [P14] to represent the "Experiment" factor in a IPD stratified model. In spite of its intuitiveness, such approach may be misleading: differences across experiments' results may not come from sole moderators, specially because other unknown variables (e.g., age, motivation, skills, treatment conformance, the materialization of some threats to validity in some experiments and not in others, etc. [59]), or confounding factors (e.g., if more than one change were purposefully introduced simultaneously across the experiments) may be also behind differences of results [60], [61].

Finally, we would also like to comment on the ability of families to increase statistical power [3], and the extent to which this is dependent upon the "severity" of the changes introduced across the experiments within families. In particular, if a certain dimension —or dimensions— are changed across experiments, this may introduce heterogeneity of results [62]. Under such circumstance, random-effects models should be preferred over fixed-effects models [21], [20]. This may have a noticeable impact on statistical power (as random-effects models are known to be more conservative than their fixed-effects counter-parts [12], [21], [63]). Thus, if changes are introduced across experiments, larger sample sizes, and a potentially larger number of experiments may be needed to reach to the same significance levels that would have been achieved with a fixed-effects model [21]. In turn, introducing many changes across experiments may be detrimental towards families' statistical power.

7 Good Practices Analyzing Families

We have elaborated a preliminary list of recommendations to analyze families after the common limitations that we observed with regard to joint data analysis. These recommendations follow:

- **Avoid Narrative synthesis if possible.** Despite its appealing, the application of Narrative synthesis is dangerous given current sample size limitations in SE experiments —and thus, the large variability of results expected. In addition, Narrative synthesis just provides a textual summary of results instead of quantitative summary (such as a p-value or effect size). Thus, the application of Narrative synthesis to aggregate experiments' results goes against the best practices of SE experimentation [17], [64], [65], [66], [7], where p-values and effect sizes are encouraged to be reported to summarize results.
- **Avoid IPD mega-trial if possible.** Even though intuitive, IPD mega-trial fails to account for heterogeneity of results across experiments. Unfortunately, heterogeneity may materialize if changes are introduced deliberately —or inadvertently— across experiments. IPD mega-trial also fails to account for the plausible correlation of participants' scores within experiments (as subjects may resemble more to each other within experiments than across experiments) and the existence of experiments with different sample sizes and missing-data [43], [42].
- **Avoid Aggregation of p-values if possible.** Mainly because with its use each experiment contributes identically to the overall conclusion independently of its quality, effect size or sample size, and because it does not provide any effect size, and thus, hinders the assessment of the relevance of results (e.g., how *large* is the joint effect? [21], [20]).
- **When data do not follow normality**. The robustness of traditional parametric statistical tests such as ANOVA or the t-test to departures from normality has been assessed over and over again even in smaller sample sizes than those typical in SE

experiments [67], [53], [55], [68], [69]. The robustness of traditionally used parametric statistical tests to departures from normality is even greater with sample sizes in the hundreds [54] —as those resulting when pooling together the raw-data of all the experiments within families. The superiority of parametric statistical tests over non-parametric tests such as the U-Mann Whitney or Wilcoxon has been widely acknowledged [70], [67], [53], specially with regard to their interpretability of results (as results can be provided in natural units —e.g., differences between means— instead of differences between mean ranks) and statistical flexibility (e.g., as multiple factors such as type of subject, experiment, etc. can be included). Even though we concede that defenders of both sides (i.e., parametric vs. non-parametric) can be found scattered along the literature [71], [72], [70], [67], [53], [55], [68], [69], we think that a good compromise is that followed by some families' authors (e.g., Laitenberger et al. [P22] or Pfahl et al. [P24]). In particular, in view that the Wilcoxon signed rank test and the dependent *t*-test provided similar results, Laitenberger et al. [P22] used the results of the *t*-test. If in spite of the robustness of parametric statistical tests to departures from normality it is still wanted to proceed with non-parametric tests, we recommend to use one of the following approaches to aggregate results:

  - **AD.** Use either (1) *non-parametric effect sizes* (e.g., Cliff's delta or Probability of Superiority, etc. [39]) or (2) *bootstrap* to elicit the standard error of the parametric effect size selected (e.g., Cohen's d or Hedge's g) [73].
  - **IPD stratified.** Use either (1) *a permutation test* (by following a similar approach to that followed by Ricca et al. [P16]); (2) a *Generalized Linear Model* to accommodate the distribution of the data (e.g., by means of logistic regressions [74], [75]); (3) use *bootstrap* [76], [77].

- **Check AD and IPD statistical assumptions.** If sample sizes are small and data are not normal or have different variances, parametric effect sizes (such as Cohen's d or Hedge's g [21]) may be unrealiable [39], [38], [57]. Similarly, if an IPD stratified statistical model is used (e.g., ANOVA), check its assumptions (e.g., normality and equality of variances) or consult specialized literature on the topic to investigate about their robustness to departures from statistical assumptions [72], [70], [53], [55], [68], [69].
- **Acknowledge heterogeneity of results** when providing joint results. If changes are introduced across experiments (e.g., different populations, experimental designs, etc.), such changes may impact the effect size being estimated in each experiment [21]. As others did before, we recommend to apply by default random-effects models over fixed-effects models as the prior reduce to the later if no heterogeneity of results has materialized [21].
- **Plan ahead the analysis procedure to follow.** We encourage researchers conducting families to plan ahead *at least* the pre-processing steps that are going to be undertaken (e.g., which procedure is going to be followed to remove outliers and graph data?), the statistical tests that are going to be used (e.g., are researchers interested in differences between means (and thus parametric tests?), and the aggregation techniques to be used (e.g., are researchers interested in providing intuitive visualizations of results such as with AD? are researchers interested in evaluating results in natural units, such as with IPD?). If the application of such procedures, tests or aggregation techniques was unfeasible after conducting the experiments within the family (e.g., due to unexpected data-losses, experimental restrictions that forced design changes, etc.), such rationale should be discussed before providing joint results or eliciting moderators.

