Reliability of Systematic Literatures Reviews on Test-Driven Development

Fernando Uyaguari[1], **Silvia T. Acuña**[2], **John W. Castro**[3*], **Oscar Dieste**[4], **Natalia Juristo**[4]

[1] Instituto Superior Tecnológico Wissen, Av. 10 de Agosto s/n y Jose Ma. Sanchez, 010107, Cuenca, Ecuador

[2] Departamento de Ingeniería Informática, Universidad Autónoma de Madrid,

Calle Francisco Tomás y Valiente 11, 28049 Madrid, Spain

[3] Departamento de Ingeniería Informática y Ciencias de la Computación, Universidad de Atacama,

Avenida Copayapu 485, 1530000 Copiapó, Chile

[4] Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, C. de los Ciruelos, 28660 Boadilla del Monte, Spain

[1]fernando.uyaguari@wissen.edu.ec, [2]silvia.acunna@uam.es, [3]john.castro@uda.cl, [4]odieste@fi.upm.es, [4]natalia@fi.upm.es

ABSTRACT

Context: Test-driven development (TDD) is a software development technique studied empirically over the last few decades. There are several systematic literature reviews (SLRs) on TDD. The reliability of these studies should not be taken for granted because SLRs are highly dependent on the context and researcher decision-making.
Objective: This study determines, analyses, and synthesizes the limited overlap between SLRs on TDD and the influence of their conclusions and results on the code quality and developer productivity response variables.
Method: A tertiary study was conducted to source SLRs on TDD from the scientific literature, and the primary studies referenced in each SLR were analysed. We compared SLRs with similar objectives, SLRs with similar response variables, and all SLRs. We analysed the differences between the selected primary studies and their impact on the conclusions and results.
Results: The overlap between SLRs with similar response variables (54%) is greater than between SLRs with similar objectives (36%). Only three per cent of the primary studies are included in all eight analysed SLRs. Conclusions regarding external quality and productivity may vary across the SLRs on TDD. While we found that SLR results are similar, these results may differ when authors classify primary studies by experiments and case studies.
Conclusion: SLRs with similar response variables tend to be more repeatable than SLRs with similar objectives and SLRs addressing the same topic. The SLR authors' criteria with respect to the consistency of evidence may influence the conclusions of SLRs on TDD. The results of SLRs where all primary studies count equally appear to be consistent. The SLR authors' criteria for selecting primary studies may influence the results classified by case studies and experiments.

Keywords: Reliability, Repeatability, Systematic Literature Review, Test-Driven Development.

1. Introduction

Evidence-based software engineering (EBSE) is a concept introduced by Kitchenham et al. [1] for the purpose of improving decision making with respect to software development and maintenance by integrating the best state-of-the-art evidence from research. Evidence can be sourced from laboratory experiments, industrial projects, observation studies, case studies, surveys and field experiments [2]. The EBSE concept led to evolutionary changes in research related to secondary studies in software engineering (SE). Secondary studies can be divided into systematic mapping studies (SMSs) and systematic literature reviews (SLRs). SMSs provide an overview of a specific topic of interest and identify the number and type of investigations and results available

on that topic [3]. According to Kitchenham and Charters, SLRs are a means of identifying, evaluating, and interpreting all the available research in response to specific research questions [3].

In this context, reliability refers to the situation where two studies on the same topic arrive at the same conclusions [4]. Previous research has attempted to comprehend reliability and repeatability of secondary studies through the comparison of two SLRs. For instance, MacDonell et al. [5] evaluated the reliability of SLRs in SE. To conduct their study, they compared two SLRs. They observed differences between the activities performed in the SLRs, which do not appear to have an effect on the conclusions. The study suggests that SLRs are a method that is robust to differences in people and processes. Kitchenham et al. [6] studied the property of SLR repeatability in SE. Six out of 32 primary studies used by the two SLRs are the same, that is, the overlap among the articles is rather small. They highlight that the missing primary studies may have a significant impact on the results of secondary studies. Wohlin et al. [4] conducted a study on the reliability of SMSs. To conduct the study, they used two SMSs in the software product line testing area. They concluded that the reliability of secondary studies should not be taken for granted, as secondary studies are highly dependent on the setting: study area, researchers conducting the study, search approach and data supplied about the primary study.

The above studies on repeatability and reliability arrive at different conclusions. MacDonell et al. [5] state that SLRs are robust to different people and processes, whereas Kitchenham et al. [6] determine that differences could have a significant impact on results, and Wohlin et al. [4] claim that the reliability of secondary studies should not be taken for granted. Unlike the above studies that used two SLRs for the purpose of their research, this study also compares two SLRs and in one case compares eight SLRs on test-driven development (TDD). TDD is a technique used since the early days of software development [7], which grew in popularity as an eXtreme Programming (XP) practice [8]. TDD has been studied since the 2000s [9]. There are many primary studies on TDD that use a range of research methods such as experiments, surveys and case studies in industrial and academic settings.

We identified eight SLRs studying the TDD technique: Kollanus [9], Sfetsos and Stamelos [10], Turhan et al. [11], Causevic et al. [12], Mäkinen and Münch [13], Munir et al. [14], Bissi et al. [15] and Abushama et al. [101]. What struck us most was that there are pronounced differences with respect to the primary studies included in the SLRs. This led us to formulate the research question: How reliable are the results of SLRs on TDD, that is, how confident can we be that the results and conclusions are stable with respect to the primary studies included throughout the process? This article reports a tertiary study on the reliability of the results and conclusions considering the eight SLRs found in the literature. Additionally, we analyse SLRs with similar objectives and SLRs with similar responsible variables. The study response variables are (QLTY) and productivity (PROD).

The remainder of the paper is structured as follows. Section 2 presents the related work. Section 3 describes the research method employed in the tertiary study. Section 4 analyses the differences and overlaps between the SLRs on TDD. Section 5 analyses the reliability of the conclusions and results of the SLRs with both similar objectives and similar response variables, also determining the effect of errors in SLRs on their results. Section 6 reports the validity threats and future work.

2. Related Work

We identified 20 secondary studies in SE that dated back to before 2007. As of 2007, many of the studies adopted the guidelines for performing SLRs proposed by Kitchenham and Charters [3]. In 2009, it was found that SLR quality had increased, and the number of secondary studies published per year was constant [16]. SLR reliability and repeatability studies were carried out in SE as of 2010. The first, published by MacDonell et al. [5], indicates that SLRs are a robust research method as they produce stable findings (findings that are unchanged) with respect to different processes and people. Later, Kitchenham et al. [6] determined that any primary studies that were not included could have an impact on the results, and Wohlin et al. [4] claimed that the results of secondary studies are highly dependent on the setting, including study area, people, search approach and data available in the primary study. We have not found any other studies on SLR reliability or repeatability in SE since 2013.

According to Zhang and Babar [17], SLR is a methodology that is relatively new to SE researchers, and its application has not been fully evaluated. Wohlin et al. [4] state that secondary study reliability can pose a challenge and, therefore, warrants further research. Although Munir et al. [14] do not focus on reliability, they do state that there are differences between SLRs on TDD with respect to the primary studies not included in existing SLRs whose causes are as follows: (i) studies published after the search date, which were not available when the study was conducted; or (ii) studies that do not meet the SLR inclusion criteria. As already noted,

* Corresponding author.

SLRs on TDD have sizeable differences with respect to the selected primary studies. There are no studies in SE that focus on explaining such differences and analysing the reliability of the results and conclusions. Unlike previous studies that compared only two SLRs, the research reported here aims to analyse reliability using all the SLRs on TDD.

3. Research Method

The tertiary study reported here applies the same methodology as any SLR, following the guidelines proposed by Kitchenham and Charters [18]. Specifically, we perform the following tasks: define the research questions, determine the review protocol, specify the search string, define the inclusion/exclusion criteria, select the secondary studies that meet the criteria, and extract and synthesize data.

3.1. Research questions

The research questions (RQ) formulated in this research are listed below: (**RQ-1**) Which SLRs on TDD have been reported in the literature?; (**RQ-2)** How much overlap is there between the SLRs on TDD in terms of the selected primary studies?; (**RQ-3**) How do the criteria used by the authors to select the primary studies affect the results and conclusions of the SLRs on TDD?

3.2. Review protocol

The review protocol was defined by the four activities below:

1. We gathered available information on the SLRs on TDD published in the scientific literature. We retrieved the articles referenced in each of the SLRs.
2. We used the information on the SLRs and their associated primary studies to determine the overlap between SLRs.
3. We analysed the reliability of the results and conclusions of the SLRs by comparing developer productivity and code quality. This activity was performed for SLRs with similar objectives, SLRs with similar response variables and all the SLRs on TDD.
4. We obtained the conclusions of the study.

3.3. Search strategy

To create the search string, we used the most significant terms for our research, divided into three components, as well as similar or related terms. The first component is related to TDD. The second component is related to the response variables of interest. The third component accounts for SLRs. The defined string was: *("test driven development" OR "test-driven development" OR TDD) AND (quality OR "code quality" OR "quality improvement" OR "internal quality" OR "external quality" OR productivity) AND ("systematic review" OR "systematic literature review" OR "systematic mapping" OR "systematic mapping study")*. Searches were run on Web of Science and Scopus. The search field used in Web of Science and Scopus was the same (Title OR Abstract OR Keywords). In Scopus, we obtained 14 articles, while in Web of Science, we obtained 4. In total, the string obtained 18 articles. The defined inclusion criteria are as follows: (i) the study analyses the effects of TDD on different response variables (for example, internal/external quality, productivity, conformance, developer opinions), or (ii) the study identifies aspects that limit TDD adoption in industry. On the other hand, the exclusion criteria are as follows: (i) the study does not study the effects of TDD on at least one response variable, (ii) the study is a multiple report, or (iii) the study is written in a language other than English.

3.4. Selected secondary studies

As mentioned above, we retrieved a total of 18 articles, of which four were excluded as duplicates, leaving 14 articles. Then, we selected the SLRs on TDD that consider quality or productivity as response variables, leaving a total of eight secondary studies.

4. SLRs on TDD

4.1. Analysed SLRs

The research was conducted based on SLRs on TDD, published from 2010 to 2021, analysing the effects of TDD on different response variables.

- Sfetsos and Stamelos [10] study the impacts of TDD on quality.

- Turhan et al. [11] analyse the impacts of TDD on productivity, internal/external quality and test quality.
- Kollanus [9] analyse the empirical evidence on the impacts of TDD on productivity and quality.
- Causevic et al. [12] identify the aspects that limit TDD adoption in industry.
- Mäkinen and Münch [13] study the effects of TDD on the following response variables: defects, code coverage, code complexity, coupling, cohesion, size, effort, external quality, productivity and maintainability.
- Munir et al. [14] evaluate TDD considering the response variables: quality, productivity, conformance and developer opinion.
- Bissi et al. [15] study the effect of TDD on the following response variables: internal quality, external quality and productivity.
- Abushama et al. [101] study the effect of TDD on project success factors (that is, cost, time and customer satisfaction/external quality).

Table 1 shows the SLRs on TDD. Column 4 lists the articles included in each SLR. Column 5 indicates the number of primary studies in the SLR. As we can see, the number of studies included in the publications varies enormously over similar time periods.

Table 1: SLRs on TDD included in this research.

Study	Year of publication	Period covered	Primary studies included	No. of primary studies
Sfetsos and Stamelos [10]	2010	Up until 2009	[19],[20],[21],[22],[23],[24],[25],[26],[27],[28],[29],[30],[31],[32],[33],[34],[35],[36]	18
Turhan et al. [11]	2010	2002-2009	[24],[26],[27],[31],[32],[33],[35],[37],[38],[39],[40],[41],[42],[43],[44],[45],[46],[47],[48],[49],[50],[51]	22
Kollanus [9]	2010	2001-2010	[19],[21],[37],[22],[23],[24],[25],[26],[27],[28],[29],[31],[32],[33],[34],[35],[36],[38],[40],[44],[45],[46],[47],[48],[49],[51],[52],[53],[54],[55],[56],[57],[58],[59],[60],[61],[62]	37
Causevic et al. [12]	2011	2002-2010	[19],[20],[21],[24],[25],[26],[27],[28],[29],[31],[32],[34],[35],[37],[38],[40],[45],[46],[47],[48],[49],[52],[55],[56],[58],[61],[63],[64],[65],[66],[67],[68],[69],[70],[71],[72],[73],[74],[75],[76],[77],[78],[79],[80],[81],[82],[83],[84]	48
Mäkinen and Münch [13]	2014	2003-2012	[19],[22],[26],[27],[28],[29],[32],[35],[37],[40],[46],[49],[54],[55],[58],[80],[85],[86],[87]	19
Munir et al. [14]	2014	2002-2012	[19],[20],[22],[24],[25],[26],[27],[28],[29],[31],[32],[33],[34],[35],[37],[38],[40],[44],[46],[47],[48],[49],[55],[58],[61],[67],[69],[85],[86],[87],[88],[89],[90],[91],[92],[93],[94],[95],[96], [97],[98]	41
Bissi et al. [15]	2016	2003-2012	[19],[20],[23],[24],[25],[26],[28],[29],[32],[33],[34],[37],[40],[46],[48],[49],[50],[54],[55],[56],[58],[61],[86],[88],[95],[99], [100]	27
Abushama et al. [101]	2020	1999-2019	[19],[24],[37],[47],[54],[57],[87],[95],[102],[103],[104],[105],[106],[107],[108],[109],[110],[111],[112],[113],[114],[115],[116],[117],[118],[119],[120],[121],[122],[123],[124]	31

The characteristics of the SLRs are reported in Table A1 of Appendix A (see https://encr.pw/Ctn2L). Table A1 lists the study objective, response variables, study type, scientific databases searched, search strings and date, date of publication, inclusion/exclusion criteria and quality criteria. The SLRs were published in 2010, 2011, 2014, 2016 and 2020, respectively. As shown in Table 1, there is a substantial difference across SLRs with respect to the number of primary studies, and two of the most recent SLRs conducted in 2014 (Mäkinen and Münch) and 2016 (Bissi et al.) include fewer primary studies (around half) than the SLR conducted in 2011.

4.2. Overlaps and differences between SLRs

In total, we identified 105 primary studies on TDD. Each primary study was referenced at least once. Table 2 lists the articles referenced by two or more SLRs (unshaded cells) and, on the diagonal (grey-shaded cells), articles referenced by only one SLR.

Table 2: Primary studies common to SLRs on TDD.

