Agile Testing: A Systematic Mapping Across Three Conferences

Understanding Agile Testing in the XP/Agile Universe, Agile, and XP Conferences

Theodore D. Hellmann, Apoorve Chokshi, Zahra Shakeri Hossein Abad, Sydney Pratte, Frank Maurer
Department of Computer Science
University of Calgary
Calgary, Alberta, Canada
{tdhellma, apoorve.chokshi, zshakeri, sapratte, frank.maurer}@ucalgary.ca

***Abstract*— Unit and acceptance testing are central to agile software development, but is that all there is to agile testing? We build on previous work to provide a systematic mapping of agile testing publications at major agile conferences. The analysis presented in this paper allows us to answer research questions like: what is agile testing used for; what types of studies on agile testing have been published; what problems do people have when performing agile testing; and what benefits do these publications offer? We additionally explore topics such as: who are the major authors in this field; in which countries do these authors work; what tools are mentioned; and is the field driven by academics, practitioners, or collaborations? This paper presents our analysis of these topics in order to better structure future work in the field of agile testing and to provide a better understanding of what this field actually entails.**

Keywords-agile software development, software testing, systematic mapping, empirical, test-driven development, testing tools.

I. INTRODUCTION

Effective software testing is one of the cornerstones of agile software development. In fact, most other agile practices rely on some form of testing [1]. In that citation, testing in an agile context means one of two things: unit testing or acceptance testing. But is this really all there is to testing within agile? In order to find out more about what testing means within agile, we decided to perform a systematic mapping study of agile testing.

Systematic mapping studies take a large number of papers as input and categorize them based on their titles and abstracts. The result of such a study is a framework for understanding a field of research at a high level – in other words, to find out what topics a given field encompasses. These studies are gaining popularity within software engineering research, but must be approached with caution [2] due to concerns about their repeatability. Last year, we performed a systematic mapping to determine what the term "agile testing" meant within the field of software engineering as a whole [3] using a primarily automated search for papers. While the results were insightful, we feel that it is important to approach this topic again from a slightly different angle. Where last year we searched all software engineering publications for papers to include, this year we restricted our initial search for papers to major agile conferences: the (now defunct) XP/Agile Universe, Agile, and XP conferences. Additionally, while the automated approach we used last year turned up 166 related papers, we felt that it could have missed important papers. Using a manual approach for this year's study, we came up with a new set of 110 papers for analysis - only 17 of which were present in last year's paperset. This makes our analysis for this year's study very relevant over and above last year's analysis, as our approach identified a significant number of new papers. However, it also identifies serious issues with the validity and reliability of systematic mappings and reviews. These are addressed in Section VI.B.

Our goal with this paper is to present a summary of what the field of agile testing is about. We accomplish this by asking research questions including: what has agile testing been used for; what types of research have been published; what problems are associated with agile testing; and what benefits do publications claim to offer? As a secondary analysis, we also answer several additional research questions: which authors lead the field; which countries produce the most research on this topic; which tools are most frequently mentioned; and do academics or practitioners (or collaborations) contribute more publications to the field?

II. BACKGROUND

Meta-analyses, such as systematic literature reviews and systematic mappings, have been carried out to investigate the state of the art for agile software development and related fields, such as test-driven development (TDD). Dybå and Dingsøyr [4] present a systematic review of empirical studies of agile software development and present implications for research as well as industry. Their review indicated a need for more empirical studies that focus on methods other than eXtreme Programming (XP). However, their review explicitly did not focus on agile testing.

In the meta-analysis by Shull et al. [5], the effectiveness of TDD is examined in terms of delivered quality, internal code quality, and productivity. However, the results reported in these meta-analyses are mixed. Reviewing the evidence for adoption of TDD, Causevic et al. [6] found the variation in reported results problematic for comparing between studies. This was also a found by Dybå and Dingsøyr [4], who identified large variation in how empirical studies were conducted and reported. Jeffries and Melnik [7] conducted a review of selected empirical studies on the effect of TDD on quality and effort. They found that TDD largely resulted in an increase in quality, but one study they identified showed

978-0-7695-5076-3/13 $26.00 © 2013 IEEE
DOI 10.1109/AGILE.2013.10

instead that TDD resulted in a strong negative impact on quality. Additionally, while they showed that TDD could reduce the amount of effort required by up to 27%, most studies found an increase in effort of up to 100%.