8 Good Practices Reporting Families

While conducting this SMS we found that assessing the appropriateness of the aggregation technique/s used within families was not straightforward. This was in part because the articles missed some relevant information to judge this (e.g., are the raw-data of all the experiments available?, are the response variables' operationalizations and scales identical across the experiments?), and in part because the raw-data of the families were not made public —and thus, re-analyzing the family with all the techniques to check the suitability of the technique/s used was unfeasible. In view of this, we have extracted a series of elements that we think are relevant for assessing the reliability of results and the suitability of the aggregation technique/s applied. We suggest to report at least the following elements in research articles reporting families:

- **Raw-data availability.** Report whether the raw-data are available at the time of analyzing the family. If the raw-data are available, techniques such as IPD stratified or AD may be preferred over the other techniques to provide joint results or elicit moderators.
- **Response variable changes.** Report whether the same response variable operationalization is used across experiments, as if this is the case, there is no need to compute standardized effect sizes (e.g., Cohen's d) to conduct AD. In this situation, *unstandardized* units can be used instead to increase the interpretability of results with AD [21], [40]. IPD stratified can be safely used either [44].
- **Experiments' sample sizes.** Report the number of subjects that participated in each experiment, since this information is useful for computing various effect sizes [21], and arguing about the generalizability of results.
- **The relationship between the participants across experiments.** Report if different subjects participated across the experiments, as otherwise, if they are the same, and this was not accounted for by the aggregation technique used, joint results may be misleading.
- **The technique/s used to analyze the family.** Motivate the selection of the aggregation technique/s

used, and why such technique/s were selected instead of others given the characteristics of the family.
- **The elements that changed across experiments.** Make explicit the changes introduced across the experiments, so it is possible assessing the suitability of the aggregation technique/s applied.

Finally, we have observed that the raw-data (i.e., a spread-sheet with the response variables and the assignment of subjects to treatments across the experiments) of the families were not accessible in most families of experiments —despite that 6 out of 39 families claimed to have published them [P14], [P35][P8], [P17], [P19], [P1]. This precludes re-analysis with other —perhaps more suitable— analysis techniques to double-check the robustness of results to different aggregation techniques, and eventually, hinders the reproducibility of results. Thus, if possible, **we encourage researchers reporting families to make their raw-data available**.

9 Threats to Validity

In this section we discuss the main threats to validity of our SMS according to Petersen et al.'s guidelines [15].

When conducting a secondary study, the search terms should identify as many relevant papers as possible to provide an accurate overview of the topic and its structure [15]. We piloted different terms and search strings along the way to reduce the risk of missing relevant primary studies. Due to the inconsistency in the terms used to refer to families of experiments in SE, we used synonyms of the term obtained from a reference set of five articles [P38], [P14], [P8], [P2], [P36]. This helped us to broaden the results of our SMS and increase the reliability of the findings.

We had to trade off the complexity of the search string (in order not to be too restrictive) against the looseness of the terms (due to the acute reduction in accuracy of the search). In particular, the words "experiment" and "aggregation", "group of studies" or "series of experiments" appeared in many different disciplines unrelated to the scope of our research, which added a large amount of noise to our results. In turn, we had no option but restrict the search space to relevant venues on SE experimentation. Still, we do not think this had a major impact on results, as the techniques that we found in our primary studies were similar to those used in other disciplines such as medicine or pharmacology [20], [21], [62].

In order to increase the precision of the results, we confined the search space to the venues surveyed in a well-known secondary study on SE experiments [78]. We acknowledge the possibility of publication bias on results (as we restricted our search to well-known venues for publishing empirical research). Even though this might have conditioned our results, we complemented the search with a backward snowballing process [17].

After identifying relevant articles, the selection of primary studies within the scope of research is crucial for achieving relevant conclusions. The protocol included the definition of explicit exclusion criteria to minimize subjectivity and prevent the omission of valid articles. Exclusion criteria were applied by two researchers in order to select valid primary studies. In the event of disagreement, a third researcher helped with the inclusion or exclusion of the article.

We acknowledge that other researchers may have gathered a different list of references for portraying the advantages and disadvantages of the aggregation techniques that we identified. Unfortunately, it was unfeasible for us to gather references systematically due the enormous list of references prompted in online databases with the terms "aggregation", "experiments", "data" or "effect sizes". However, with the aim of minimizing this shortcoming we tried to recur to well-known references in mature experimental disciplines such as medicine and pharmacology [62], [79], seminal works on vote-counting and narrative synthesis [80], [81], and meta-analysis and reproducibility of results [21], [56], [14] to make our findings representative of the current state of knowledge.