	Sfetsos & Stamelos [10]	**Turhan et al. [11]**	**Kollanus [9]**	**Causevic et al. [12]**	**Mäkinen & Münch [13]**	**Munir et al. [14]**	**Bissi et al. [15]**	**Abushama et al. [101]**
Sfetsos & Stamelos [10]	[30]	[24],[26], [27],[31], [32],[33], [35]	[19],[21], [22],[23], [24],[25], [26],[27], [28],[29], [31],[32], [33],[34], [35],[36]	[19],[20], [21],[24], [25],[26], [27],[28], [29],[31], [32],[34], [35]	[19],[22], [26],[27], [28],[29], [32],[35]	[19],[20], [22],[24], [25],[26], [27],[28], [29],[31], [32],[33], [34],[35]	[19],[20], [23],[24], [25],[26], [28],[29], [32],[33], [34]	[19],[24]
Turhan et al. [11]		[39],[41], [42],[43]	[24],[26], [27],[31], [32],[33], [35],[37], [38],[40], [44],[45], [46],[47], [48],[49], [51]	[24],[26], [27],[31], [32],[35], [37],[38], [40],[45], [46],[47], [48],[49]	[26],[27], [32],[35], [37],[40], [46],[49]	[24],[26], [27],[31], [32],[33], [35],[37], [38],[40], [44],[46], [47],[48], [49]	[24],[26], [32],[33], [37],[40], [46],[48], [49],[50]	[24],[37], [47]
Kollanus [9]			[53],[57], [59],[60], [62]	[19],[21], [24],[25], [26],[27], [28],[29], [31],[32], [34],[35], [37],[38], [40],[45], [46],[47], [48],[49], [52],[55], [56],[58], [61]	[19],[22], [26],[27], [28],[29], [32],[35], [37],[40], [46],[49], [54],[55], [58]	[19],[22], [24],[25], [26],[27], [28],[29], [31],[32], [33],[34], [35],[37], [38],[40], [44],[46], [47],[48], [49],[55], [58],[61]	[19],[23], [24],[25], [37],[40], [26],[28], [29],[32], [33],[34], [46],[48], [49],[54], [55],[56], [58],[61]	[19],[24], [37],[47], [54],[57]

	Sfetsos & Stamelos [10]	**Turhan et al. [11]**	**Kollanus [9]**	**Causevic et al. [12]**	**Mäkinen & Münch [13]**	**Munir et al. [14]**	**Bissi et al. [15]**	**Abushama et al. [101]**
Causevic et al. [12]				[63],[64], [65],[66], [68],[70], [71],[72], [73],[74], [75],[76], [77],[78], [79],[81], [82],[83], [84]	[19],[26], [27],[28], [29],[32], [35],[37], [40],[46], [49],[55], [58],[80]	[19],[20], [24],[25], [26],[27], [28],[29], [31],[32], [34],[35], [37],[38], [40],[46], [47],[48], [49],[55], [58],[61], [67],[69]	[19],[20], [24],[25], [26],[28], [29],[32], [34],[37], [40],[46], [48],[49], [55],[56], [58],[61]	[19],[24], [37],[47]
Mäkinen & Münch [13]						[19],[22], [26],[27], [28],[29], [32],[35], [37],[40], [46],[49], [55],[58], [85],[86], [87]	[19],[26], [28],[29], [32],[37], [40],[46], [49],[54], [55],[58], [86]	[19],[37], [47],[54], [87]
Munir et al. [14]						[89],[90], [91],[92], [93],[94], [96],[97], [98]	[19],[20], [24],[25], [26],[28], [29],[32], [33],[34], [37],[40], [46],[48], [49],[55], [58],[61], [86],[88], [95]	[19],[24], [37],[87], [95]
Bissi et al. [15]							[99],[100]	[19],[24], [37],[54],

	Sfetsos & Stamelos [10]	Turhan et al. [11]	Kollanus [9]	Causevic et al. [12]	Mäkinen & Münch [13]	Munir et al. [14]	Bissi et al. [15]	Abushama et al. [101]
								[95]
Abushama et al. [101]								[102],[103], [104],[105], [106],[107], [108],[109], [110],[111], [112],[113], [114],[115], [116],[117], [118],[119], [120],[121], [122],[123], [124]

With respect to studies that appear in two or more SLRs, Kollanus [9] and Causevic et al. [12] share 25; Causevic et al. [12] and Munir et al. [14] share 24 studies, and Kollanus [9] and Munir et al. [14] also share 24 studies. The overlap between SLRs is smaller in the case of Sfetsos and Stamelos [10] and Turhan et al. [11], which share only seven studies; Sfetsos and Stamelos [10] and Mäkinen and Münch [13], which share eight studies. There is very little intersection between the studies referenced by Abushama et al. [101] and other SLRs (fewer than seven matches in all cases). Strikingly, there is very little overlap between the primary studies included in the SLRs, which exhibit some sizeable differences. We have termed the primary studies that appear in only one SLR specific studies. Mäkinen and Münch [13] do not have any specific studies, all their primary studies overlap with other SLRs; Sfetsos and Stamelos [10] have one specific primary study, whereas Munir et al. [14] have nine, Causevic et al. [12] include 19 and Abushama et al. [101] reference 23 specific studies, respectively.

Clearly, there are differences across the SLRs with respect to the studies that they include. These differences could influence the conclusions and results of the secondary studies. For the purpose of analysing the inclusion or exclusion of each of the primary studies in the SLRs, we built Table 3. Due to space restrictions, only an excerpt from Table 3 is included. The complete Table 3 can be found at https://acortar.link/JwvQxI. The SLRs are listed in the columns and the primary studies in the rows of Table 3. If the SLR includes the primary study, the reference number (or identifier, for example, S3) used in the SLR appears in the respective cell. In Turhan et al. [11], there is an X in place of the reference number, since the references are not numbered. We list the possible reasons why the primary study was not included in the SLR below:

1. **NEW**: The primary study was published after the SLR. Some primary studies are not included in the SLRs because they were published after the SLR search date.

2. **OK**: There is a proper reason for not including the article, which has nothing to do with the author of the SLR. Non-inclusion could be due one or more of the following reasons: (i) The primary study was not selected because it does not address the SLR response variable, for example, the SLR studies conformance and the primary study does not; (ii) Some primary studies are reported in more than one article, and SLRs generally do not include multiple reports; (iii) The primary study uses a research method that is not considered in the SLR, and it is, therefore, not included in the review. For example, the secondary study by Munir et al. [14] does not include experience reports; Turhan et al. [11] does not include questionnaires; Kollanus [9], Sfetsos and Stamelos [10], Causevic et al. [11] and Bissi et al. [15] do not include reviews;

only Bissi et al. [15] include simulations. None of the secondary studies include expert opinions. The exclusion of the article from the SLR is justified in all the above cases.

3. **SLR**: The article is excluded on the grounds of the criteria used in the SLR. The author considered that the article does not comply with the inclusion criteria, for example, all the SLRs accept journal (JCR) or conference papers only, except Turhan et al. [11] who included technical reports and theses.

4. **NK**: The grounds on which the authors did not include the primary study in their SLR are not known.

As Table 3 shows, the letters **EXP** appear next to the reference to some of the studies included in the SLRs (these table cells are shaded grey). In our opinion, these studies should not have been included in the SLRs, because they do not focus on the TDD technique. For example, (i) they focus on TDD as technique for improving testing quality, (ii) they study how to improve agile development course teaching, or (iii) they examine the influence of TDD on the implementation of spreadsheets. Their inclusion appears to be the result of a wrong decision made by the SLR authors. We attempt to determine whether this affects their conclusions later. The last but one column of Table 3 states the grounds for article inclusion, and the far-right column notes on the reasons for article exclusion from the respective SLRs. Figure 1 is a bar graph that illustrates how many primary studies were included and the main reasons why the articles missing from SLRs were not included. We used the following nomenclature:

1. **REF:** articles correctly referenced in the SLR.
2. **EXP:** articles referenced by the SLR that do not, in our opinion, focus on TDD.
3. **SR:** articles not included on grounds attributable to the SLR.
4. **OK:** articles not included in the SLR on justified grounds.
5. **NK:** articles not included for unknown reasons.
6. **NEW:** articles not included because they were published after the SLR search date.

Table 3: Fragment of the studies included in the eight secondary studies.

ID	Author	Research method	Year	[10]	[11]	[9]	[12]	[13]	[14]	[15]	[101]	Why was the article included in the SLRs?	Why was the article not included in the SLRs?
[PS1]	Abrahamsson et al. [63]	Case study	2005	OK	OK	OK	[8]	OK	OK	OK	OK	The article conforms to Causevic et al. [12], which is the only SLR that contemplates measures of perception.	The other SLRs do not include this study, as they do not accept perceptions on TDD.
[PS2]	Aniche & Gerosa [88]	Questionnaire	2010	NEW	NEW	NEW	NEW	OK	[72]	[40]	OK	This article conforms to the SLRs by Munir et al. [14] and Bissi et al. [15], as they accept questionnaires.	This article does not conform to Mäkinen and Münch [13] or Abushama et al. [101], as these secondary studies do not include questionnaires.
[PS3]	Aniche & Gerosa [99]	Qualitative study	2012	NEW	NEW	NEW	NEW	OK	OK	[38]	OK	Bissi et al. [15] include this article even though it is written in a language other than English (Portuguese).	This article does not conform to any SLR, as it is a study written in a language other than English (Portuguese).
[PS4]	Bannerman & Martin [89]	Case study	2011	NEW	NEW	NEW	NEW	OK	[60] EXP	OK	OK	Munir et al. [14] included this article, even though it does not focus on the TDD technique.	This article does not conform to any SLR as it studies TWD (Test with Development), a technique that is based on but, strictly speaking, not TDD.
[PS7]	Bhat & Nagappan [19]	Case study	2006	[16] S3	OK	[9]	[9]	[12]	[64]	[33]	[44]	The article conforms to most of the SLRs and is included in 7 out of the 8 SLRs.	Turhan et al.'s SLR [11] does not include this article but references another article that reports this study [PS81].
[PS8]	Canfora et al. [37]	Controlled experiment	2006	NK	X	[37]	[11]	[13]	[43]	[32]	[7]	The article conforms to 7 out of the 8 SLRs.	It is not known why Sfetsos and Stamelos [10] did not include this article.
[PS9]	Canfora et al. [52]	Experiment	2006	OK	OK	[36]	[10]	OK	OK	OK	OK	Kollanus [9] and Causevic et al. [12] cite this duplicated reference. [PS8] and [PS9] report the same study.	The study was already reported in [PS8].
[PS10]	Cao & Ramesh [64]	Case study	2008	OK	OK	OK	[12] EXP	OK	OK	OK	OK	Causevic et al. [12] included this article that does not	The study focuses on requirements rather than the TDD technique.

ID	Author	Research method	Year	[10]	[11]	[9]	[12]	[13]	[14]	[15]	[101]	Why was the article included in the SLRs?	Why was the article not included in the SLRs?
												appear to focus on the TDD technique.	
[PS14]	Chien et al. [65]	Experiment	2008	OK	OK	OK	[13] EXP	OK	OK	OK	OK	Causevic et al. [12] included this article that does not appear to focus on the TDD technique.	The study reported in this article evaluates a TDD-based training method.
[PS15]	Crispin [90]	Expert opinion	2006	OK	OK	OK	OK	OK	[65] EXP	OK	OK	Munir et al. [14] includes this article that does not appear to conform to the secondary studies.	This study is not considered by the other SLRs, as it does not report results, it is an expert opinion.

As Figure 1 shows, 22, 20, 12, 12 and 32 primary studies were excluded on unknown grounds (**NK**) from the SLRs by Mäkinen and Münch [13], Bissi et al. [15], Sfetsos and Stamelos [10], Turhan et al. [11], and Abushama et al. [101], respectively. On the other hand, 26, 29, 24 and 24 primary studies were excluded on the grounds that they were published after the SLR search date (**NEW**) from SLRs by Sfetsos and Stamelos [10], Turhan et al. [11], Kollanus [9] and Causevic et al. [12], respectively. The SLRs by Causevic et al. [12] and Munir et al. [14] included eight and seven studies, respectively, that we believe should not have been included (**EXP**). The conditions of the SLR conducted by Causevic et al. [12] led to the exclusion of nine articles. The differences between the SLRs on TDD have different causes, where most are unknown (**NK**), and others depend on the SLR criteria.

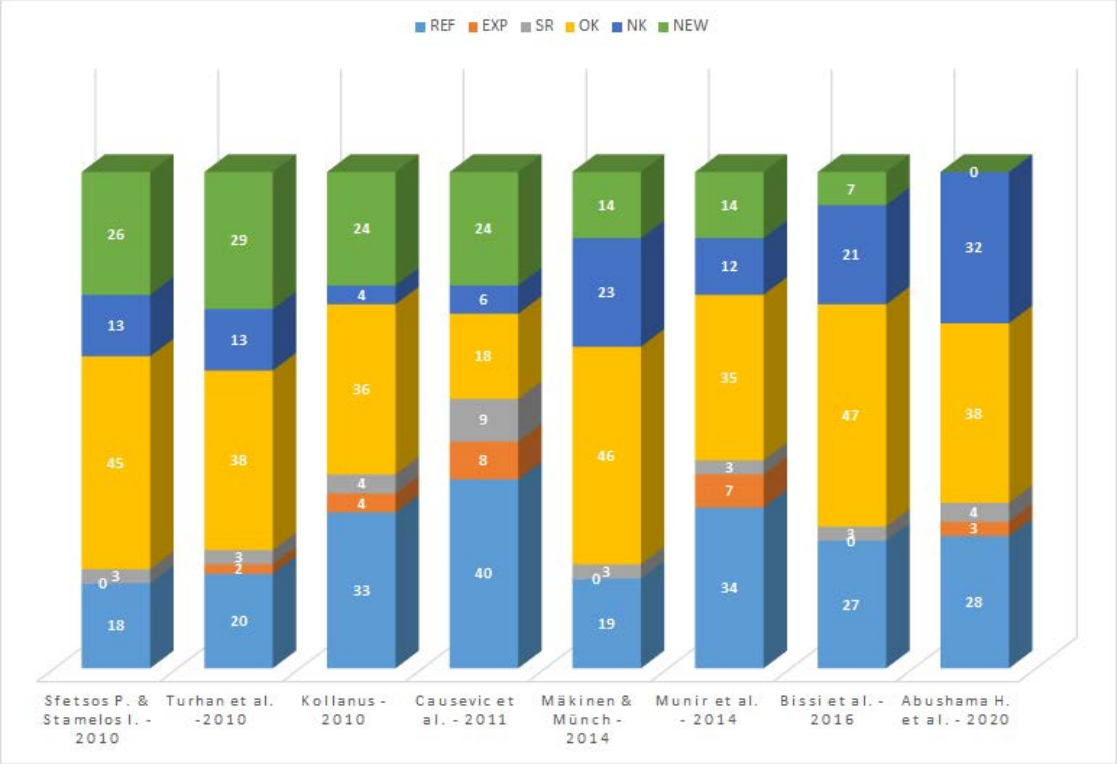

Figure 1: Number of primary studies included in or missing from the SLRs.

5. Influence of the Differences on the Conclusions and Results of the SLRs

For the purpose of analysing the stability of the conclusions and results with respect to the external quality and productivity response variables, we compared the following SLRs: (i) SLRs with similar objectives, (ii) SLRs with similar response variables, and (iii) all SLRs on TDD.

5.1. Stability analysis of the conclusions and results of the SLRs with similar objectives

According to the characteristics of the SLRs on TDD (Table A1 of Appendix A), the objectives of the SLRs differ. According to Sfetsos and Stamelos [10], their SLR aims to evaluate the empirical results with respect to the quality of agile practices. Turhan et al. [11] and Mäkinen and Münch [13] claim that their objective is to offer an updated overview of evidence on the impacts of TDD. Kollanus [9] specifies that his objective is to analyse current empirical evidence on the effects of TDD, whereas Bissi et al. [15] aim to analyse the conclusions of articles addressing the effects of TDD. Munir et al. [14] investigate what changes rigour and relevance bring about in SLR conclusions. Causevic et al. [12] claim that they set out to identify the factors limiting the adoption of TDD. Abushama et al. [101] define their objective as determining the impact of the use of TDD and BDD on project success factors. The SLRs by Turhan et al. [11] and Mäkinen and Münch [13] have common objectives, namely, both offer a full overview of the evidence of the effects of TDD, in both studies this objective has been written explicitly and coincidentally. In this study, we analyse the stability of the results and

conclusions considering that the secondary studies have similar objectives and differences with respect to the primary studies that they include.

5.1.1. Comparison of SLRs with similar objectives

Table 4 shows the characteristics of the SLRs by Turhan et al. [11] and Mäkinen and Münch [13], and the last column specifies their differences. Although the objectives are similar, we find that the SLRs differ in terms of the type of studies included. For instance, Turhan et al. [11] include technical reports and theses, whereas Mäkinen and Münch [13] do not. Each SLR has particular inclusion criteria and neither details the exclusion criteria. Finally, there are differences with regard to the response variables (Table 4, column 4).