A major consideration with meta-analyses, such as the ones referenced above, are issues of repeatability. Kitchenham et al. [8] carried out a case study to investigate the extent to which systematic literature reviews are repeatable. They found that the experience level of researchers impacted whether comparable sets of studies are selected for review, which subsequently impacted conclusions drawn by the researchers. For these reasons, it becomes imperative that methods used in systematic literature reviews and systematic mapping studies are clearly and transparently presented [8].

Following the advice by Kitchenham et al. [8], we followed the methods by Petersen et al. [9] and Dybå and Dingsøyr [4] in order to maintain transparency in our methods. However, we modified these approaches to use a manual search for papers. This allowed us to use a much more detailed search strategy than the one employed in last year's study.

III. Methodology

This study extends and builds upon previous work [3] which was in turn based on the guidelines provided by Petersen et al. [9]. This section provides details on how we conducted each step of the present systematic mapping. At least two of the authors participated in each step of the analysis of publications in order to allow us to immediately resolve disagreements as to how publications should be keyworded through discussion.

A. Define Research Questions

Specific research questions are important for constraining the scope of a systematic mapping study. For this study, we wished to investigate what agile testing means with respect to research papers published at major agile software development conferences. We defined four primary research questions to guide this investigation:

PRQ1. What has agile testing been used for?
PRQ2. What types of papers on agile testing have been published?
PRQ3. What problems with agile testing do publications mention?
PRQ4. What benefits do publications propose their research will produce?

We also chose to investigate several secondary research questions. These questions are no less important than our primary research questions, but the analysis required to answer them is more cursory. These questions are:

SRQ1. Which authors lead the field of agile testing?
SRQ2. In which countries do the leading authors work?
SRQ3. What tools have been used for agile testing?
SRQ4. Is the field led by practitioners, academics, or collaborations?

B. Conduct Search

After defining these research questions, the next step in our research was to conduct a search for publications relevant to our research questions. In last year's study, we chose to do this automatically by performing an automated search of the major paper repositories IEEE Xplore and SciVerse Scopus. However, this search did not turn up as many papers from leading conferences on agile development methodologies as could have been expected. Based on this, we decided to conduct the search for this study by manually searching through the XP/Agile Universe, Agile, and XP conference proceedings to identify papers on the topic of software testing.

The total number of papers in these conferences between the years 2002 and 2012 was 993 – almost exactly 100 papers more than the 894 initially discovered for last year's set. The breakdown of these papers by conference can be seen in Figure 1.

C. Screening Papers

After collecting this initial set of 993 papers, the next step was to screen out those publications not relevant to our research questions. This was done in two passes. During the first pass, papers were eliminated on the basis of their titles.

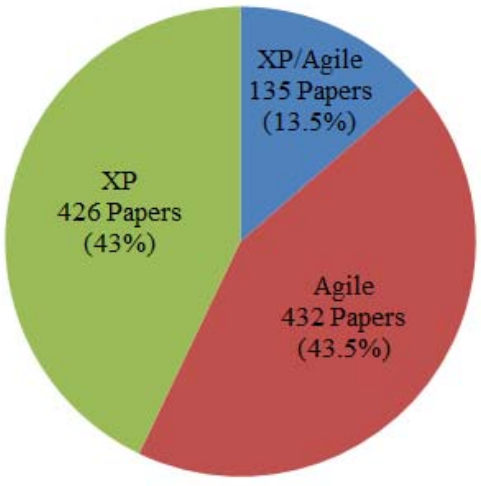

Figure 1: Breakdown of initial paperset by conference.

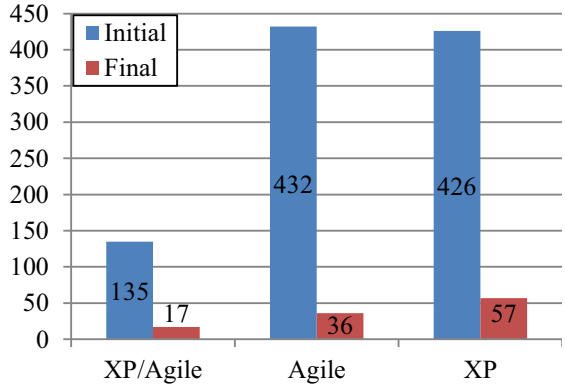

Figure 2: Initial and final paperset by conference.