Even though we used the same naming conventions used in medicine to categorize the aggregation techniques used in SE (e.g., AD, IPD, etc.), this does not imply that those techniques were just applied in medicine: the same aggregation techniques —albeit with different names— have been also applied in other disciplines such as econometrics, ecology, biology, or the social sciences, just to name a few [21], [20], [82]. Instead, in this article we just recurred to the naming conventions used in medicine as a way of categorizing the techniques applied in SE into broad groups, and as SE experimental research has previously imported from medicine much of the data analysis procedures used to analyze individual experiments [83], [64], and the research methods followed to come up with empirical evidence (e.g., the procedures for conducting SLRs [7] or Evidence Based Software Engineering [66] among others).

Finally, in only 7% of the identified families of experiments the raw-data were accessible. This hindered re-analysis, and thus, a formal assessment on the suitability of the techniques applied to analyze both individual experiments and the families as whole. As a result, providing fine tailored recommendations to each family on the suitability of the techniques applied was unfeasible.

10 Conclusion

Families of experiments (i.e., groups of interrelated experiments with the same goal [1]) are on the rise in SE. Characteristics, implications and particularities of families over other study types (such as groups of experiments gathered by means of SLRs) are to be defined yet in SE. Fine-tuning the term family of experiments may help SE researchers to apply consistent techniques to analyze them.

We define a family of experiments as a group of replications where access to the raw-data is guaranteed, the settings of the experiments are known by researchers, and at least three experiments are conducted to assess the performance of at least two different technologies on the same response variable. Several techniques have been used to aggregate experiments' results within families in SE: Narrative synthesis, AD, IPD mega-trial, IPD stratified and Aggregation of p-values. The rationale to select aggregation techniques within families is rarely discussed in research articles. This may impact the reliability of joint results and undermine their potential to elicit moderators.

46% of the families used Narrative synthesis to provide joint conclusions or study moderators. SE experiments' small sample sizes may impact the reliability of such families' results. 38% of the families used AD. Some of the families applying AD resorted to parametric effect sizes (e.g., Cohen's d or Hedge's g) despite acknowledging small sample sizes and that data did not follow normality. This may have impacted the reliability of such families' findings. 33% of the identified families used IPD mega-trial. Data unbalance, and the multiple changes commonly introduced across experiments within families may have lead to unreliable results. 15% of the families used IPD stratified to aggregate experimental results following similar procedures to those followed in medicine [20]. Finally, Aggregation of p-values was used in just 7% of the families. Aggregation of p-values weights identically each experiment towards the joint result and does not provide a joint effect size. This may limit the interpretability of results in such families.

The aggregation technique/s used within families may impact the reliability of joint results. Transparent and reliable aggregation techniques such as AD or IPD stratified seem suitable to analyze SE families. However, the reliability of AD and IPD stratified comes at a cost: their statistical assumptions need to be checked before interpreting results. We recommend to minimize the use of Narrative synthesis as it does not provide a quantitative summary of results, involves subjective judgment, and does not take advantage of the raw-data when providing joint results or eliciting moderators. We recommend to minimize the use of Aggregation of p-values as it does not provide any way of assessing the relevance of results, and in its basic form, weights identically each experiment towards the joint conclusion —regardless of their effect sizes, sample sizes, quality or experimental designs. IPD mega-trial should be avoided as it may provide biased results if data is unbalanced across or within experiments or if subjects resemble more to each other within-experiments than across experiments.

Finally, increasing numbers of experiments are being included within families (an average of five per family). This could lead SE to a next stage of maturity. However, reporting deficiencies and the blind application of unsuitable aggregation techniques could limit the reliability of joint results. Researchers should assess the suitability of the aggregation techniques used, and provide a rationale on why such techniques were used to analyze their respective families. Finally, we urge researchers conducting families of experiments in SE to plan ahead their data analyses, and improve reporting transparency with respect to data, statistical assumptions and data analysis techniques.

ACKNOWLEDGMENTS

This research was developed with the support of the Spanish Ministry of Science and Innovation project TIN2014-60490-P. We would like to thank the anonymous reviewers who aided to polish this research article.

REFERENCES

[1] V. R. Basili, F. Shull, and F. Lanubile, "Building knowledge through families of experiments," *Software Engineering, IEEE Transactions on*, vol. 25, no. 4, pp. 456–473, 1999.

[2] O. S. Gomez, N. Juristo, and S. Vegas, "Understanding replication of experiments in software engineering: A classification," *Information and Software Technology*, vol. 56, no. 8, pp. 1033–1048, 2014.

[3] E. Fernández, O. Dieste, P. M. Pesado, and R. García Martínez, "The importance of using empirical evidence in software engineering," 2011.

[4] J. C. Carver, N. Juristo, M. T. Baldassarre, and S. Vegas, "Introduction to special issue on replications of software engineering experiments," *Empirical Softw. Engg.*, vol. 19, no. 2, pp. 267–276, Apr. 2014.

[5] M. Ciolkowski, F. Shull, and S. Biffl, "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*, 2002.

[6] N. Juristo and S. Vegas, "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*. IEEE Computer Society, 2009, pp. 356–366.

[7] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.

[8] N. Juristo and S. Vegas, "The role of non-exact replications in software engineering experiments," *Empirical Software Engineering*, vol. 16, no. 3, pp. 295–324, 2011.