For the purpose of this comparison, we considered that the primary studies [PS8] and [PS9] report the same study. Therefore, only one of the two was taken into account. The same applies to [PS16] and [PS17]; [PS23] and [PS24]; [PS35], [PS36] and [PS37]; [PS38] and [PS39]. With regard to the publication dates of the SLRs, there is a difference of four years. On this ground, primary studies included in Mäkinen and Münch's SLR [13] that were published later than March 2009 were excluded from the comparison. These are: (i) the experiments by Desai et al. [22], Madeyski [58], and Pancur and Ciglaric [86]; (ii) the case study by Dogša and Batić [85]; and (iii) the quasi experiment by Wilkerson et al. [87].

Table 4: Characteristics of the SLRs with similar objectives.

	Turhan et al. [11]	**Mäkinen & Münch [13]**	**Turhan et al. [11] vs. Mäkinen & Münch [13]**
Objective	Provide an overview of the current evidence on TDD impacts.	Gain an overview of the observed effects of TDD.	They are similar.
Response variables	Internal software quality (complexity, reuse, coupling, cohesion), external quality and productivity, test quality (test coverage, test density, test productivity, test effort).	Defects, code coverage, code complexity, coupling, cohesion, size, effort, external quality, productivity, maintainability.	Mäkinen & Münch [13] include size, effort and maintainability. Turhan et al. [11] include reuse, test density, test productivity and test effort.
Study types	Controlled experiments, pilot study, commercial projects in industry.	Experiments, case studies.	They are similar.
Digital libraries	ACM, IEEE, Elsevier and grey literature (technical reports, thesis).	IEEE, ACM, Springer, Elsevier.	Turhan et al. [11] include technical reports and theses.
Search strings	Unspecified, but they do state that they include official variants of TDD like: ATDD (acceptance-test-driven development), BDD and STDD (story-test-driven development).	Test driven development, test driven, test first programming and test first.	They are not specified by Turhan et al. [11].
Search date	First quarter of 2009	2012	Studies published later than the first quarter of 2009 included in Mäkinen & Münch [13] were not taken into account in this study.
Date of publication	2010	2014	Mäkinen & Münch [13] published 4 years after Turhan et al. [11].
Inclusion criteria	We infer that they consider study quality, including design, setting, participants, treatments, controls and results.	The main criterion was whether the publication included the empirical results of test-based development and reported industrial or academic results.	They each use their own inclusion criteria.

	Turhan et al. [11]	Mäkinen & Münch [13]	Turhan et al. [11] vs. Mäkinen & Münch [13]
Exclusion criteria	Multiply reported articles. They offer no further information on the exclusion criteria.	They used publication quality and overall relevance as exclusion criteria in some cases.	Exclusion criteria are not detailed in either case.

We find that Turhan et al. [11] select 21 articles, whereas Mäkinen and Münch [13] include 13 articles. Only nine of the articles in the two SLRs are the same, whereas Turhan et al. [11] have 12 and Mäkinen and Münch [13] have four specific articles, respectively. Only nine articles (36%) included in the two SLRs are the same, and the remaining 64% articles are specific to each of the SLRs. Although the SLRs have similar objectives, the overlap is 36% (which we consider to be a poor overlap).

Table 5 shows the differences with regard to the selected primary studies. The last two columns specify why the articles were not included in each of SLRs. Looking at the last column of Table 5, we find that the main difference is that Mäkinen and Münch's SLR [13] did not include seven experiments and two case studies that were selected by Turhan et al. [11].

Table 5: Differences between the SLRs with similar objectives with respect to the included studies.

ID	Author	Turhan et al. [11]	Mäkinen & Münch [13]	Studies not included in [11]	Studies not included in [13]
[PS8]	Canfora et al. [37]	X	[13]		
[PS25]	Erdogmus et al. [24]	X	NK		It is not known why this experiment on quality and productivity was not included.
[PS29]	Flohr & Schneider [38]	X	NK		It is not known why this experiment on productivity was not included.
[PS35] / [PS36]	[39] / [54]	X	--- / [15]		
[PS38] / [PS39]	Geras et al. [40] / Geras [41]	X / X	[16] / ---		
[PS40]	Gupta & Jalote [26]	X	[24]		
[PS44]	Huand & Holcombe [27]	X	[25]		
[PS46]	Janzen & Saiedian [55]	NK	[26]	It is not known why this experiment on productivity was not included.	
[PS48]	Janzen [42]	X	OK		The decision not to include this PhD thesis appears to be correct as it is not a research paper.
[PS52]	Janzen & Saiedian [28]	NK	[20]	It is not known why this case study on the impact of TDD on	

ID	Author	Turhan et al. [11]	Mäkinen & Münch [13]	Studies not included in [11]	Studies not included in [13]
				internal quality was not included.	
[PS55]	Kaufmann & Janzen [43]	X EXP	NK		The decision not to include this case study that evaluates internal quality and productivity appears to be correct as it is a 2-page article.
[PS66]	Madeyski [44]	X	NK		It is not known why this experiment that evaluates external quality was not included.
[PS67]	Madeyski [45]	X	NK		It is not known why this experiment that evaluates object-oriented design was not included.
[PS69]	Madeyski & Szala [46]	X	[21]		
[PS73]	Maximilien & Williams [29]	NK	[17]	It is not known why this case study that evaluates external quality was not included.	
[PS78]	Müller & Hagner [31]	X	NK		It is not known why this experiment that evaluates test case reliability was not included.
[PS79]	Müller & Höfer [80]	NK	[22]	It is not known why this quasi experiment that evaluates test quality was not included.	
[PS7] / [PS81]	Nagappan et al. [32] / Bhat & Nagappan [19]	X / X	[18] / [12]		
[PS82]	Pancur et al. [33]	X	NK		It is not known why this experiment that evaluates external quality was not included.
[PS94]	Siniaalto & Abrahamsson [47]	X	NK		It is not known why this case study that evaluates internal quality was not included.
[PS95]	Slyngstad et al. [48]	X	NK		It is not known why this case study that evaluates TDD on framework development was not included.

ID	Author	Turhan et al. [11]	Mäkinen & Münch [13]	Studies not included in [11]	Studies not included in [13]
[PS98]	Vu et al. [49]	X	[29]		
[PS101]	Williams et al. [35]	X	[19]		
[PS103]	Yenduri & Perkins [50]	X	NK		It is not known why this case study that evaluates internal quality and productivity was not included.
[PS105]	Zhang et al. [51]	X	NK		It is not known why this experiment that studies several response variables was not included, possibly because it is a 2-page report.

5.1.2. Stability of the conclusions and results of SLRs with similar objectives

Having identified the differences between the SLRs, we analysed the differences in the conclusions and results. Table 6 shows the differences with respect to the quality (QLTY) and productivity (PROD) response variables. Turhan et al.'s conclusions [11] suggest that the evidence is not consistent with respect to either of the measures. Mäkinen and Münch [13] claim that TDD may reduce the number of defects, although development time may be longer. The conclusions of the SLRs with respect to quality and productivity appear to be different. The conclusions of the SLRs with similar objectives may be influenced by the authors' opinion with respect to the resulting evidence. Looking at Table 6, we find that Turhan et al. [11] state that 13 out of the 22 studies (59%) report an improvement in external quality, whereas the percentage calculated by Mäkinen and Münch [13] is 66%. The results for external quality are similar, and they both claim that TDD use improves external quality. As regards the effect of TDD on productivity, the results of the SLRs are again consistent: both suggest that the results are inconclusive. Although there are sizeable differences between the SLRs with similar objectives, we find that the results of the secondary studies on TDD are stable.

Table 6: Comparison of the conclusions and results of SLRs with similar objectives.

	Turhan et al. [11]	Mäkinen & Münch [13] up until 2009	Turhan et al. [11] vs. Mäkinen & Münch [13] up until 2009
QLTY CONCLUSION	There are many question marks surrounding the impacts of TDD. The evidence with respect to the impacts of TDD is not consistent for any of the measures.	TDD may reduce the number of defects.	**There are differences.** Turhan et al. [11] state that the evidence is not consistent with respect to any of the measures. Mäkinen & Münch [13] state that TDD can reduce the number of defects.
PROD CONCLUSION	There are still many question marks surrounding the impacts of TDD. The evidence with respect to the impacts of TDD is not consistent for any of the measures.	While code maintenance may be faster, development time may be longer.	**There are differences.** Turhan et al. [11] state that the evidence is not consistent. Mäkinen & Münch [13] state that development may take longer.
QLTY RESULTS	In 13 out of 22 studies, TDD improves external quality (59%). In 3 studies, TDD reduces external quality. The results are inconclusive or no different in 6 studies.	The SLR evaluates the effect of TDD on defect reduction. We find that TDD reduces defects in 4 studies (66%), whereas 2 are inconclusive.	Results are similar with respect to quality improvement in several studies and the presence of inconclusive results.
QLTY RESULTS - Case Studies	The author indicates that pilot and industry studies suggest that TDD produces better external quality.	The author indicates that several industrial cases studies show a relatively large reduction in defects (It is noted that	Similar results, TDD improves external quality.

	Turhan et al. [11]	Mäkinen & Münch [13] up until 2009	Turhan et al. [11] vs. Mäkinen & Münch [13] up until 2009
		all 4 case studies suggest that TDD reduces defects).	
QLTY RESULTS - Experiments	The authors state that the results of the controlled experiments are inconclusive. Out of 6 experiments, 3 are inconclusive, TDD improves quality in 1, and quality drops in 2.	The authors state that, in the experiments reported in the SLR, TDD did not produce better results than other development methods (there are no differences between the effects of TDD in 2 experiments).	Mäkinen and Münch [13] reference only 2 experiments compared to the 6 experiments referenced by Turhan et al. [11]. The results are similar.
PROD RESULTS	TDD does not have a consistent impact on productivity.	The results are found not to be inconclusive. TDD reduces productivity in 2 studies, and the other 8 are inconclusive.	Results are similar.
PROD RESULTS - Case Studies	The pilot studies return mixed evidence. The industrial studies suggest that TDD leads to lower productivity.	No case studies were taken into account.	Mäkinen & Münch [13] do not have case studies for comparison.
PROD RESULTS - Experiments	The authors state that the evidence from the controlled experiments suggests that TDD use improves productivity. Of the referenced controlled experiments, 3 suggest that productivity increases with TDD use, whereas there is no difference in 1 experiment.	They find that the results are inconclusive. TDD reduces productivity in 2 experiments, and the results are inconclusive in 6 experiments.	**Contradictory results**. Turhan et al. [11] suggest an improvement in productivity, whereas Mäkinen & Münch [13] state that the results are inconclusive.

Most of the SLRs report results classified by research method, that is, they describe the results of the observed effect in experiments on the one hand and the effect observed by case studies on the other. This applies to Sfetsos and Stamelos [10], Turhan et al. [11], Kollanus [9] and Bissi et al. [15]. We analyse the stability of the results classified by research method (case studies and experiments) below. With regard to the results classified by case studies, both of the above SLRs agree, according to Table 6, that TDD improves external quality, that is, the results are stable. With regard to productivity, Mäkinen and Münch [13] do not describe any case studies on productivity, and, therefore, the results are not comparable. With regard to the results classified by experiments, both SLRs with similar objectives agree that TDD improves external quality, that is, the results are stable (see Table 6). With respect to productivity, Turhan et al. [11] state that TDD improves productivity, whereas Mäkinen and Münch [13] report inconclusive results with respect to the effects of TDD on productivity. The results of the SLRs with similar objectives classified by experiments report differences with regard to productivity. The results appear to be influenced by the inclusion criteria used by authors of SLRs.

5.2. Analysis of the stability of the conclusions and results of the SLRs with similar response variables

The SLRs on TDD study the effects on different response variables. Sfetsos and Stamelos [10] investigate the effect of TDD on quality. Turhan et al. [11] analyse the impacts of TDD on internal/external quality, productivity and test quality. Kollanus [9] and Bissi et al. [15] study the impacts of TDD on productivity and external/internal quality. Munir et al. [14] analyse the impact of TDD on quality, developer opinion, conformance, size, robustness and productivity. Causevic et al. [12] consider the development time, experience/knowledge, design, refactoring, test skills, adherence, code quality, cost, code coverage, complexity, feedback time, domain and specific tools, code size, perceptions, communication and (customer) collaboration, legacy code, defect reproduction and documentation response variables. Abushama et al. [101] analyse the impacts of TDD on time, cost, customer satisfaction and external quality. For the purpose of analysing the stability of the conclusions and SLRs with similar response variables, we used the studies by Kollanus [9] and Bissi et al. [15]. Both SLRs analyse the impact of TDD on the response variables: productivity and internal/external quality. We focus on the stability of the conclusions with respect to external quality and productivity.

5.2.1. Comparison between SLRs with similar response variables

Table 7 shows a comparison of the general characteristics of the SLRs by Kollanus [9] and Bissi et al. [15]. There are several differences which are described in the last column of Table 7. The publication year of the SLR by Bissi et al. [15] is 2014, whereas Kollanus [9] published in 2010. For the purposes of comparison, we did not include the following articles referenced by Bissi et al. [15], which were published after the date of Kollanus' search [9]: (i) the survey by Aniche and Gerosa [88], which analyses external quality; (ii) the qualitative study by Aniche and Gerosa [99], which analyses internal quality; and (iii) the experiment by Pancur and Ciglaric [86], which studies quality and productivity. With respect to the primary studies reported in more than one article: the primary study by Bhat and Nagappan [PS7] referenced by Kollanus [9] is also reported in Nagappan et al. [PS81]. The primary study by Geras et al. [40] is also reported in Geras et al. [PS38] and the thesis by Geras [PS39]. Canfora et al. [PS9] cited by Kollanus [9] is part of Canfora et al. [PS8]. Edwards [PS24] reports the same study as Edwards [PS23], as do George and Williams [PS36] and George and Williams [PS37]. Damm and Lundberg [PS16] and Damm and Lundberg [PS17] also report the same study. For the purposes of this comparison, we consider only one of each of these reports.

Table 7: Characteristics of the SLRs with similar response variables.

	Kollanus [9]	**Bissi et al. [15]**	**Kollanus [9] vs. Bissi et al. [15]**
Objective	Analyse current empirical evidence on TDD.	Analyse the conclusions of articles about the effects of TDD.	The objectives are related.
Response variables	Productivity, internal quality, external quality	Internal/external software quality and productivity.	Similar.
Study types	Experiments, case studies and surveys.	Experiments, case studies, surveys and simulations.	Bissi et al. [15] include simulations.
Digital libraries	ACM, IEEE, Springer.	ACM, IEEE, CiteSeerx, Science Direct and Wiley.	Similar.
Search strings	"TDD" and "test-driven development".	Software Engineers, Software Developers, Programmers — Test Driven Development, Test-Driven Development, TDD, Test First Development. — Code Quality, Quality Improvement, Design Improvement, Improved Software Development, Internal Quality, External Quality, Productivity.	Bissi et al. [15] include more synonyms.
Search date	Not specified.	IEEE Xplore 18/12/2014 ScienceDirect 18/12/2014 ACM Digital Library 19/12/2014 CiteSeerx 20/12/2014 Wiley Online Library 20/12/2014	Kollanus [9] does not specify the search date but includes articles published up until 2009.
Date of publication	September 2010 (paper submitted 11 April 2010).	Accepted: 15 February 2016. Available online: 24 February 2016.	Bissi et al. [15] published 6 years after Kollanus [9].
Inclusion criteria	Articles that include any sort of empirical evidence on TDD. In an optimal situation, high quality studies should be selected, but the	The full text of the articles must accessible, and the articles should have been published in scientific format (journals or conference proceedings). The articles must	Bissi et al. [15] specify more inclusion criteria than Kollanus [9].