For inclusion, papers titles needed to make reference to testing or to an implicitly testing-centric process (like continuous integration). Papers could also be excluded if they did not relate to testing within the context of software development. After this step, our paperset included 139 papers. For the second step, papers were eliminated on the basis of their type. Where it was unclear whether a paper should be included, the abstract was also consulted. Because we wished to focus on research, we excluded workshop proposals, tutorials, demos, and other non-research papers. During this pass, a further 29 papers were eliminated to leave us with our final set of 110 (compare to the 166 from last year; only 17 papers are duplicated between these sets). The breakdown of these final papers by conference can be seen in Figure 2.

D. Keywording Using Abstracts

After we finished screening papers by title, we keyworded each paper based on its title and abstract in order to develop a framework for understanding the field of agile testing. This keywording was done using both open and closed coding – as in grounded theory [10]. In open coding, the keywords used to describe a paper are drawn out of the source material itself. This approach was used for PRQs 1, 3, and 4. In closed coding, the list of keywords is pre-defined and keywords are simply picked from the list (see Section IV.B) to apply to papers. This approach was used for PRQ2. These keywords were then used as the basis for the rest of our analysis.

E. Data Extraction and Mapping

We used EndNote (www.endnote.com) to manage citations and keep track of the keywords applied to our paperset. EndNote makes grouping of keywords easier, but is not ideal for analysis. The analysis and visualization preparation was done in Excel (office.microsoft.com/excel).

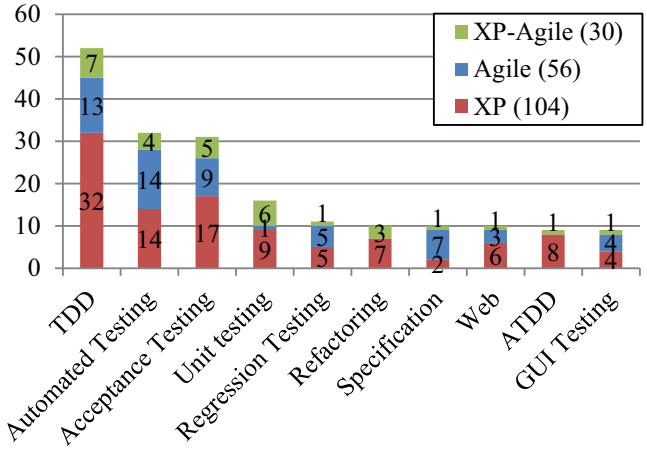

Figure 3: Breakdown of topics by conference.

IV. Primary Results

A. What has agile testing been used for?

During the course of keywording our paperset, we identified 51 different keywords relating to the uses to which testing was put. Due to space limitations, we limit the present discussion to the top 10 most frequently-occurring keywords. The frequency of these keywords across the three conferences can be seen in Figure 2.

Unsurprisingly, nearly half (52) of our 110 papers discussed test-driven development. Over a third discussed automated testing or acceptance testing (32 and 31, respectively). These topics are incredibly important to agile testing to the point where, when speaking of tests in an agile context, it's often unnecessary to specify that the tests are automated (outside the special case of GUI testing). Surprisingly, however, only 16 papers specifically mentioned unit testing. This may be because, as with automated testing, tests are assumed to be at the unit level unless otherwise