[9] F. J. Shull, J. C. Carver, S. Vegas, and N. Juristo, "The role of replications in empirical software engineering," *Empirical Software Engineering*, vol. 13, no. 2, pp. 211–218, 2008.

[10] B. Kitchenham, "The role of replications in empirical software engineeringa word of warning," *Empirical Software Engineering*, vol. 13, no. 2, pp. 219–221, 2008.

[11] H. Cooper and E. A. Patall, "The relative benefits of meta-analysis conducted with individual participant data versus aggregated data." *Psychological methods*, vol. 14, no. 2, p. 165, 2009.

[12] J. Lau, J. P. Ioannidis, and C. H. Schmid, "Summing up evidence: one answer is not always enough," *The lancet*, vol. 351, no. 9096, pp. 123–127, 1998.

[13] A. Haidich, "Meta-analysis in medical research," *Hippokratia*, vol. 14, no. 1, pp. 29–37, 2011.

[14] J. Ioannidis, N. Patsopoulos, and H. Rothstein, "Research methodology: reasons or excuses for avoiding meta-analysis in forest plots," *BMJ: British Medical Journal*, vol. 336, no. 7658, pp. 1413–1415, 2008.

[15] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th international conference on evaluation and assessment in software engineering*, vol. 17, no. 1. sn, 2008, pp. 1–10.

[16] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[17] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2014, p. 38.

[18] F. Shull, M. G. Mendoncça, V. Basili, J. Carver, J. C. Maldonado, S. Fabbri, G. H. Travassos, and M. C. Ferreira, "Knowledge-sharing issues in experimental software engineering," *Empirical Software Engineering*, vol. 9, no. 1-2, pp. 111–137, 2004.

[19] J. C. Carver, N. Juristo, M. T. Baldassarre, and S. Vegas, "Replications of software engineering experiments," 2014.

[20] A. Whitehead, *Meta-analysis of controlled clinical trials*. John Wiley & Sons, 2002, vol. 7.

[21] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. John Wiley & Sons, 2011.

[22] D. Jackson, R. Riley, and I. R. White, "Multivariate meta-analysis: Potential and promise," *Statistics in medicine*, vol. 30, no. 20, pp. 2481–2498, 2011.

[23] C. Anello, R. T. ONeill, and S. Dubey, "Multicentre trials: a us regulatory perspective," *Statistical Methods in Medical Research*, vol. 14, no. 3, pp. 303–318, 2005.

[24] J. A. Lewis, "Statistical principles for clinical trials (ich e9): an introductory note on an international guideline," *Statistics in medicine*, vol. 18, no. 15, pp. 1903–1942, 1999.

[25] L. A. Stewart, M. Clarke, M. Rovers, R. D. Riley, M. Simmonds, G. Stewart, and J. F. Tierney, "Preferred reporting items for a systematic review and meta-analysis of individual participant data: the prisma-ipd statement," *Jama*, vol. 313, no. 16, pp. 1657–1665, 2015.

[26] L. Bero and D. Rennie, "The cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care," *Jama*, vol. 274, no. 24, pp. 1935–1938, 1995.

[27] A. Field, *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[28] M. C. Simmonds, J. P. Higginsa, L. A. Stewartb, J. F. Tierneyb, M. J. Clarke, and S. G. Thompson, "Meta-analysis of individual patient data from randomized trials: a review of methods used in practice," *Clinical Trials*, vol. 2, no. 3, pp. 209–217, 2005.

[29] J. Popay, H. Roberts, A. Sowden, M. Petticrew, L. Arai, M. Rodgers, N. Britten, K. Roen, and S. Duffy, "Guidance on the conduct of narrative synthesis in systematic reviews," *A product from the ESRC methods programme Version*, vol. 1, p. b92, 2006.

[30] M. Rodgers, A. Sowden, M. Petticrew, L. Arai, H. Roberts, N. Britten, and J. Popay, "Testing methodological guidance on the conduct of narrative synthesis in systematic reviews: effectiveness of interventions to promote smoke alarm ownership and function," *Evaluation*, vol. 15, no. 1, pp. 49–73, 2009.

[31] R. Rosenthal, H. Cooper, and L. Hedges, "Parametric measures of effect size," *The handbook of research synthesis*, pp. 231–244, 1994.

[32] S. Nakagawa, "A farewell to bonferroni: the problems of low statistical power and publication bias," *Behavioral Ecology*, vol. 15, no. 6, pp. 1044–1045, 2004.

[33] R. Coe, "It's the effect size, stupid: What effect size is and why it is important," 2002.

[34] J. Cohen, "Statistical power analysis for the behavioral sciences lawrence earlbaum associates," *Hillsdale, NJ*, pp. 20–26, 1988.

[35] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. Sjøberg, "A systematic review of effect size in software engineering experiments," *Information and Software Technology*, vol. 49, no. 11, pp. 1073–1086, 2007.

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

[37] M. R. Hess and J. D. Kromrey, "Robust confidence intervals for effect sizes: A comparative study of cohensd and cliffs delta under non-normality and heterogeneous variances," in *Annual Meeting of the American Educational Research Association*, 2004, pp. 12–16.