	Kollanus [9]	Bissi et al. [15]	Kollanus [9] vs. Bissi et al. [15]
	research reports offer limited information in this case.	report results on TDD practice in software development. They must properly describe the setting in which the research was conducted. The results must be based on a clear measurement criterion.	
Exclusion criteria	Articles that do not study TDD. Articles that address ATDD or related concepts. Articles on XP that do not address TDD. Publications that are not research articles. Conference proceedings were regarded as research articles, but IEEE Software short papers were excluded.	Multiple reports. Articles not related to TDD practice in software development. SLRs on TDD practice. Articles not written in English. Articles not published in academic format. Articles whose full text is not accessible.	Bissi et al. [15] provide more specific exclusion criteria.

Kollanus [9] selects 34 articles, whereas Bissi et al. [15] include 23 articles in their SLR. These SLRs have 20 articles in common, Kollanus [9] has 14 specific articles and Bissi et al. [15] has only 3 specific articles. There is a 54% overlap, which is equivalent to the articles reviewed by both SLRs, and the remaining 46% are articles that are specific to one or other of the SLRs. The SLRs with similar response variables have more articles in common (54%) (that is, there is a bigger overlap) than the SLRs with similar objectives (36%). The SLRs with similar response variables are more repeatable than the SLRs with similar objectives. Table 8 shows the differences between the SLRs by Kollanus [9] and Bissi et al. [15] with respect to the selected studies. The last two columns list the reasons why each of the articles were not included in each SLR. The major differences between the two SLRs are: (i) five experiments and three case studies that appear in Kollanus [9] are missing in Bissi et al. [15], and (ii) Kollanus [9] includes six primary studies ([PS65][PS67][PS84][PS85][PS94][PS105]), which, in our opinion, do not focus on the TDD technique.

Table 8: Differences between the SLRs with similar response variables with respect to the selected studies.

ID	Author	Kollanus [9]	Bissi et al. [15]	Studies not included in [9]	Studies not included in [15]
[PS7]	Bhat & Nagappan [19]	[9]	[33]		
[PS8]	Canfora et al. [37] / Canfora [52]	[37] / [36]	[32]		
[PS16] / [PS17]	Damm & Lundberg [20] / Damm & Lundberg [21]	--- / [28]	[41] / ---		
[PS19]	Desai et al. [22]	[34]	NK		It is not known why this experiment on QLTY and PROD was not included.
[PS23] / [PS24]	Edwards [23] / Edwards [53]	[32] / [7]	[30] / ---		
[PS25]	Erdogmus et al. [24]	[5]	[36]		

ID	Author	Kollanus [9]	Bissi et al. [15]	Studies not included in [9]	Studies not included in [15]
[PS29]	Flohr & Schneider [38]	[40]	NK		It is not known why this experiment on PROD was not included.
[PS37] / [PS36]	George & Williams [25] / George [54]	18] / [47]	22] / [24]		
[PS38]	Geras et al. [40]	[44]	[21]		
[PS40]	Gupta & Jalote [26]	[19]	[23]		
[PS44]	Huang & Holcombe [27]	[20]	NK		It is not known why this experiment on QLTY and PROD was not included.
[PS46]	Janzen & Saiedian [55]	[6]	[35]		
[PS49]	Janzen et al. [56]	[45]	[29]		
[PS51]	Janzen & Saiedian [95]	NK	[31]	It is not known why this experiment on PROD was not included.	
[PS52]	Janzen & Saiedian [28]	[10]	[26]		
[PS65]	Lui & Chan [57]	[30] EXP	OK		The authors appear to be right not to include this 4-page article reporting a case study that studies the effect of TDD on team performance.
[PS66]	Madeyski [44]	[23]	NK		It is not known why this experiment that focuses on both TDD and pair programming was not included, as it reports quality-related results.
[PS67]	Madeyski [45]	[38]	OK		The authors appear to be right not to include this experiment focused on the effect of TDD on software design. Bissi et al. [15] do not study this response variable.
[PS69]	Madeyski & Szala [46]	[31]	[43]		
[PS71]	Madeyski [58]	[39]	[42]		
[PS73]	Maximilien & Williams [29]	[24]	[25]		
[PS78]	Müller & Hagner [31]	[22]	NK		It is not known why this experiment that studies the impact of TDD on productivity and quality was not included.
[PS81]	Nagappan et al. [32]	[8]	[44]		

ID	Author	Kollanus [9]	Bissi et al. [15]	Studies not included in [9]	Studies not included in [15]
[PS82]	Pancur et al. [33]	[21]	[28]		
[PS84]	Rahman [59]	[33] EXP	OK		The authors appear to be right not to include this 2-page article that reports a case study because it focuses on TBC (a method based on TDD) and not on TDD.
[PS85]	Rendell [60]	[41]	OK		The authors appear to be right not to include this experience report.
[PS89]	Sanchez et al. [34]	[26]	[37]		
[PS93]	Siniaalto & Abrahamsson [61]	[42]	[20]		
[PS94]	Siniaalto & Abrahamsson [47]	[43]	OK		The authors appear to be right not to include this case study on the effect of TDD on design. Bissi et al. [15] do not study this response variable.
[PS95]	Slyngstad et al. [48]	[25]	[39]		
[PS97]	Turnu et al. [100]	OK	[34]	Kollanus [9] does not include simulations.	
[PS98]	Vu et al. [49]	[46]	[27]		
[PS101]	Williams et al. [35]	[27]	NK		It is not known why this case study that studies the impact of TDD on external quality was not included.
[PS102]	Xu & Li [62]	[35]	NK		It is not known why this case study that studies the impact of TDD on productivity and external quality was not included.
[PS103]	Yenduri & Perkins [50]	NK	[45]	It is not known why this case study that studies the impact of TDD on internal and external quality and productivity was not included.	
[PS104]	Ynchausti [36]	[29]	NK		It is not known why this case study on quality and productivity was not included.
[PS105]	Zhang et al. [51]	[48]	OK		The authors appear to be right not to include this 2-page article that reports an experiment. The experiment focuses on several response variables, including productivity.

5.2.2. Stability of the conclusions and results of SLRs with similar response variables

Having identified the differences between the SLRs with similar response variables, we analysed the stability of the conclusions and results with respect to quality and productivity. As Table 9 shows, the conclusions on productivity are similar for the two SLRs. With regard to the quality study, Kollanus [9] and Bissi et al. [15] also concluded that TDD has a positive impact. The conclusions suggest that TDD may improve external quality, and there is moderate evidence of decreased productivity. The conclusions of the SLRs with similar response variables are stable.

Table 9: Comparison of the conclusions and results of SLRs with similar response variables.

	Kollanus [9]	**Bissi et al. [15] up until 2009**	**Kollanus [9] vs. Bissi et al. [15] up until 2009**
QLTY CONCLUSION	TDD might improve external quality. However, it is questionable whether TDD is the factor that really explains the results.	Most of the studies suggest an increase in external quality.	Conclusions are similar.
PROD CONCLUSION	There is moderate evidence of decreased productivity with TDD.	Of the studies, 44% suggest a decrease in productivity using TDD.	Conclusions are similar.
QLTY RESULTS	The author states that TDD might improve external quality. A total of 16 out of the 22 studies (i.e., 72%) suggest that TDD improves external quality.	According to the article and considering studies up until 2009, 75% of the studies identify a significant increase in external quality.	Results are similar.
QLTY RESULTS - Case Studies	The author states that most of the case studies quite consistently report better external quality after the implementation of TDD (13 favourable to TDD, 1 no difference).	Based on the report in the article, all the case studies, 5 in total, suggest that TDD increases external quality.	Results are similar.
QLTY RESULTS - Experiments	The author states that there is no difference in most of the controlled experiments. TDD increases quality in 2 experiments, there is no difference in 4 experiments, and TDD reduces quality in 1 experiment.	TDD increases external quality in 6 experiments, there is no difference in 1 experiment, and TDD decreases quality in 1 experiment.	Results are similar.
PROD RESULTS	The author indicates that, on the whole, the studies suggest that TDD either increases the required development effort (11) or there is not difference in the effect (7). (TDD is found to increase productivity in 5 studies).	Based on the article, considering all the studies up until 2009, 47% of the studies suggest that productivity is lower using TDD than TLD (test last development).	Results are similar.
PROD RESULTS – Case Studies	The author states that case study results are more consistent. Of all the case studies, 8 report decreased productivity and 2 do not find any difference.	Based on the classification of experiments and case studies in the article, we calculated that TDD improves productivity in 2 studies, there is no difference in 1, and TDD decreases productivity in 3 studies.	Results are different.
PROD RESULTS - Experiments	The author states that most of the controlled experiments output different results.	The experiments output contradictory results. TDD improves to productivity in 2 experiments, there is no difference in 4 experiments, and productivity drops in 4 experiments.	Results are similar.

As regards the results for external quality, 72% of the primary studies included by Kollanus [9] result in improved quality, whereas 75% of the studies referenced by Bissi et al. [15] indicate increased quality. The results of both SLRs suggest that TDD improves external quality. With respect to productivity, Kollanus [9] states that TDD might increase the required development effort, whereas Bissi et al. [15] estimate that productivity drops with TDD use in 47% of the studies. The results of both SLRs are similar. Although there are differences regarding the articles included in the SLRs with similar response variables, we observe that the results are stable with respect to quality and productivity. The SLRs by Kollanus [9] and Bissi et al. [15] classify the studies by research method. Kollanus [9] reports results by experiments and case studies. In the following, we analyse the stability of the results classified by research method: experiments and case studies. With respect to case studies, external quality improves with the use of TDD in both SLRs. Considering case studies alone, the results are similar for SLRs with similar response variables. The case studies referenced in Kollanus [9] suggest that TDD reduces productivity, whereas the case studies referenced in Bissi et al. [15] output contradictory results (TDD improves productivity in two cases studies, there is no difference in another, and quality drops with the use of TDD in the other three studies). Taking into account case studies only, the results of the SLRs with similar response variables with regard to productivity differ.

Looking at the results for experiments, Kollanus [9] states that most of the controlled experiments suggest that there is no difference in the results for quality, whereas the experiments considered in Bissi et al. [15] suggest that TDD improves quality. Considering experiments only, the results of the SLRs with similar response variables differ with regard to external quality. In the SLRs by Kollanus [9] and Bissi et al. [15], we find that the results of the experiments with respect to the effects of TDD on productivity are contradictory. With respect to productivity, the results for experiments reported in the SLRs with similar response variables are similar. Analysing the stability of the results classified by research method —case studies and experiments—, the results may differ. These results depend on the inclusion criteria used by the authors to select experiments and case studies for their SLRs.

5.3. Stability analysis of the conclusions and results with respect to all the SLRs that study TDD

This analysis took into account all eight SLRs on TDD. We carried out a comparison to identify their overlaps and differences. We then analysed the stability of the conclusions and results.

5.3.1. Comparison of SLRs

As specified in Section 4.1, the SLRs have different dates of publication ranging from 2010 to 2020. The first SLRs on TDD were conducted by Sfetsos and Stamelos [10] and Turhan et al. [11], which included primary studies published up until the year 2009. The other five publications [12][13][14][15][101] appeared later, and, for the sake of comparability, we consider only the primary studies published up until 2009. Figure 2 shows the number of primary studies by SLR. The primary studies published up until 2009 (primary studies considered in the analysis) are shaded blue and the studies published later than 2009 are shaded red. As Figure 2 shows, even though the analysis only accounts for contemporary primary studies, there are still substantial differences with respect to the number of articles included in the SLRs.

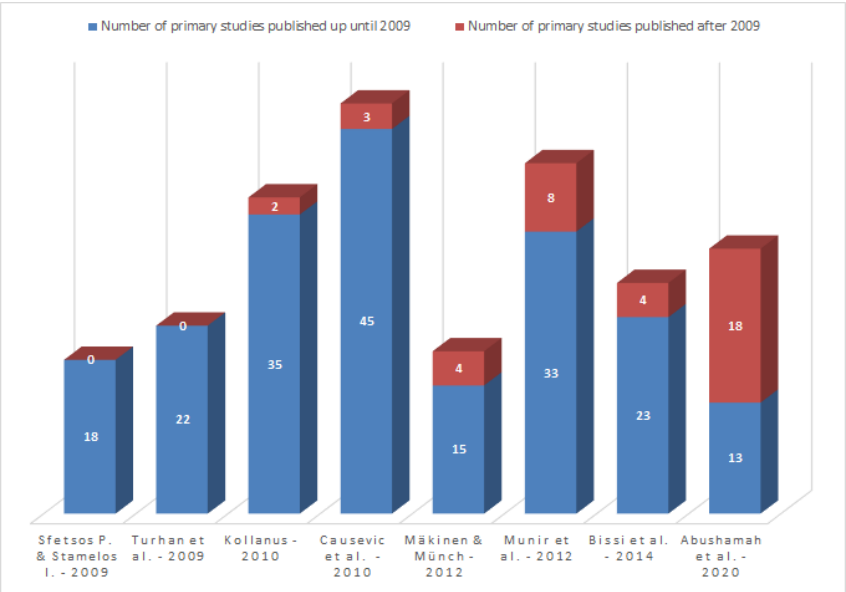

Figure 2: Number of primary studies published up until 2009 by SLR.

In order to highlight the studies that the SLRs have in common, we built Table 10. The SLRs with the biggest overlap in terms of studies are Causevic et al. [12] and Kollanus [9], which share 24 articles, Causevic et al. [12] and Munir et al. [14], which share 23 studies, and Kollanus [9] and Munir et al. [14], which share 23 studies. The SLRs with the smallest overlap in terms of studies are: Sfetsos and Stamelos [10] and Turhan et al. [11], which share seven articles, Sfetsos and Stamelos [10] and Mäkinen and Münch [13], which share eight studies, like Turhan et al. [11] and Mäkinen and Münch [13]. Abushama et al. [101] has the least number of studies in common with the other SLRs. Table 10 shows the specific studies that appear in only one SLR (shaded grey). At one end of the scale, Mäkinen and Münch [13] have no specific studies, and all their articles are shared with other SLRs, and Sfetsos and Stamelos [10] have one specific study. At the other end of the scale, Causevic et al. [12] have 17 and Munir et al. [14] have 6 specific studies, respectively, which suggests that the authors used different inclusion criteria to select the primary studies. Figure 3 shows the percentage of primary studies specific to each SLR and the number of primary studies shared by more than one SLR (2, 3, 4, 5, 6, 7 and 8 SLRs). Of the primary studies, 51% are referenced in only one SLR (Figure 3), whereas 49% are shared by two or more SLRs. None of the primary studies were included in all eight SLRs.

Table 10: Primary studies (up until 2009) shared by the SLRs on TDD.

	Sfetsos & Stamelos [10]	**Turhan et al. [11]**	**Kollanus [9]**	**Causevic et al. [12]**	**Mäkinen & Münch [13]**	**Munir et al. [14]**	**Bissi et al. [15]**	**Abushama et al. [101]**
Sfetsos & Stamelos [10]	[30]	[24],[26], [27],[31], [32],[33], [35]	[19],[21], [22],[23], [24],[25], [26],[27], [28],[29], [31],[32], [33],[34], [35],[36]	[19],[20], [21],[24], [25],[26], [27],[28], [29],[31], [32],[34], [35]	[19],[22], [26],[27], [28],[29], [32],[35]	[19],[20], [22],[24], [25],[26], [27],[28], [29],[31], [32],[33], [34],[35]	[19],[20], [23],[24], [25],[26], [28],[29], [32],[33], [34]	[19],[24]
Turhan et al. [11]		[39],[41], [42],[43]	[24],[26], [27],[31], [32],[33], [35],[37], [38],[40], [44],[45], [46],[47], [48],[49], [51]	[24],[26] [27],[31] [32],[35] [37],[38], [40],[45], [46],[47], [48],[49]	[26],[27], [32],[35], [37],[40], [46],[49]	[24],[26], [27],[31], [32],[33], [35],[37], [38],[40], [44],[46], [47],[48], [49]	[24],[26], [32],[33], [37],[40], [46],[48], [49],[50]	[24],[37], [47]
Kollanus [9]			[53],[57], [59],[60]	[19],[21], [24],[25], [26],[27], [28],[29], [31],[32],	[19],[22], [26],[27], [28],[29], [32],[35], [37],[40],	[19],[22], [24],[25], [26],[27], [28],[29], [31],[32],	[19],[23], [24],[25], [37],[40], [26],[28], [29],[32],	[19],[24], [37],[47], [54],[57]

	Sfetsos & Stamelos [10]	**Turhan et al. [11]**	**Kollanus [9]**	**Causevic et al. [12]**	**Mäkinen & Münch [13]**	**Munir et al. [14]**	**Bissi et al. [15]**	**Abushama et al. [101]**
				[34],[35], [37],[38], [40],[45], [46],[47], [48],[49], [52],[55], [56],[61]	[46],[49], [54],[55]	[33],[34], [35],[37], [38],[40], [44],[46], [47],[48], [49],[55], [61]	[33],[34], [46],[48], [49],[54], [55],[56], [61]	
Causevic et al. [12]				[63],[64], [65],[66], [68],[71], [72],[73], [74],[75], [76],[77], [79],[81], [82],[83], [84]	[19],[26], [27],[28], [29],[32], [35],[37], [40],[46], [49],[55], [58],[80]	[19],[20], [24],[25], [26],[27], [28],[29], [31],[32], [34],[35], [37],[38], [40],[46], [47],[48], [49],[55], [61],[67], [69]	[19],[20], [24],[25], [26],[28], [29],[32], [34],[37], [40],[46], [48],[49], [55],[56], [61]	[19],[24], [37],[47]
Mäkinen & Münch [13]						[19],[22], [26],[27], [28],[29], [32],[35], [37],[40], [46],[49], [55]	[19],[26], [28],[29], [32],[37], [40],[46], [49],[54], [55]	[19],[37], [47],[54]
Munir et al. [14]						[90],[91], [93],[94], [96],[98]	[19],[20], [24],[25], [26],[28], [29],[32], [33],[34], [37],[40], [46],[48],	[19],[24], [37],[95]

	Sfetsos & Stamelos [10]	Turhan et al. [11]	Kollanus [9]	Causevic et al. [12]	Mäkinen & Münch [13]	Munir et al. [14]	Bissi et al. [15]	Abushama et al. [101]
							[49],[55], [61],[95]	
Bissi et al. [15]							[100]	[19],[24], [37],[54], [95]
Abushama et al. [101]								[104],[111], [112],[119], [120]

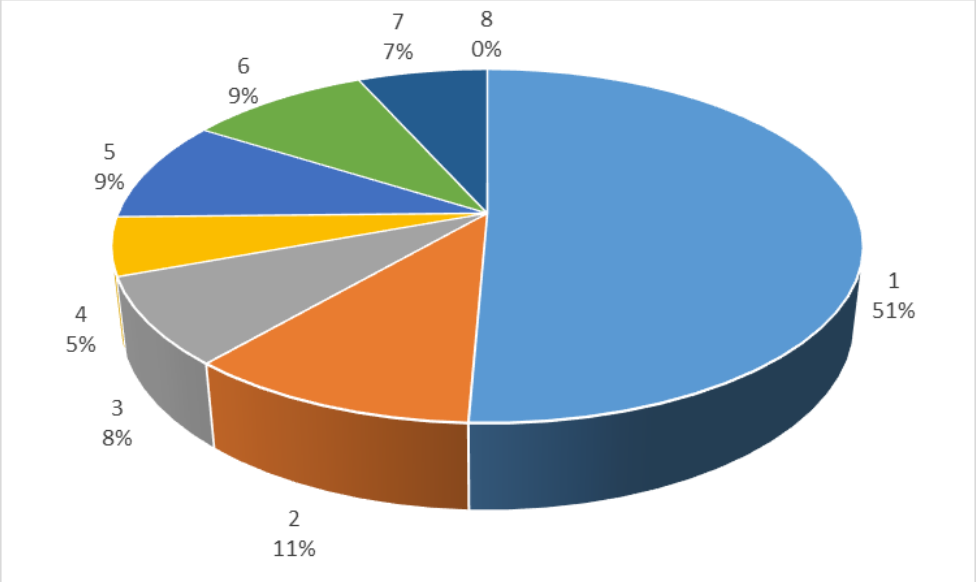

Figure 3: Percentage of specific studies and number of studies shared by the SLRs.

5.3.2. Stability of conclusions

Table 11 (rows 1 and 2) show the conclusions of the SLRs with respect to the impact of TDD on productivity and quality. The last column contains a comment on the conclusions. As regards external quality, four secondary studies (Kollanus [9], Mäkinen & Münch [13], Bissi et al. [15], and Munir et al. [14]) report similar results, stating that TDD may improve external quality, Sfetsos and Stamelos [10] claim that TDD improves external quality in the industrial setting, whereas Turhan et al. [11] suggest that the evidence is inconsistent. Causevic et al. [12] conclude that external quality is not one of the seven factors that limit the adoption of TDD in industry, whereas Abushama et al. [101] report a negative result. Therefore, the conclusions on external quality may differ. Kollanus [9], Bissi et al. [15], and Abushama et al. [101] claim that there is moderate evidence with respect to a drop in productivity. Sfetsos and Stamelos [10] state that the results are contradictory, and Turhan et al. [11] suggest that the evidence is inconsistent. Causevic et al. [12] state that the longer development time has been considered as a factor that limits the adoption of TDD, and Munir et al. [14] classify the results by different settings. The conclusions of the SLRs with regard to the effects of TDD on productivity differ. With respect to both quality and productivity, the conclusions of the SLRs may differ, and the criteria used by the authors to select evidence appear to influence the conclusions of the SLRs.

Table 11: Conclusions and results of the SLRs on TDD.

	Sfetsos & Stamelos [10]	**Turhan et al. [11]**	**Kollanus [9] up until 2009**	**Causevic et al. [12] up until 2009**	**Mäkinen & Münch [13] up until 2009**	**Munir et al. [14] up until 2009**	**Bissi et al. [15] up until 2009**	**Abushama et al. [101] up until 2009**	**Comparison of the 8 secondary studies**
QLTY CONCLUSION	The authors conclude that software quality improves in an industrial setting, while the effect in academia is unclear.	The impacts of TDD still raise many question marks. The evidence is not consistent with respect to the effects of TDD on any of the quality measures.	TDD might improve external quality. However, it is questionable whether TDD is the factor that really explains the results.	External quality is not one of the 7 factors that limit the adoption of TDD in industry.	TDD may reduce the number of defects.	The authors classify the studies by rigour and relevance and state that they reached different conclusions in each category. Therefore, the results would be biased if the studies were considered as a whole.	Most of the studies suggest an increase in external quality.	The authors state that TDD has more negative effects than TLD on external quality. TDD places more emphasis on customer commitment.	4 studies indicate that external quality can improve with the use of TDD or reduce the number of defects. 1 study claims that the use of TDD improves external quality in industry. 1 study indicates that evidence is inconsistent. 1 study focuses on conclusions with respect to limiting factors.
PROD CONCLUSION	The effects of TDD on productivity are contradictory and differ across different settings.	The impacts of TDD still raise many question marks. The evidence is not consistent with respect to the effects of TDD on any of the productivity measures.	There is moderate evidence that TDD decreases productivity.	Longer development time was considered as a factor that limits TDD adoption.	Although code maintenance may take less time, the development can take longer.	The authors classify the studies by rigour and relevance and indicate that the conclusions for each category differed. Therefore, if the studies were considered as a whole, the	44% studies indicate a drop in productivity when TDD is used.	Most studies show that TDD takes longer than TLD.	The conclusions differ. 1 study indicates that the evidence is inconsistent. 2 studies indicate there is moderate or little evidence of a drop productivity. 1 study indicates that the results are contradictory. 1 study does not consider all the

	Sfetsos & Stamelos [10]	Turhan et al. [11]	Kollanus [9] up until 2009	Causevic et al. [12] up until 2009	Mäkinen & Münch [13] up until 2009	Munir et al. [14] up until 2009	Bissi et al. [15] up until 2009	Abushama et al. [101] up until 2009	Comparison of the 8 secondary studies
						results would be biased.			studies as a whole, and they are classified by rigour and relevance. 1 study indicates that a longer development time was considered as a factor that limits TDD adoption.
QLTY RESULTS	The authors claim that the results of most of the experiments and case studies show that TDD improves external quality.	The authors state that if all the studies, controlled experiments and case studies count equally, TDD improves external quality. TDD improves external quality in 13 out of 22 studies (59%). TDD reduces external quality in 3 studies. The results are inconclusive or report no difference in 6 studies.	The author states that TDD could improve external quality. 15 out of a total of 21 studies suggest that TDD improves external quality (15 studies are equivalent to 71%).	The authors evaluate the effect of TDD on code quality and consider that code quality is not a factor that limits TDD adoption. For our part, we observed that TDD improves code quality in 12 out of 17 studies (70%). TDD has a negative effect in 1 study. There is no difference in 2 studies.	The SLR evaluates the effect of TDD on defect reduction. We observe that TDD reduces defects in 4 studies (70%), and 2 are inconclusive.	The authors state that the studies with high rigour and relevance (A) report results that show clear external quality improvements (5 favourable). The observation is similar for studies with high relevance and low rigour (B1) (4 favourable). Of the studies with low relevance and high rigour (B2), 4 report no difference, and 3 are favourable to TDD. Of the studies with low	According to the article and considering the studies up until 2009, 75% of the studies identify a significant increase in external quality.	This article includes 11 studies on external quality. TDD improves the code quality in 5 of the studies, whereas quality drops in the other 6.	The results with respect to external quality are similar: TDD increases external quality. The percentage of studies that report improvement varies.

	Sfetsos & Stamelos [10]	Turhan et al. [11]	Kollanus [9] up until 2009	Causevic et al. [12] up until 2009	Mäkinen & Münch [13] up until 2009	Munir et al. [14] up until 2009	Bissi et al. [15] up until 2009	Abushama et al. [101] up until 2009	Comparison of the 8 secondary studies
						rigour and relevance (C), 1 is favourable and 1 reports no difference.			
QLTY RESULTS – Case Studies	7 out of 8 case studies report a sharp increase in external quality.	The authors indicate that pilot and industrial studies suggest that TDD produces better external quality.	The author states that the case studies fairly consistently report better external quality after the implementation of TDD. (13 favourable to TDD, with no difference in 1).	They do not classify studies by case studies.	The authors state that several industrial case studies show a relatively large drop in defects. (We found that 4 case studies suggest that TDD reduces defects).	Category A studies include 5 case studies (71%) and 2 surveys. Studies in category B1 include 4 case studies (100%). Category A and category B1 studies report results that improve external quality.	Based on this report, we found that TDD increases external quality in all 5 case studies.	1 of 3 case studies report a positive result and the others return negative results.	Comparing the results for case studies, they are similar across the different SLRs. Only Abushama et al. [101] arrives at a contradictory conclusion.
QLTY RESULTS - Experiments	TDD improves external quality in 6 out of 8 experiments.	The authors indicate that the results of the controlled experiments are inconclusive. Of 6 experiments, 3 are inconclusive, TDD improves quality in 1 and reduces quality in 2 experiments, respectively.	The author states that there is no difference in most of the controlled experiments. TDD increases quality in 2 experiments. There is no difference in 4 experiments and TDD reduces quality in 1 experiment.	They do not classify studies by experiments.	The authors state that, in the experiments included in the SLR, TDD is no better than other development methods in terms of results (there is no difference between the effects of TDD in 2 experiments).	Studies in category B2 include 16 experiments (94%) and 1 case study. Studies in category C is 1 experiment studying quality. Category B2 studies do not report differences. There is no difference in 1 experiment in category C.	TDD increases external quality in 6 experiments, there is no difference in 1 experiment, and TDD reduces quality in 1 experiment.	The results of the TDD experiments are different. TDD increases quality in 4 experiments and reduces quality in 4 experiments.	The experiments cited in Sfetsos and Stamelos [10] and Bissi et al. [15] indicate that TDD improves quality. The results of the other SLRs suggest that there is no difference in the effects of TDD or that the results are contradictory.

	Sfetsos & Stamelos [10]	**Turhan et al. [11]**	**Kollanus [9] up until 2009**	**Causevic et al. [12] up until 2009**	**Mäkinen & Münch [13] up until 2009**	**Munir et al. [14] up until 2009**	**Bissi et al. [15] up until 2009**	**Abushama et al. [101] up until 2009**	**Comparison of the 8 secondary studies**
PROD RESULTS	The effect of TDD on productivity is contradictory, and the results differ for the different settings.	TDD does not have a consistent effect on productivity.	The author states that, as a whole, the studies appear to suggest that TDD might improve the required development effort (11) or that there is no difference in the effect (7). (We observed that TDD increases productivity in 4 studies).	The authors do not evaluate the effect of TDD on productivity. They state that the increase in development time is a factor that limits TDD use. The authors state that the results suggest that 8 primary studies report negative experiences with respect to development time (5 in industrial and 3 in academic settings). 5 studies report positive effects with respect to development time.	The results are found to be inconclusive. TDD reduces productivity in 2 studies and the other 8 are inconclusive.	In the studies with high rigour and high relevance (A), there is no difference (1 study). For studies with high relevance and low rigour (B1), there is a loss in productivity. 9 studies with low relevance and high rigour (B2) do not report any difference, and 1 is favourable to TDD. 1 study with low rigour and relevance (C) returns 1 negative result for TDD on productivity.	Based on the article considering all the studies published up until 2009, 47% of the studies indicate that productivity is lower using TDD than with TLD.	66% of the selected studies suggest that TDD has negative effects on time, and only 33% suggest a positive effect.	Results are inconclusive or contradictory.
PROD RESULTS – Case Studies	1 case study reports a decrease in productivity, and there is no significant difference in 1 study.	The pilot studies provide mixed evidence. The industrial studies suggest that productivity is worse with TDD.	The author states that the results of the case studies are more consistent. 8 case studies report a drop in productivity, and 2 studies find no difference.	They do not classify studies by case studies.	They did not consider case studies.	The category A studies do not report any difference (1 study). The category B1 studies are 4 case studies (100%) that report productivity loss.	Based on the classification made in the article, we calculated that TDD is favourable to productivity in 2 studies, there is no difference in 1 study and TDD reduces	5 case studies report that TDD reduces productivity (takes longer). TDD increases productivity (saves time) in 2 studies.	Except for [15], who report contradictory results, the results of the case studies referenced in the SLRs suggest that there is no difference or lower productivity. This difference is attributed

	Sfetsos & Stamelos [10]	Turhan et al. [11]	Kollanus [9] up until 2009	Causevic et al. [12] up until 2009	Mäkinen & Münch [13] up until 2009	Munir et al. [14] up until 2009	Bissi et al. [15] up until 2009	Abushama et al. [101] up until 2009	Comparison of the 8 secondary studies
							productivity in 3 studies.		to the fact that the referenced studies [46], [50]) were not cited by the other SLRs.
PROD RESULTS - Experiments	2 experiments report that TDD increases productivity.	The authors state that the evidence from controlled experiments suggests that productivity improves using TDD. 3 controlled experiments suggest an improvement, whereas there is no difference in 1.	The author states that most of the controlled experiments return different results.	They do not classify studies by experiments.	The results were found to be inconclusive. TDD leads to lower productivity in 2 experiments, and the results are inconclusive in 6 experiments.	The category B2 studies report no difference. The category C studies do report an effect equivalent to 1 negative result on TDD productivity.	The experiments return contradictory results. TDD improves productivity in 2 experiments, there is no difference in 4 experiments, and TDD reduces productivity in 4 experiments.	5 out of 14 TDD experiments report positive results. 9 have negative results.	Sfetsos & Stamelos [10] refer to only 2 experiments that increase productivity, Huang & Holcombe [27], Gupta & Jalote [26]), and Turhan et al. [11] confirm a productivity increase. The experiments reported by the other SLRs return contradictory or inconclusive results.