[38] C.-Y. J. Peng and L.-T. Chen, "Beyond cohen's d: Alternative effect size measures for between-subject designs," *The Journal of Experimental Education*, vol. 82, no. 1, pp. 22–50, 2014.

[39] C. O. Fritz, P. E. Morris, and J. J. Richler, "Effect size estimates: current use, calculations, and interpretation." *Journal of Experimental Psychology: General*, vol. 141, no. 1, p. 2, 2012.

[40] L. A. Stewart and J. F. Tierney, "To ipd or not to ipd? advantages and disadvantages of systematic reviews using individual patient data," *Evaluation & the health professions*, vol. 25, no. 1, pp. 76–97, 2002.

[41] H. Quené and H. Van den Bergh, "On multi-level modeling of data from repeated measures designs: A tutorial," *Speech Communication*, vol. 43, no. 1-2, pp. 103–121, 2004.

[42] G. Abo-Zaid, B. Guo, J. J. Deeks, T. P. Debray, E. W. Steyerberg, K. G. Moons, and R. D. Riley, "Individual participant data meta-analyses should not ignore clustering," *Journal of clinical epidemiology*, vol. 66, no. 8, pp. 865–873, 2013.

[43] H. C. Kraemer, "Pitfalls of multisite randomized clinical trials of efficacy and effectiveness," *Schizophrenia Bulletin*, vol. 26, no. 3, pp. 533–541, 2000.

[44] T. Debray, K. G. Moons, G. Valkenhoef, O. Efthimiou, N. Hummel, R. H. Groenwold, and J. B. Reitsma, "Get real in individual participant data (ipd) meta-analysis: a review of the methodology," *Research synthesis methods*, vol. 6, no. 4, pp. 293–309, 2015.

[45] G. B. Stewart, D. G. Altman, L. M. Askie, L. Duley, M. C. Simmonds, and L. A. Stewart, "Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice," *PloS one*, vol. 7, no. 10, p. e46042, 2012.

[46] A. Birnbaum, "Combining independent tests of significance," *Journal of the American Statistical Association*, vol. 49, no. 267, pp. 559–574, 1954.

[47] G. Leandro, *Meta-analysis in Medical Research: The handbook for the understanding and practice of meta-analysis*. John Wiley & Sons, 2008.

[48] J. Miller, "Applying meta-analytical procedures to software engineering experiments," *Journal of Systems and Software*, vol. 54, no. 1, pp. 29–39, 2000.

[49] H. Brown and R. Prescott, *Applied mixed models in medicine*. John Wiley & Sons, 2014.

[50] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, p. 365, 2013.

[51] T. Dybå, V. B. Kampenes, and D. I. Sjøberg, "A systematic review of statistical power in software engineering experiments," *Information and Software Technology*, vol. 48, no. 8, pp. 745–755, 2006.

[52] S. E. Maxwell, M. Y. Lau, and G. S. Howard, "Is psychology suffering from a replication crisis? what does failure to replicate really mean?" *American Psychologist*, vol. 70, no. 6, p. 487, 2015.

[53] A. J. Vickers, "Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data," *BMC medical research methodology*, vol. 5, no. 1, p. 35, 2005.

[54] T. Lumley, P. Diehr, S. Emerson, and L. Chen, "The importance of the normality assumption in large public health data sets," *Annual review of public health*, vol. 23, no. 1, pp. 151–169, 2002.

[55] M. W. Fagerland, "t-tests, non-parametric tests, and large studiesa paradox of statistical practice?" *BMC Medical Research Methodology*, vol. 12, no. 1, p. 78, 2012.

[56] M. Egger, G. Davey-Smith, and D. Altman, *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons, 2008.

[57] G. Macbeth, E. Razumiejczyk, and R. D. Ledesma, "Cliff's delta calculator: A non-parametric effect size program for two groups of observations," *Universitas Psychologica*, vol. 10, no. 2, pp. 545–555, 2011.

[58] O. Dieste, G. Raura, P. Rodríguez *et al.*, "Professionals are not superman: failures beyond motivation in software experiments," in *Conducting Empirical Studies in Industry (CESI), 2017 IEEE/ACM 5th International Workshop on*. IEEE, 2017, pp. 27–32.

[59] O. Dieste, A. M. Aranda, F. Uyaguari, B. Turhan, A. Tosun, D. Fucci, M. Oivo, and N. Juristo, "Empirical evaluation of the effects of experience on code quality and programmer productivity: an exploratory study," *Empirical Software Engineering*, vol. 22, no. 5, pp. 2457–2542, 2017.

[60] R. H. Groenwold, A. R. T. Donders, G. J. van der Heijden, A. W. Hoes, and M. M. Rovers, "Confounding of subgroup analyses in randomized data," *Archives of internal medicine*, vol. 169, no. 16, pp. 1532–1534, 2009.

[61] B. Dijkman, B. Kooistra, and M. Bhandari, "How to work with a subgroup analysis," *Canadian Journal of Surgery*, vol. 52, no. 6, p. 515, 2009.

[62] J. P. Higgins and S. Green, *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2011, vol. 4.

[63] V. Huta, "When to use hierarchical linear modeling," *Quant Methods Psychol*, vol. 10, no. 1, pp. 13–28, 2014.

[64] N. Juristo and A. M. Moreno, *Basics of software engineering experimentation*. Springer Science & Business Media, 2011.