5.3.3. Stability of the results

To analyse the results, we considered all the studies, except Munir et al. [14], to be equal. As Table 11 (QLTY RESULTS row) shows, all the SLRs agree that TDD increases external quality, except Abushama et al. [101], whose TDD sample is probably too small. There are differences with respect to the percentage of studies that suggest a quality increase. For example, Turhan et al. [11] conclude that 13 out of the 22 primary studies (59%) show that TDD improves quality, whereas Bissi et al. [15] claims that 88% of the primary studies report improved quality (considering articles published up until 2009, that is, 75%). Abushama et al. [101] claim that 5 out of the 11 studies (45%) on TDD concluded that TDD improved external quality. With respect to the results of the SLRs concerning the effect of TDD on productivity (PROD RESULTS row), all the SLRs agree that the results are either inconclusive or contradictory. The results reported in the SLRs on external quality and productivity are similar. The results of the SLRs appear to be more stable than the conclusions.

5.3.4. Stability of the results by case studies

As mentioned above, several SLRs report results classified by research method. This applies to Sfetsos and Stamelos [10], Turhan et al. [11], Kollanus [9] and Bissi et al. [15]. We analyse the stability of the results with respect to case studies below. Munir et al. [14] classify primary studies by categories according to their rigour or relevance. They establish four categories: high rigour and high relevance (category A), low rigour and high relevance (category B1), high rigour and low relevance (category B2), low rigour and low relevance (category C). Considering the studies published up until 2009, we observed that: category A includes five case studies (71%) and two surveys, category B1 includes four case studies (100%), that is, case studies are predominant in A and B1. On the other hand, category B2 accommodates 16 experiments (94%) and one case study, and, finally, category C includes five experiments (83%) and one survey, that is, experiments are predominant in categories B2 and C. We use the results of categories A and B1 to analyse the results by case studies.

As Table 11 shows, taking into account only the case studies (QLTY RESULTS – Case Studies row), all the SLRs (except Abushama et al. [101]) agree that TDD improves external quality, that is, the results are similar for all SLRs. With respect to productivity (PROD RESULTS - Case Studies row), the SLRs by Sfetsos and Stamelos [10], Turhan et al. [11], Kollanus [9] and Munir et al. [14] show that either there is no difference or productivity drops. On the other hand, Bissi et al. [15] reveal that the case studies output contradictory results (two studies indicate that TDD increases productivity, three indicate that TDD reduces productivity, and one case study claims that there is no difference). The results by case studies on productivity differ. When the SLRs output results on productivity classified by case studies, the results may be influenced by the inclusion criteria used by the authors to select the case studies.

5.3.5. Stability of the results by experiments

Comparing the results of experiments on external quality (QLTY RESULTS - Experiments row), we find that the SLRs by Sfetsos and Stamelos [10] and Bissi et al. [15] agree that TDD improves quality, whereas SLRs by Turhan et al. [11], Kollanus [9], Munir et al. [14] and Abushama et al. [101] report that the effects of TDD do not differ or that the results are contradictory. The results of the SLRs by experiments on external quality differ. The stability of these results cannot be taken for granted. With respect to productivity (PROD RESULTS - Experiments row), the SLR by Sfetsos and Stamelos [10] indicates that TDD increases productivity. Turhan et al. [11] also state that TDD increases productivity. On the other hand, the experiments analysed by Kollanus [9], Munir et al. [14] (considering categories B2 and C where experiments are predominant) and Bissi et al. [15] find that the effect of TDD on productivity is contradictory or inconclusive. With respect to experiments, the SLRs may report different results, which means that these results are not reliable. The results reported by the authors classified by research method sometimes differ. The inclusion criteria used by the authors to select the primary studies in the SLR may have an influence on the results classified by experiments and case studies.

5.4. Effect of errors on the results of SLRs

Some SLRs included primary studies that have different approaches to software development with TDD, for example, they study TDD as a technique for improving testing quality, look at how to improve agile development course teaching quality or investigate the influence of TDD on spreadsheet implementation. Their inclusion appears to be the product of wrong decision making by the authors. All in all, two such articles were included in Turhan et al. [11], four in Kollanus [9], eight in Causevic et al. [12], seven in Munir et al. [14] and three in Abushama et al. [101]. They are labelled with (EXP) (shaded grey) at the side of the reference number in Table 3, which includes an explanation of the reason why these primary studies should not have been considered in the far-right column. Below, we analyse the influence that these incorrect decisions could have on the results of the SLRs on TDD. We analyse two SLRs: Kollanus [9] and Munir et al. [14]. We clustered the results of the SLRs, removing the primary studies that did not, in our opinion, focus on TDD, and observed whether the result of the SLR changed. With respect to the results on external quality, we found that the SLR

by Kollanus [9] includes 16 primary studies that specify that TDD increases quality (72%), five studies that indicate that there is no difference, and one primary study that reports that TDD decreases quality. We then deleted three primary studies that do not focus on TDD —Lui and Chan [PS65], Rahman [PS84] and Xu and Li [PS102]—, with the following results: 13 studies suggest that TDD increases quality (68%), five studies indicate that there is no difference, and one study reports that TDD reduces external quality. We find that the results do not differ substantially.

We proceeded similarly with the SLR by Munir et al. [14] for the results corresponding to category A, where seven studies indicate that TDD increases quality. If we remove the article that we consider should not have been taken into account —Bannerman & Martin [PS4]—, we find that there are six studies in favour of TDD. Again, the results are unchanged. The articles included in SLRs by wrong decision making do not appear to have a major impact on the results on external quality. With regard to productivity, five articles in the SLR by Kollanus [9] indicate that TDD increases productivity, seven that there is no difference, and 11 that TDD reduces productivity. If we remove the two studies that do not, in our opinion, focus on TDD —Zhang et al. [PS105] and Xu and Li [PS102]—, we find that three studies indicate that TDD increases productivity, seven that there is no difference, and 11 that TDD reduces productivity. Therefore, we get similar results that are unaffected by errors.

In the SLR by Munir et al. [14], there are very few articles in category A, where there is no difference with respect to the effect of TDD on productivity in one study and a negative result for TDD in two studies. If we remove the study that does not focus on TDD —Bannerman & Martin [PS4]—, we find that one study says that there is no difference with respect to the influence of TDD on productivity and one study that suggests that TDD has a negative impact. In this case, wrong decision making does have an influence, but this is not a representative result, as there are not enough primary studies on productivity. The studies included in the SLRs as a result of wrong decision making do not appear to affect the results of the SLRs on TDD with respect to the effects on quality and productivity.

6. Conclusions and Future Work

It is assumed that SLRs on TDD should be based on applicable primary studies existing in the literature. It appears, however, that authors each use their own inclusion criteria for selecting primary studies, which means that there is a sizeable difference with respect to the number of articles across SLRs. The information included in SLRs and primary studies on TDD is not comprehensive. This is an obstacle to the research. A fuller report on experiments and case studies in primary studies and a fuller description of the quality criteria, search date, bibliographic databases, and article inclusion and exclusion criteria in SLRs would facilitate research. According to our findings, SLRs with similar objectives have less overlap than SLRs with similar response variables. The SLRs with similar response variables comply better with the SLR replication principle. With regard to the differences found in the conclusions on SLRs, they are influenced by the criteria used by each author. What some authors regard as negligible consistent evidence is sufficient for other researchers to claim that TDD may have an effect on the response variable. There is no agreement or criteria for evaluating the quality of the evidence on a particular effect. Several authors report results classified by the research method. These results with respect to both case studies and experiments do not always match and do not, therefore, appear to be reliable. The results are more stable with respect to both quality and productivity when results are not classified by research method (experiment or case study).

The evidence selected in a SLR is composed of all the applicable primary studies and should be similar for all the SLRs that study one and the same topic. However, we found that the overlap between the primary studies included in all the SLRs that study TDD is no more than three per cent. The SLRs that address TDD differ with respect to the study objectives and the response variables. We found that there is a bigger overlap between SLRs with similar response variables (54%) than for SLRs with similar objectives (36%). The SLRs with similar response variables are more repeatable.

Little is known about the impact of the different overlaps between SLRs that address the same topic. The results and conclusions appear to differ depending on the criteria used by the authors to select evidence (primary studies). The conclusions of the SLRs that address the same topic are not stable: they are influenced by the results, and the results are influenced by the criteria applied by the authors with respect to the consistency of the selected evidence. When the authors report results classified by research method (case studies and experiments), the results may differ, which means that they are not very reliable. The criteria applied by authors with respect to experiment and case study selection appears to influence the results. We found that some authors wrongly selected a number of primary studies that do not focus on TDD as a development technique. However, this error does not appear to influence the results.

We now analyse the different threats that might affect the validity of the results. One potential internal validity threat to this research is the possibility of missing some SLRs. To reduce this threat, we conducted a rigorous search process by identifying the key words and composing search strings and running the search according to the instructions provided by each scientific database. Additionally, we applied inclusion and exclusion criteria for the final selection of articles. Any dispute during the final selection process was settled by negotiation in order to reduce partiality. With respect to external validity, the study is considered to be reproducible because a systematic protocol was enacted to search, collect and evaluate data. With regard to conclusion validity, we generated the tables and figures illustrated in this article directly from the data. Therefore, our results are fully traceable. The data of the secondary studies were meticulously and systematically extracted, guaranteeing a high level of reliability, as the conclusions drawn from this research can be related directly to the data and, therefore, are reproducible by other researchers.

The aim of future research is primarily to identify the criteria for evaluating the quality of the evidence gathered about an effect on a particular response variable. Additionally, it is necessary to continue the research effort to propose additional guidelines to determine the quality of SLRs.

Acknowledgements

This research was funded by grant PID2022-137846NB-I00 funded by MCIN/AEI/10.13039/501100011033, by "ERDF A way of making Europe" and the FINESSE project, Spain (PID2021-122270OB-I00). This research was also supported by the Madrid Region R&D program, Spain (project FORTE, P2018/TCS-4314) and the SATORI-UAM project (TED2021-129381B-C21).

References

[1] B. A. Kitchenham, T. Dyba, M. Jorgensen, Evidence-based software engineering, in: *Proc. 26th Intern. Conf. on Software Engineering (ICSE'04)*, Edinburgh, UK, pp. 273-281, 2004. https://doi.org/10.1109/ICSE.2004.1317449

[2] T. Dyba, B. A. Kitchenham, M. Jorgensen, Evidence-based software engineering for practitioners, *IEEE Software* 22(1):58-65, 2005. https://doi.org/10.1109/MS.2005.6

[3] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, *Technical Report* EBSE-2007-01 version 2.3, Keele University and Durham University Joint Report, UK, 2007.

[4] C. Wohlin, P. Runeson, P. A. da Mota, E. Engstrom, I. do Carmo, E. S. de Almeida, On the reliability of mapping studies in software engineering, *Journal of Systems and Software* 86(10):2594-2610, 2013. https://doi.org/10.1016/j.jss.2013.04.076

[5] S. MacDonell, M. Shepperd, B. Kitchenham, E. Mendes, How reliable are systematic reviews in empirical software engineering? *IEEE Transactions on Software Engineering* 36(5):676-687, 2010. https://doi.org/10.1109/TSE.2010.28

[6] B. Kitchenham, P. Brereton, Z. Li, D. Budgen, A. Burn, Repeatability of systematic literature reviews, in: *Proc. 15th Conf. on Evaluation & Assessment in Software Engineering (EASE'11)*, Durham, UK, pp. 46-55, 2011. https://doi.org/10.1049/ic.2011.0006

[7] K. Beck, Aim, fire, *IEEE Software* 18(5):87-89, 2001. https://doi.org/10.1109/52.951502

[8] K. Beck, C. Andres, *Extreme programming explained: Embrace change*, 2nd edition, Addison-Wesley Professional, 1999.

[9] S. Kollanus, Test-driven development - Still a promising approach? in: *Proc. 2010 Seventh Intern. Conf. on the Quality of Inform. and Communications Technology (QUATIC'10)*, Porto, Portugal, pp. 403-408, 2010. https://doi.org/10.1109/QUATIC.2010.73

[10] P. Sfetsos, I. Stamelos, Empirical studies on quality in agile practices: A systematic literature review, in: *Proc. 2010 Seventh Intern. Conf. on the Quality of Inform. and Communications Technology (QUATIC'10)*, Porto, Portugal, pp. 44-53, 2010. https://doi.org/10.1109/QUATIC.2010.17

[11] B. Turhan, L. Layman, M. Diep, H. Erdogmus, F. Shull, How effective is test-driven development, in: A. Oram, G. Wilson (eds.) *Making Software: What Really Works, and Why We Believe It*, Chapter 12, pp. 207-217, O'Reilly Press, 2010.

[12] A. Causevic, D. Sundmark, S. Punnekkat, Factors limiting industrial adoption of test-driven development: A systematic review, in: *Proc. 2011 Fourth IEEE Intern. Conf. on Software Testing, Verification and Validation*, Berlin, Germany, pp. 337-346, 2011. https://doi.org/10.1109/ICST.2011.19

[13] S. Mäkinen, J. Münch, Effects of test-driven development: A comparative analysis of empirical studies, in: D. Winkler, S. Biffl, J. Bergsmann (eds.) *Software Quality. Model-Based Approaches for Advanced Software and Systems Engineering*. SWQD 2014 pp. 155-169. Lecture Notes in Business Information Processing, vol 166. Springer, Cham, 2014. https://doi.org/10.1007/978-3-319-03602-1_10

[14] H. Munir, M. Moayyed, K. Petersen, Considering rigor and relevance when evaluating test driven development: A systematic review, *Information and Software Technology* 56(4):375-394, 2014. https://doi.org/10.1016/j.infsof.2014.01.002

[15] W. Bissi, A. G. Serra, M. C. Figueiredo, The effects of test-driven development on internal quality, external quality and productivity: A systematic review, *Information and Software Technology* 74:45-54, 2016. https://doi.org/10.1016/j.infsof.2016.02.004

[16] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering - A systematic literature review, *Information and Software Technology* 51(1):7-15, 2009. https://doi.org/10.1016/j.infsof.2008.09.009

[17] H. Zhang, M. A. Babar, Systematic reviews in software engineering: An empirical investigation, *Information and Software Technology* 55(7):1341-1354, 2013. DOI: 10.1016/j.infsof.2012.09.008

[18] B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, *Information and Software Technology* 55(12):2049-2075, 2013. https://doi.org/10.1016/j.infsof.2013.07.010

[19] T. Bhat, N. Nagappan, Evaluating the efficacy of test-driven development: Industrial case studies, in: *Proc. 2006 ACM/IEEE Intenn. Sympos. on Empirical Software Engineering (ISESE'06)*, Rio de Janeiro, Brazil, pp. 356-363, 2006. https://doi.org/10.1145/1159733.1159787

[20] L.-O. Damm, L. Lundberg, Results from introducing component-level test automation and test-driven development, *Journal of Systems and Software* 79(7): 1001-1014, 2006. https://doi.org/10.1016/j.jss.2005.10.015

[21] L.-O. Damm, L. Lundberg, Quality impact of introducing component-level test automation and test-driven development, in: In: P. Abrahamsson, N. Baddoo, T. Margaria, R. Messnarz (eds.) *Software Process Improvement*. EuroSPI 2007 pp. 187-199. Lecture Notes in Computer Science, vol 4764. Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-75381-0_17

[22] C. Desai, D. S. Janzen, J. Clements, Implications of integrating test-driven development into CS1/CS2 curricula, in: *Proc. 40th ACM Techn. Sympos. on Computer Science Education (SIGCSE'09)*, Chattanooga, Tennessee, USA, pp. 148-152, 2009. https://doi.org/10.1145/1508865.1508921

[23] S. H. Edwards, Using test-driven development in the classroom: Providing students with automatic, concrete feedback on performance, in: *Proc. Intern. Conf. on Education and Information Systems: Technologies and Applications Vol. 3 (EISTA'03)*, Orlando, Fl, USA, pp. 1-6, 2003.

[24] H. Erdogmus, M. Morisio, M. Torchiano, On the effectiveness of the test-first approach to programming, *IEEE Transactions on Software Engineering* 31(3):226-237, 2005. https://doi.org/10.1109/TSE.2005.37

[25] B. George, L. Williams, A structured experiment of test-driven development, *Information and Software Technology* 46(5): 337-342, (2004. https://doi.org/10.1016/j.infsof.2003.09.011

[26] A. Gupta, P. Jalote, An experimental evaluation of the effectiveness and efficiency of the test-driven development, in: *Proc. First Intern. Sympos. on Empirical Software Engineering and Measurement (ESEM'07)*, Madrid, Spain, pp. 285-294, 2007. https://doi.org/10.1109/ESEM.2007.41

[27] L. Huang, M. Holcombe, Empirical investigation towards the effectiveness of test first programming, *Information and Software Technology* 51(1):182-192, 2009. https://doi.org/10.1016/j.infsof.2008.03.007

[28] D. Janzen, H. Saiedian, Does test-driven development really improve software design quality? *IEEE Software* 25(2):77-84, 2008). https://doi.org/10.1109/MS.2008.34

[29] E. M. Maximilien, L. Williams, Assessing test-driven development at IBM, in: *Proc. 25th Intern. Conf. on Software Engineering (ICSE'03)*, Portland, OR, USA, pp. 564-569, 2003. https://doi.org/10.1109/ICSE.2003.1201238

[30] G. Melnik, F. Maurer, A cross-program investigation of students' perceptions of agile methods, in: *Proc. 27th Intern. Conf. on Software Engineering (ICSE'05)*, St. Louis, MO, USA, pp. 481-488, 2005. https://doi.org/10.1109/ICSE.2005.1553593

[31] M. M. Müller, O. Hagner, Experiment about Test-First programming, *IEE Proceedings - Software* 149(5):131-136, 2002. https://doi.org/10.1049/ip-sen:20020540

[32] N. Nagappan, E. M. Maximilien, T. Bhat, L. Williams, Realizing quality improvement through test driven development: Results and experiences of four industrial teams, *Empirical Software Engineering* 13(3):289-302, 2008. https://doi.org/10.1007/s10664-008-9062-z

[33] M. Pancur, M. Ciglaric, M. Trampus, T. Vidmar, Towards empirical evaluation of test-driven development in a university environment, in: *The IEEE Region 8 EUROCON 2003. Computer as a Tool*, Ljubljana, Slovenia, pp. 83-86, 2003. https://doi.org/10.1109/EURCON.2003.1248153

[34] J. C. Sanchez, L. Williams, E. M. Maximilien, On the sustained use of a test-driven development practice at IBM, in: *Proc. Agile Conference (AGILE'07)*, Washington, DC, USA, pp. 5-14, 2007. https://doi.org/10.1109/AGILE.2007.43

[35] L. Williams, E. M. Maximilien, M. Vouk, Test-driven development as a defect-reduction practice, in: *Proc. 14th Intern. Sympos. on Software Reliability Engineering (ISSRE'03)*, Denver, CO, USA, pp. 34-45, 2003. https://doi.org/10.1109/ISSRE.2003.1251029

[36] R. A. Ynchausti, Integrating unit testing into a software development team's process, in: *Proc. 2nd Intern. Conf. on Extreme Programm. and Flexible Processes in Software Engineering (XP'01)*, Sardinia, Italy, pp. 79-83, 2001.

[37] G. Canfora, A. Cimitile, F. Garcia, M. Piattini, C. A. Visaggio, Evaluating advantages of test-driven development: A controlled experiment with professionals, in: *Proc. 2006 ACM/IEEE Intern. Sympos. on Empirical Software Engineering (ISESE'06)*, Rio de Janeiro Brazil, pp. 364-371, 2006. https://doi.org/10.1145/1159733.1159788

[38] T. Flohr, T. Schneider, Lessons learned from an XP experiment with students: Test-first needs more teachings, in: J. Münch, M. Vierimaa (eds.) *Product-Focused Software Process Improvement*. PROFES 2006 pp. 305-318. Lecture Notes in Computer Science, vol 4034. Springer, Berlin, Heidelberg, 2006. https://doi.org/10.1007/11767718_26

[39] B. George, Analysis and quantification of test-driven development approach, *Master's thesis*, Computer Science, North Carolina State University, USA, 2002.

[40] A. Geras, M. Smith, J. Miller, A prototype empirical evaluation of test-driven development, in: *Proc. 10th Intern. Sympos. on Software Metrics*, Chicago, Il, USA, pp. 405-416, 2004. https://doi.org/10.1109/METRIC.2004. 1357925

[41] A. Geras, The effectiveness of test-driven development, *Master's thesis*, University of Calgary, Calgary, 2004.

[42] D. S. Janzen, An empirical evaluation of the impact of test-driven development on software quality, *Ph.D. thesis*, Computer Science, University of Kansas, USA, 1993.

[43] R. Kaufmann, D. Janzen, Implications of test-driven development: A pilot study, in: *Companion of the 18th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA'03)*, Anaheim, CA, USA, 2003, pp. 298-299, 2003. https://doi.org/10.1145/949344.949421

[44] L. Madeyski, Preliminary analysis of the effects of pair programming and test-driven development on the external code quality, in: K. Zieliński, T. Szmuc (eds.) *Software Engineering: Evolution and Emerging Technologies*. Series: Frontiers in Artificial Intelligence and Applications vol. 130, pp. 113-123, 2005.

[45] L. Madeyski, The impact of pair programming and test-driven development on package dependencies in object-oriented design - An experiment, in: J. Münch, M. Vierimaa (eds.) *Product-Focused Software Process Improvement*. PROFES 2006 pp. 278-289. Lecture Notes in Computer Science, vol 4034. Springer, Berlin, Heidelberg, 2006. https://doi.org/10.1007/11767718_24

[46] L. Madeyski, L. Szala, The impact of test-driven development on software development productivity - An empirical study, in: P. Abrahamsson, N. Baddoo, T. Margaria, R. Messnarz (eds.) *Software Process Improvement*. EuroSPI 2007 pp. 200-211, Lecture Notes in Computer Science, vol 4764. Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-75381-0_18

[47] Siniaalto, M. and P. Abrahamsson. Does test-driven development improve the program code? Alarming results from a comparative case study, in: B. Meyer, J. R. Nawrocki, B. Walter (eds.) *Balancing Agility and Formalism in Software Engineering*. CEE-SET 2007 pp. 143-156. Lecture Notes in Computer Science, vol 5082. Springer, Berlin, Heidelberg, 2008. https://doi.org/10.1007/978-3-540-85279-7_12

[48] O. P. N. Slyngstad, J. Li, R. Conradi, H. Rønneberg, E. Landre, H. Wesenberg, The impact of test-driven development on the evolution of a reusable framework of components - An industrial case study, in: *Proc. 2008 the Third Intern. Conf. on Software Engineering Advances (ICSEA'08)*, Sliema, Malta, pp. 214-223, 2008. https://doi.org/10.1109/ICSEA.2008.8

[49] J. H. Vu, N. Frojd, C. Shenkel-Therolf, D. S. Janzen, Evaluating test-driven development in an industry-sponsored capstone project, in: *Proc. 2009 Sixth Intern. Conf. on Information Technology: New Generations*, Las Vegas, NV, USA, pp. 229-234, 2009. https://doi.org/10.1109/ITNG.2009.11

[50] S. Yenduri, L. A. Perkins, Impact of using test-driven development: A case study, in: *Proc. Intern. Conf. on Software Engineering Research and Practice & Conf. on Programming Languages and Compilers (SERP'06)*, Las Vegas, Nevada, USA, pp. 126-129, 2006.

[51] L. Zhang, S. Akifuji, K. Kawai, T. Morioka, Comparison between test driven development and waterfall development in a small-scale project, in: P. Abrahamsson, M. Marchesi, G. Succi (eds.) *Extreme Programming and Agile Processes in Software Engineering*. XP 2006 pp. 211-212. Lecture Notes in Computer Science, vol 4044. Springer, Berlin, Heidelberg, 2006. https://doi.org/10.1007/11774129_29

[52] G. Canfora, A. Cimitile, F. Garcia, M. Piattini, C. A. Visaggio, Productivity of test-driven development: A controlled experiment with professionals, in: J. Münch, M. Vierimaa (eds.) *Product-Focused Software Process Improvement*. PROFES 2006 pp. 383-388. Lecture Notes in Computer Science, vol 4034. Springer, Berlin, Heidelberg, 2006. https://doi.org/10. 1007/11767718_32

[53] S. H. Edwards, Using software testing to move students from trial-and-error to reflection-in-action, in: *Proc. 35th SIGCSE Techn. Sympos. on Computer Science Education (SIGCSE'04)*, Norfolk, Virginia, USA, pp. 26-30, 2004. https://doi.org/10.1145/1028174.971312

[54] B. George, L. Williams, An initial investigation of test-driven development in industry, in: *Proc. 2003 ACM Sympos. on Applied Computing (SAC'03)*, Melbourne, Florida, USA, pp. 1135-1139, 2003. https://doi.org/10.1145/952532.952753

[55] D. S. Janzen, H. Saiedian, On the influence of test-driven development on software design, in: *Proc. 19th Conf. on Software Engineering Education Training (CSEET'06)*, Turtle Bay, HI, USA, pp. 141-148, 2006. https://doi.org/10.1109/CSEET.2006.25

[56] D. S. Janzen, C. S. Turner, H. Saiedian, Empirical software engineering in industry short courses, in: *Proc. 20th Conf. on Software Engineering Education & Training (CSEET'07)*, Dublin, Ireland, pp. 89-96, 2007. https://doi.org/10.1109/CSEET.2007.20

[57] K. M. Lui, K. C. C. Chan, Test driven development and software process improvement in China, in: J. Eckstein and H. Baumeister (eds.) *Extreme Programming and Agile Processes in Software Engineering*. XP 2004 pp. 219-222. Lecture Notes in Computer Science, vol 3092. Springer-Verlag Berlin Heidelberg, 2004. https://doi.org/10.1007/978-3-540-24853-8_27

[58] L. Madeyski, The impact of test-first programming on branch coverage and mutation score indicator of unit tests: An experiment, *Information and Software Technology* 52(2):169-187, 2010. https://doi.org/10.1016/j.infsof.2009.08.007

[59] S. M. Rahman, Applying the TBC method in introductory programming courses, in: *Proc. 2007 37th Annual Frontiers in Education Conf. – Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, Milwaukee, WI, USA, pp. T1E-20-T1E-21, 2007. https://doi.org/10.1109/FIE.2007.4418120

[60] A. Rendell, Effective and pragmatic test-driven development, in: *Proceedings of the Agile 2008 Conference (AGILE'08)*, Toronto, ON, Canada, pp. 298-303, 2008. https://doi.org/10.1109/Agile.2008.45

[61] M. Siniaalto, P. Abrahamsson, A comparative case study on the impact of test-driven development on program design and test coverage, in: *Proc. First Intern. Sympos. on Empirical Software Engineering and Measurement (ESEM'07)*, Madrid, Spain, pp. 275-284, 2007. https://doi.org/10.1109/ESEM.2007.35

[62] S. Xu, T. Li, Evaluation of test-driven development: An academic case study, in: R. Lee, N. Ishii (eds.) *Software Engineering Research, Management and Applications 2009* (pp. 229-238). Studies in Computational Intelligence, vol 253. Springer, Berlin, Heidelberg, 2009. https://doi.org/10.1007/978-3-642-05441-9_20

[63] P. Abrahamsson, A. Hanhineva, J. Jäälinoja, Improving business agility through technical solutions: A case study on test-driven development in mobile software development, in: R. L. Baskerville, L. Mathiassen, J. Pries-Heje, J. I. DeGross (eds.) *Business Agility and Information Technology Diffusion*. TDIT 2005 pp. 227. IFIP International Federation for Information Processing, vol 180. Springer, Boston, MA, 2005. https://doi.org/10.1007/0-387-25590-7_14

[64] L. Cao, B. Ramesh, Agile requirements engineering practices: An empirical study, *IEEE Software* 25(1):60-67, 2008. https://doi.org/10.1109/MS.2008.1

[65] L.-R. Chien, D. J. Buehrer, C.-Y. Yang, C.-M. Chen, An evaluation of TDD training methods in a programming curriculum, in: *Proc. IEEE Intern. Sympos. on IT in Medicine and Education (ITME'08)*, Xiamen, Fujian, China, pp. 660-665, 2008. https://doi.org/10.1109/ITME.2008.4743948

[66] M. A. Domino, R. W. Collins, A. R. Hevner, C. F. Cohen, Conflict in collaborative software development, in: *Proc. 2003 SIGMIS Conf. on Computer Personnel Research: Freedom in Philadelphia-Leveraging Differences and Diversity in the IT Workforce, (SIGMIS CPR'03)*, Philadelphia, Pennsylvania, USA, pp. 44-51, 2003. https://doi.org/10.1145/761849.761856

[67] M. A. Domino, R. W. Collins, A. R. Hevner, Controlled experimentation on adaptations of pair programming, *Information Technology and Management* 8(4):297-312, 2007. https://doi.org/10.1007/s10799-007-0016-8

[68] W. P. Paula Filho, Quality gates in use-case driven development, in: *Proc. 2006 Intern. Workshop on Software Quality (WoSQ'06)*, Shanghai, China, pp. 33-38, 2006. https://doi.org/10.1145/1137702.1137710

[69] T. Flohr, T. Schneider, An XP experiment with students - Setup and problems, in: F. Bomarius, S. Komi-Sirviö (eds.) *Product Focused Software Process Improvement*. PROFES 2005 pp. 474-486. Lecture Notes in Computer Science, vol 3547. Springer, Berlin, Heidelberg, 2005. https://doi.org/10.1007/11497455_37

[70] A. Höfer, M. Philipp, An empirical study on the TDD conformance of novice and expert pair programmers, in: P. Abrahamsson, M. Marchesi, F. Maurer (eds.) *Agile Processes in Software Engineering and Extreme Programming*. XP 2009 pp. 33-42. Lecture Notes in Business Information Processing, vol 31. Springer-Verlag Berlin Heidelberg, 2009. https://doi.org/10.1007/978-3-642-01853-4_6

[71] L. Huang, C. Thomson, M. Holcombe, How good are your testers? an assessment of testing ability, in: *Proc. Testing: Academic and Industrial Conf. Practice and Research Techniques – MUTATION (TAICPART-MUTATION'07)*, Windsor, UK, pp. 82-88, 2007. https://doi.org/10.1109/TAIC.PART.2007.16

[72] O. Kobayashi, M. Kawabata, M. Sakai, E. Parkinson, Analysis of the interaction between practices for introducing XP effectively, in: *Proc. 28th Intern. Conf. on Software Engineering (ICSE'06)*, Shanghai, China, 2006, pp. 544-550, 2006. https://doi.org/10.1145/1134285.1134361

[73] S. Kollanus, V. Isomöttönen, Understanding TDD in academic environment: Experiences from two experiments, in: *Proc. 8th Intern. Conf. on Computing Education Research (Koli'08)*, Koli, Finland, 2008, pp. 25-31, 2008. https://doi.org/10.1145/1595356.1595362

[74] L. Layman, L. Williams, L. Cunningham, Motivations and measurements in an agile case study, *Journal of Systems Architecture* 52(11):654-667, 2006. https://doi.org/10.1016/j.sysarc.2006.06.009

[75] N. F. LeJeune, Teaching software engineering practices with extreme programming, *Journal of Computing Sciences in Colleges* 21(3):107-117, 2006.