[65] T. Dyba, B. A. Kitchenham, and M. Jorgensen, "Evidence-based software engineering for practitioners," *IEEE software*, vol. 22, no. 1, pp. 58–65, 2005.

[66] B. A. Kitchenham, T. Dyba, and M. Jorgensen, "Evidence-based software engineering," in *Proceedings of the 26th international conference on software engineering*. IEEE Computer Society, 2004, pp. 273–281.

[67] J. C. De Winter, "Using the student's t-test with extremely small sample sizes." *Practical Assessment, Research & Evaluation*, vol. 18, no. 10, 2013.

[68] E. Schmider, M. Ziegler, E. Danay, L. Beyer, and M. Bühner, "Is it really robust?" *Methodology*, 2010.

[69] U. Wadgave and M. R. Khairnar, "Parametric tests for likert scale: For and against," *Asian journal of psychiatry*, vol. 24, pp. 67–68, 2016.

[70] G. Norman, "Likert scales, levels of measurement and the laws of statistics," *Advances in health sciences education*, vol. 15, no. 5, pp. 625–632, 2010.

[71] A. Arcuri and L. Briand, "A practical guide for using statistical tests to assess randomized algorithms in software engineering," in *Software Engineering (ICSE), 2011 33rd International Conference on*. IEEE, 2011, pp. 1–10.

[72] E. Whitley and J. Ball, "Statistics review 6: Nonparametric methods," *Critical care*, vol. 6, no. 6, p. 509, 2002.

[73] S. Nakagawa and I. C. Cuthill, "Effect size, confidence interval and statistical significance: a practical guide for biologists," *Biological reviews*, vol. 82, no. 4, pp. 591–605, 2007.

[74] C. E. McCulloch and J. M. Neuhaus, *Generalized linear mixed models*. Wiley Online Library, 2001.

[75] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White, "Generalized linear mixed models: a practical guide for ecology and evolution," *Trends in ecology & evolution*, vol. 24, no. 3, pp. 127–135, 2009.

[76] S. Ren, H. Lai, W. Tong, M. Aminzadeh, X. Hou, and S. Lai, "Nonparametric bootstrapping for hierarchical data," *Journal of Applied Statistics*, vol. 37, no. 9, pp. 1487–1498, 2010.

[77] T. C. Hesterberg, "What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum," *The American Statistician*, vol. 69, no. 4, pp. 371–386, 2015.

[78] D. I. Sjøberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N.-K. Liborg, and A. C. Rekdal, "A survey of controlled experiments in software engineering," *Software Engineering, IEEE Transactions on*, vol. 31, no. 9, pp. 733–753, 2005.

[79] D. B. Petitti *et al.*, *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine*. OUP USA, 2000, no. 31.

[80] L. V. Hedges and I. Olkin, "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*, 1979.

[81] ——, "Vote-counting methods in research synthesis." *Psychological bulletin*, vol. 88, no. 2, p. 359, 1980.

[82] F. Harrison, "Getting started with meta-analysis," *Methods in Ecology and Evolution*, vol. 2, no. 1, pp. 1–10, 2011.

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

PRIMARY STUDIES

[P1] G. Scanniello, C. Gravino, G. Tortora, M. Genero, M. Risi, J. A. Cruz-Lemus, and G. Dodero, "Studying the effect of uml-based models on source-code comprehensibility: Results from a long-term investigation," in *International Conference on Product-Focused Software Process Improvement*. Springer, 2015, pp. 311–327.

[P2] N. Juristo, S. Vegas, M. Solari, S. Abrahao, and I. Ramos, "Comparing the effectiveness of equivalence partitioning, branch testing and code reading by stepwise abstraction applied by subjects," in *2012 IEEE Fifth International Conference on Software Testing, Verification and Validation*. IEEE, 2012, pp. 330–339.

[P3] J. L. Krein, L. Prechelt, N. Juristo, A. Nanthaamornphong, J. C. Carver, S. Vegas, C. D. Knutson, K. D. Seppi, and D. L. Eggett, "A multi-site joint replication of a design patterns experiment using moderator variables to generalize across contexts," *IEEE Transactions on Software Engineering*, vol. 42, no. 4, pp. 302–321, 2016.

[P4] G. Canfora, F. García, M. Piattini, F. Ruiz, and C. A. Visaggio, "A family of experiments to validate metrics for software process models," *Journal of Systems and Software*, vol. 77, no. 2, pp. 113–129, 2005.

[P5] M. E. Manso, J. A. Cruz-Lemus, M. Genero, and M. Piattini, "Empirical validation of measures for uml class diagrams: A meta-analysis study," in *International Conference on Model Driven Engineering Languages and Systems*. Springer, 2008, pp. 303–313.

[P6] J. A. Cruz-Lemus, M. Genero, M. E. Manso, S. Morasca, and M. Piattini, "Assessing the understandability of uml statechart diagrams with composite statesa family of empirical studies," *Empirical Software Engineering*, vol. 14, no. 6, pp. 685–719, 2009.

[P7] E. Figueiredo, A. Garcia, M. Maia, G. Ferreira, C. Nunes, and J. Whittle, "On the impact of crosscutting concern projection on code measurement," in *Proceedings of the tenth international conference on Aspect-oriented software development*. ACM, 2011, pp. 81–92.