[76] L. Madeyski, On the effects of pair programming on thoroughness and fault-finding effectiveness of unit tests, in: J. Münch, P. Abrahamsson (eds.) *Product-Focused Software Process Improvement*. PROFES 2007 pp. 207-221. Lecture Notes in Computer Science, vol 4589. Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-73460-4_20

[77] L. Madeyski, Impact of pair programming on thoroughness and fault detection effectiveness of unit test suites, *Software Process: Improvement and Practice* 13(3):281-295, 2008. https://doi.org/10.1002/spip.382

[78] A. Marchenko, P. Abrahamsson, T. Ihme, Long-term effects of test-driven development a case study, in: P. Abrahamsson, M. Marchesi, F. Maurer (eds.) *Agile Processes in Software Engineering and Extreme Programming*. XP 2009 pp. 13-22. Lecture Notes in Business Information Processing, vol 31. Springer, Berlin, Heidelberg, 2009. https://doi.org/10.1007/978-3-642-01853-4_4

[79] V. B. Mišić, Perceptions of extreme programming: An exploratory study, *ACM SIGSOFT Software Engineering Notes* 31(2):1-8, 2006. https://doi.org/10.1145/1118537.1118542

[80] M. M. Muller, A. Höfer, The effect of experience on the test-driven development process, *Empirical Software Engineering* 12(6):593-615, 2007. https://doi.org/10.1007/s10664-007-9048-2

[81] O. Salo, P. Abrahamsson, An iterative improvement process for agile software development, *Software Process: Improvement and Practice* 12(1):81-100, 2007. https://doi.org/10.1002/spip.305

[82] P. Sfetsos, L. Angelis, I. Stamelos, Investigating the extreme programming System - An empirical study, *Empirical Software Engineering* 11(2):269-301, 2006. https://doi.org/10.1007/s10664-006-6404-6

[83] L. B. Sherrell, J. J. Robertson, Pair programming and agile software development: Experiences in a college setting, *Journal of Computing Sciences in Colleges* 22(2):145-153, 2006.

[84] H. Wasmus, H.-G. Gross, Evaluation of test-driven development: An industrial case study, in: *Proc. Second Intern. Conf. on Evaluation of Novel Approaches to Software Engineering (ENASE'07)*, Barcelona, Spain, pp. 103-110, 2007. https://doi.org/10.5220/0002584401030110

[85] T. Dogša, D. Batić, The effectiveness of test-driven development: An industrial case study, *Software Quality Journal* 19(4):643-661, 2011. https://doi.org/10.1007/s11219-011-9130-2

[86] M. Pancur, M. Ciglaric, Impact of test-driven development on productivity, code and tests: A controlled experiment, *Information and Software Technology* 53(6):557-573, 2011. https://doi.org/10.1016/j.infsof.2011.02.002

[87] J. W. Wilkerson, J. F. Nunamaker, R. Mercer, Comparing the defect reduction benefits of code inspection and test-driven development, *IEEE Transactions on Software Engineering* 38(3):547-560, 2012. https://doi.org/10.1109/TSE.2011.46

[88] M. F. Aniche, M. A. Gerosa, Most common mistakes in test-driven development practice: Results from an online survey with developers, in: *Proc. 2010 Third Intern. Conf. on Software Testing, Verification, and Validation Workshops (ICSTW'10)*, IEEE, 2010, pp. 469-478, 2007. https://doi.org/10.1109/ICSTW.2010.16

[89] S. Bannerman, A. Martin, A multiple comparative study of test-with development product changes and their effects on team speed and product quality, *Empirical Software Engineering* 16(2):177-210, 2011. https://doi.org/10.1007/s10664-010-9137-5

[90] L. Crispin, Driving software quality: How test-driven development impacts software quality, *IEEE Software* 23(6):70-71, 2006. https://doi.org/10.1109/MS.2006.157

[91] C. Desai, D. Janzen, K. Savage, A survey of evidence for test-driven development in academia, *ACM SIGCSE Bulletin* 40(2):97-101, 2008. https://doi.org/10.1145/1383602.1383644

[92] M. Ficco, R. Pietrantuono, S. Russo, Bug localization in test-driven development, *Advances in Software Engineering* 2011:1-18, article 492757, 2011. https://doi.org/10.1155/2011/492757

[93] L. Huang, M. Holcombe, Empirical assessment of test-first approach, in: *Proc. Testing: Academic Industrial Conf. - Practice and Research Techniques (TAIC PART'06)*, Windsor, UK, pp. 197-202, 2006. https://doi.org/10.1109/TAIC-PART. 2006.7

[94] D. S. Janzen, H. Saiedian, A leveled examination of test-driven development acceptance, in: *Proc. 29th Intern. Conf. on Software Engineering (ICSE'07)*, Minneapolis, MN, USA, pp. 719-722, 2007. https://doi.org/10.1109/ICSE.2007.8

[95] D. Janzen, H. Saiedian, Test-driven learning in early programming courses, in: *Proc. 39th SIGCSE Techn. Sympos. on Computer Science Education (SIGCSE'08)*, Portland, Oregon, USA, pp. 532-536, 2008. https://doi.org/10.1145/1352135.1352315

[96] S. Kollanus, V. Isomöttönen, Test-driven development in education: Experiences with critical viewpoints, in: *Proc. 13th Annual Conf. on Innovation and Technology in Computer Science Education (ITiCSE'08)*, Madrid, Spain, pp. 124-127, 2008. https://doi.org/10.1145/1384271.1384306

[97] N. Laranjeiro, M. Vieira, Extending test-driven development for robust web services, in: *Proc. 2009 Second Intern. Conf. on Dependability (DEPEND'09)*, Athens, Greece, pp. 122-127, 2009. https://doi.org/10.1109/DEPEND.2009.25

[98] K. McDaid, A. Rust, B. Bishop, Test-driven development: Can it work for spreadsheets? in: *Proc. 4th Intern. Workshop on End-User Software Engineering (WEUSE'08)*, Leipzig, Germany, pp. 25-29, 2008. https://doi.org/10.1145/1370847.1370853

[99] M. F. Aniche, M. A. Gerosa, How the practice of TDD influences class design in object-oriented systems: Patterns of unit tests feedback, in: *Proc. 2012 26th Brazilian Sympos. on Software Engineering (SBES'12)*, Natal, Brazil, pp. 1-10, 2012. https://doi.org/10.1109/SBES.2012.14

[100] I. Turnu, M. Melis, A. Cau, A. Setzu, G. Concas, K. Mannaro, Modeling and simulation of open-source development using an agile practice, *Journal of Systems Architecture* 52(11):610-618, 2006. https://doi.org/10.1016/j.sysarc.2006.06.005

[101] H. M. Abushama, H. A. Alassam, F. A. Elhaj, The effect of test-driven development and behavior-driven development on project success factors: A systematic literature review-based study, in: *Proc. 2020 Intern. Conf. on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE'20)*, Khartoum, Sudan, pp. 1-9, 2020. https://doi.org/10.1109/ICCCEEE49695.2021.9429593

[102] Y. Rafique, V. B. Mišić, The effects of test-driven development on external quality and productivity: A meta-analysis, *IEEE Transactions on Software Engineering* 39(6):835-856, 2013. https://doi.org/10.1109/TSE.2012.28

[103] D. Fucci, H. Erdogmus, B. Turhan, M. Oivo, N. Juristo, A dissection of the test-driven development process: Does it really matter to test-first or to test-last? *IEEE Transactions on Software Engineering* 43(7):597-614, 2017. https://doi.org/10.1109/TSE.2016.2616877

[104] D. Saff, M. D. Ernst, An experimental evaluation of continuous testing during development, in: *Proc. 2004 ACM SIGSOFT Intern. Sympos. on Software Testing and Analysis (ISSTA'04)*, Boston, MA, USA, pp. 76-85, 2004. https://doi.org/10.1145/1007512.1007523

[105] J. M. Chimento, W. Ahrendt, G. Schneider, Testing meets static and runtime verification, in: *Proc. 2018 IEEE/ACM 6th Intern. FME Workshop on Formal Methods in Software Engineering (FormaliSE'18)*, Gothenburg, Sweden, pp. 30-39, 2018. https://doi.org/10.1145/3193992.3194000

[106] N. Agarwal, P. Deep, Obtaining better software product by using test first programming technique, in: *Proc. 2014 5th Intern. Conf.-Confluence the Next Generation Information Technology Summit (Confluence'14)*, Noida, India, pp. 742-747, 2014. https://doi.org/10.1109/CONFLUENCE.2014.6949233

[107] R. Chaipraserth, A. Leelasantitham, S. Kiattisin, A test automation framework in POCT system using TDD techniques, in: *Proc. 2013 13th Intern. Sympos. on Communications and Information Technologies (ISCIT'13)*, Surat Thani, Thailand, pp. 600-604, 2013. https://doi.org/10.1109/ISCIT.2013.6645931

[108] O. Dagenais, D. Deugo, TODD: Test-oriented development and debugging, in: *Proc. 5th ACIS Intern. Conf. on Software Engineering Research, Management & Applications (SERA'07)*, Busan, Korea (South), pp. 839-846, 2007. https://doi.org/10.1109/SERA.2007.127

[109] M. Rilee, T. Clune, Towards test driven development for computational science with pFUnit, in: *Proc. 2nd Intern. workshop on Software Engineering for High Performance Computing in Computational Science and Engineering (SE-HPCCSE'14)*, New Orleans, Louisiana, USA, pp. 20-27, 2014. https://doi.org/10.1109/SE-HPCCSE.2014.5

[110] R. Swamidurai, B. Dennis, U. Kannan, Investigating the impact of peer code review and pair programming on test-driven development, in: *Proc. IEEE SOUTHEASTCON'14*, Lexington, KY, USA, pp. 1-5, 2014. https://doi.org/10.1109/SECON.2014.6950664

[111] B. S. Mattu, R. Shankar, Test driven design methodology for component-based system, in: *Proc. 2007 1st Annual IEEE Systems Conf.*, Honolulu, HI, USA, pp. 1-7, 2007. https://doi.org/10.1109/SYSTEMS.2007.374646

[112] T. Hadzic, H. R. Andersen, A BDD-based polytime algorithm for cost-bounded interactive conFiguretion, in: *Proc. 21st National Conf. on Artificial Intelligence (AAA'06)*, Boston, Massachusetts, USA, pp. 62-67, 2006.

[113] A. Scandaroli, R. Leite, A. H. Kiosia, S. A. Coelho, Behavior-driven development as an approach to improve software quality and communication across remote business stakeholders, developers and QA: Two case studies, in: *Proc. 14th Intern. Conf. on Global Software Engineering (ICGSE'19)*, Montreal, QC, Canada, pp. 105-110. 2019. https://doi.org/10.1109/ICGSE.2019.00030

[114] H. Munir, W. Krzysztof, K. Petersen, M. Moayyed, An experimental evaluation of test-driven development vs. test-last development with industry professionals, in: *Proc. 18th Intern. Conf. on Evaluation and Assessment in Software Engineering (EASE'14)*, London, England, UK, article 50, pp. 1-10, 2014. https://doi.org/10.1145/2601248.2601267

[115] T. A. Majchrzak, A. Simon, Using spring Roo for the test-driven development of Web applications, in: *Proc. 27th Annual ACM Sympos. on Applied Computing (SAC'12)*, Riva del Garda, Italy, pp. 664-671, 2012. https://doi.org/10.1145/2245276.2245404

[116] D. Fucci, S. Romano, M. T. Baldassarre, D. Caivano, G. Scanniello, B. Turhan, N. Juristo, A longitudinal cohort study on the retainment of test-driven development, in: *Proc. 12th ACM/IEEE Intern. Sympos. on Empirical Software Engineering and Measurement (ESEM'18)*, Oulu, Finland, article 18, pp. 1-10, 2018. https://doi.org/10.1145/3239235.3240502

[117] D. Fucci, B. Turhan, M. Oivo, Conformance factor in test-driven development: initial results from an enhanced replication, in: *Proc. 18th Intern. Conf. on Evaluation and Assessment in Software Engineering (EASE'14)*, London, England, UK, article 22, pp. 1-4, 2014. https://doi.org/10.1145/2601248.2601272

[118] H. Bünder, H. Kuchen, A model-driven approach for behavior-driven GUI testing, in: *Proc. 34th ACM/SIGAPP Sympos. on Applied Computing (SAC'19)*, Limassol, Cyprus, pp. 1742-1751, 2019. https://doi.org/10.1145/3297280.3297450

[119] M. M. Müller, F. Padberg, On the economic evaluation of XP projects, in: *Proc. 9th European Software Engineering Conf. Held Jointly with 11th ACM SIGSOFT International Symposium on Foundations of Software Engineering (ESEC/FSE'11)*, Helsinki, Finland, pp. 168- 177, 2003. https://doi.org/10.1145/940071.940094

[120] M. Karlesky, G. Williams, W. Bereza, M. Fletcher, Mocking the embedded world: Test-driven development, continuous integration, and design patterns, in: *Proc. Embedded Systems Conf. Silicon Valley (ESC'07)*, San Jose, CA, USA, pp. 1-15, 2007.

[121] D. Fucci, B. Turhan, N. Juristo, O. Dieste, A. Tosun-Misirli, M. Oivo, Towards an operationalization of test-driven development skills: An industrial empirical study, *Information and Software Technology* 68:82-97, 2015. https://doi.org/10.1016/j.infsof.2015.08.004

[122] A. Tosun, O. Dieste, D. Fucci, S. Vegas, B. Turhan, H. Erdogmus, A. Santos, M. Oivo, K. Toro, J. Järvinen, N. Juristo, An industry experiment on the effects of test-driven development on external quality and productivity, *Empirical Software Engineering* 22:2763-2805, 2017. https://doi.org/10.1007/s10664-016-9490-0

[123] M. Kulas, J. L. Borelli, W. Gässler, D. Peter, S. Rabien, G. Orban de Xivry, L. Busoni, M. Bonaglia, T. Mazzoni, G. Rahmer, Practical experience with test-driven development during commissioning of the multi-star AO system ARGOS, in: *Proc. SPIE Vol. 9152, Software and Cyberinfrastructure for Astronomy III*, Montreal, QC, Canada, pp. 1-10. 2014. https://doi.org/10.1117/12.2056218

[124] A. Santos, J. Järvinen, J. Partanen, M. Oivo, N. Juristo, Does the performance of TDD hold across software companies and premises? A group of industrial experiments on TDD, in: M. Kuhrmann, K. Schneider, D. Pfahl, S. Amasaki, M. Ciolkowski, R. Hebig, P. Tell, J. Klünder, S. Küpper (eds.) *Product-Focused Software Process Improvement*. PROFES 2018 pp. 227-242. Lecture Notes in Computer Science, vol 11271. Springer, Cham, 2018. https://doi.org/10.1007/978-3-030-03673-7_17