[P8] S. Abrahao, C. Gravino, E. Insfran, G. Scanniello, and G. Tortora, "Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: Results from a family of five experiments," *IEEE Transactions on Software Engineering*, vol. 39, no. 3, pp. 327–342, 2013.

[P9] N. Salleh, E. Mendes, and J. Grundy, "Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments," *Empirical Software Engineering*, vol. 19, no. 3, pp. 714–752, 2014.

[P10] M. Ceccato, M. Di Penta, P. Falcarin, F. Ricca, M. Torchiano, and P. Tonella, "A family of experiments to assess the effectiveness and efficiency of source code obfuscation techniques," *Empirical Software Engineering*, vol. 19, no. 4, pp. 1040–1074, 2014.

[P11] M. Ceccato, A. Marchetto, L. Mariani, C. D. Nguyen, and P. Tonella, "Do automatically generated test cases make debugging easier? an experimental assessment of debugging effectiveness and efficiency," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 25, no. 1, p. 5, 2015.

[P12] M. Staron, L. Kuzniarz, and C. Wohlin, "Empirical assessment of using stereotypes to improve comprehension of uml models: A set of experiments," *Journal of Systems and Software*, vol. 79, no. 5, pp. 727–742, 2006.

[P13] L. Muñoz, J.-N. Mazón, and J. Trujillo, "A family of experiments to validate measures for uml activity diagrams of etl processes in data warehouses," *Information and Software Technology*, vol. 52, no. 11, pp. 1188–1203, 2010.

[P14] F. Ricca, M. Di Penta, M. Torchiano, P. Tonella, and M. Ceccato, "How developers' experience and ability influence web application comprehension tasks supported by uml stereotypes: A series of four experiments," *IEEE Transactions on Software Engineering*, vol. 36, no. 1, pp. 96–118, 2010.

[P15] S. Mouchawrab, L. C. Briand, Y. Labiche, and M. Di Penta, "Assessing, comparing, and combining state machine-based testing and structural testing: a series of experiments," *IEEE Transactions on Software Engineering*, vol. 37, no. 2, pp. 161–187, 2011.

[P16] F. Ricca, G. Scanniello, M. Torchiano, G. Reggio, and E. Astesiano, "Assessing the effect of screen mockups on the comprehension of functional requirements," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 24, no. 1, p. 1, 2014.

[P17] G. Scanniello, C. Gravino, M. Genero, J. Cruz-Lemus, and G. Tortora, "On the impact of uml analysis models on source-code comprehensibility and modifiability," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 23, no. 2, p. 13, 2014.

[P18] J. Gonzalez-Huerta, E. Insfran, S. Abrahão, and G. Scanniello, "Validating a model-driven software architecture evaluation and improvement method: A family of experiments," *Information and Software Technology*, vol. 57, pp. 405–429, 2015.

[P19] G. Scanniello, C. Gravino, M. Risi, G. Tortora, and G. Dodero, "Documenting design-pattern instances: a family of experiments on source-code comprehensibility," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 24, no. 3, p. 14, 2015.

[P20] A. M. Fernández-Sáez, M. Genero, D. Caivano, and M. R. Chaudron, "Does the level of detail of uml diagrams affect the maintainability of source code?: a family of experiments," *Empirical Software Engineering*, vol. 21, no. 1, pp. 212–259, 2016.

[P21] A. Porter and L. Votta, "Comparing detection methods for software requirements inspections: A replication using professional subjects," *Empirical software engineering*, vol. 3, no. 4, pp. 355–379, 1998.

[P22] O. Laitenberger, K. El Emam, and T. G. Harbich, "An internally replicated quasi-experimental comparison of checklist and perspective based reading of code documents," *IEEE Transactions on Software Engineering*, vol. 27, no. 5, pp. 387–421, 2001.

[P23] B. George and L. Williams, "A structured experiment of test-driven development," *Information and software Technology*, vol. 46, no. 5, pp. 337–342, 2004.

[P24] D. Pfahl, O. Laitenberger, G. Ruhe, J. Dorsch, and T. Krivobokova, "Evaluating the learning effectiveness of using simulations in software project management education: results from a twice replicated experiment," *Information and software technology*, vol. 46, no. 2, pp. 127–147, 2004.

[P25] L. Reynoso, M. Genero, M. Piattini, and E. Manso, "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)*. IEEE, 2005, pp. 10–pp.

[P26] F. Ricca, M. Di Penta, M. Torchiano, P. Tonella, M. Ceccato, and C. A. Visaggio, "Are fit tables really talking?" in *2008 ACM/IEEE 30th International Conference on Software Engineering*. IEEE, 2008, pp. 361–370.

[P27] C. G. Von Wangenheim, M. Thiry, and D. Kochanski, "Empirical evaluation of an educational game on software measurement," *Empirical Software Engineering*, vol. 14, no. 4, pp. 418–452, 2009.

[P28] J. A. Cruz-Lemus, M. Genero, D. Caivano, S. Abrahão, E. Insfrán, and J. A. Carsí, "Assessing the influence of stereotypes on the comprehension of uml sequence diagrams: A family of experiments," *Information and Software Technology*, vol. 53, no. 12, pp. 1391–1403, 2011.

[P29] M. Jørgensen, "Contrasting ideal and realistic conditions as a means to improve judgment-based software development effort

estimation," *Information and Software Technology*, vol. 53, no. 12, pp. 1382–1390, 2011.
[P30] M. A. Teruel, E. Navarro, V. López-Jaquero, F. Montero, J. Jaen, and P. González, "Analyzing the understandability of requirements engineering languages for cscw systems: A family of experiments," *Information and Software Technology*, vol. 54, no. 11, pp. 1215–1228, 2012.
[P31] P. Runeson, A. Stefik, A. Andrews, S. Gronblom, I. Porres, and S. Siebert, "A comparative analysis of three replicated experiments comparing inspection and unit testing," in *Replication in Empirical Software Engineering Research (RESER), 2011 Second International Workshop on*. IEEE, 2011, pp. 35–42.
[P32] T. Kosar, M. Mernik, and J. C. Carver, "Program comprehension of domain-specific and general-purpose languages: comparison using a family of experiments," *Empirical software engineering*, vol. 17, no. 3, pp. 276–304, 2012.
[P33] I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. Susi, "Comparing the comprehensibility of requirements models expressed in use case and tropos: Results from a family of experiments," *Information and Software Technology*, vol. 55, no. 10, pp. 1823–1843, 2013.
[P34] C. R. L. Neto, I. do Carmo Machado, V. C. Garcia, and E. S. de Almeida, "Analyzing the effectiveness of a system testing tool for software product line engineering (s)," in *SEKE*, 2013.
[P35] A. Fernandez, S. Abrahão, and E. Insfran, "Empirical validation of a usability inspection method for model-driven web development," *Journal of Systems and Software*, vol. 86, no. 1, pp. 161–186, 2013.
[P36] P. Runeson, A. Stefik, and A. Andrews, "Variation factors in the design and analysis of replicated controlled experiments," *Empirical Software Engineering*, vol. 19, no. 6, pp. 1781–1808, 2014.
[P37] S. Ali, T. Yue, and I. Rubab, "Assessing the modeling of aspect state machines for testing from the perspective of modelers," in *2014 14th International Conference on Quality Software*. IEEE, 2014, pp. 234–239.
[P38] S. T. Acuña, M. N. Gómez, J. E. Hannay, N. Juristo, and D. Pfahl, "Are team personality and climate related to satisfaction and software quality? aggregating results from a twice replicated experiment," *Information and Software Technology*, vol. 57, pp. 141–156, 2015.
[P39] A. M. Fernández-Sáez, M. Genero, M. R. Chaudron, D. Caivano, and I. Ramos, "Are forward designed or reverse-engineered uml diagrams more helpful for code maintenance?: A family of experiments," *Information and Software Technology*, vol. 57, pp. 644–663, 2015.

Adrian Santos received his MSc in Software and Systems and MSc in Software Project Management at the Technical University of Madrid, Spain, and his MSc in IT Auditing, Security and Government at the Autonomous University of Madrid, Spain. He is a PhD student at the University of Oulu, Finland. His research interests include empirical software engineering, agile methodologies, statistical analysis and data mining techniques. He is a member of the American Statistical Association (ASA) and the International Society for Bayesian Analysis (ISBA). For more information and details, please visit http://www.adriansantosparrilla.com

Omar Gómez received his BSc in Computer Engineering from the University of Guadalajara, his MSc in Software Engineering from the Center for Mathematical Research (CIMAT) and his PhD in Software and Systems from the Technical University of Madrid. He was with the University of Guadalajara (as adjunct assistant professor), the Autonomous University of Yucatan (as adjunct associate professor), the University of Oulu (as a postdoctoral research fellow) and the Technical School of Chimborazo (as a Prometeo-Senescyt researcher). He is currently adjunct associate professor with the Technical School of Chimborazo. His main research interest is software engineering experimentation.

Natalia Juristo has been full professor of software engineering with the School of Computer Engineering at the Technical University of Madrid (UPM) since 1997. She was awarded a FiDiPro (Finland Distinguished Professor Program) professorship at the University of Oulu, starting in January 2013. She was the Director of the MSc in Software Engineering from 1992 to 2002 and coordinator of the Erasmus Mundus European Master on SE (with the participation of UPM, University of Bolzano, University of Kaiserslautern and Blege Institute of Technology) from 2006 to 2012. Her main research interests are experimental software engineering, requirements and testing. Back in 2001 she co-authored the book Basics of Software Engineering Experimentation (Kluwer). Natalia is a member of the editorial boards of IEEE Transactions on SE and Empirical SE, and Software: Testing, Verification and Reliability. She has served on several congress program committees (ICSE, RE, REFSQ, ESEM, ISESE, etc.), and has been congress program chair (EASE13, ISESE04 and SEKE97), as well as general chair (ESEM07, SNPD02 and SEKE01). Natalia was co-chair of ICSE Technical Briefings 2015 and co-chair of the Software Engineering in Practice (SEIP) track at ICSE 2017. She began her career as a developer at the European Space Agency (Rome) and the European Center for Nuclear Research (Geneva). She was a resident affiliate at the Software Engineering Institute in Pittsburgh in 1992. In 2009, Natalia was awarded an honorary doctorate by Blekinge Institute of Technology in Sweden. For more information and details, please visit http://grise.upm.es/miembros/natalia/