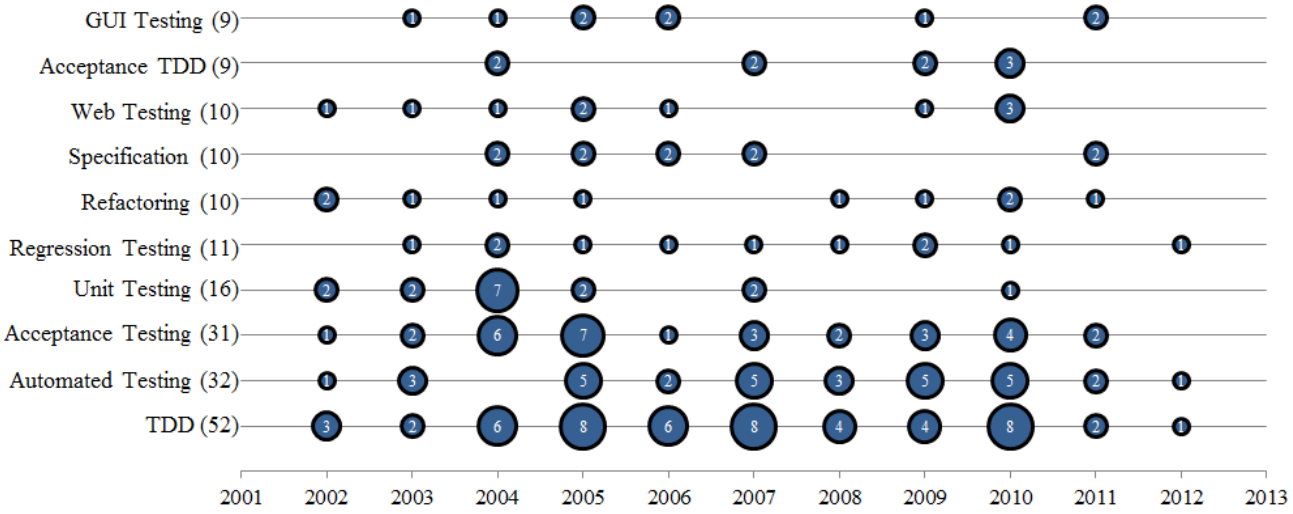

Figure 4: Breakdown of uses of testing by year.

specified. Similarly, regression testing was mentioned only 11 times, yet this is the use to which tests are normally put in an agile context. Using tests for refactoring specifically, using tests to specify features for future development, and writing tests for web systems tied with 10 occurrences each, while acceptance test driven development (ATDD) and GUI testing close out the top 10 with 9 occurrences each. Keywords from these top 10 occurred 190 times, while all other keywords for PRQ1 only occurred 92 times together.

A breakdown of these results by conference can be seen in Figure 3, while a breakdown of these topics by year can be seen in Figure 4. Figure 3 shows us how focus on topics varies between conferences. For example, papers from XP are more likely to focus on the topic of TDD, while papers on ATDD and use of tests for refactoring have not been presented at Agile. Figure 4, on the other hand, shows us that interest in all topics has dropped sharply after 2010. Interest in unit testing spiked in 2004, but has been almost completely absent since that time. Similarly, while interest in acceptance testing was high in 2004 and 2005, it has been far less prominent recently. Interest in GUI testing has been consistently low – only three publications in the past six years – despite GUI testing remaining a difficult, expensive process.

There are some similarities between the results from the previous and current study. In that study, the top 10 terms (in order of decreasing frequency) were: TDD, specification, unit testing, acceptance testing, web testing, formal specification, continuous integration, functional testing, ATDD, and a three-way tie for 10th between database, GUI, and performance testing. Since 7 of these topics are also highly important in this study, this reinforces the validity of the results in both papers. However, the remaining results do not appear in the top 10 for this paper, which shows that there is some variability between the papers published to these agile conferences and the field as a whole.

B. *What types of papers on agile testing have been published?*

In [11], five basic types of paper are identified that we reuse for the present study:

- Solution – a novel solution is proposed, but only a proof of concept (if any) is offered
- Validation – further investigates a solution, but is not evaluated in practice or rigorously
- Philosophical – provides a new framework for understanding a field
- Opinion – presents the author's personal opinions without much evidence to back up claims
- Experience – describes the author's experience on a project in industry
- Evaluation – investigation of a problem in practice with a large, rigorous study

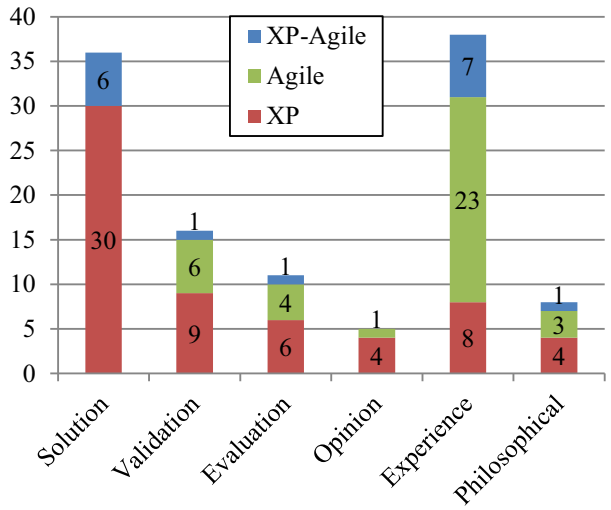

Figure 5: Breakdown of topics by conference.

We applied these paper types as keywords to each publication in our paperset in order to find out what kinds of research were used most frequently in this field. The results are broken down by conference in Figure 5 and by year in Figure 6.

From Figure 5, we can see that almost all publications accepted to XP fall into the Solution category; almost all papers at Agile fall into the Experience category. XP/Agile Universe was an even mix of those two types. Given the distinct separation between the types of papers Agile and XP attract, it would be in the interest of the field of agile testing

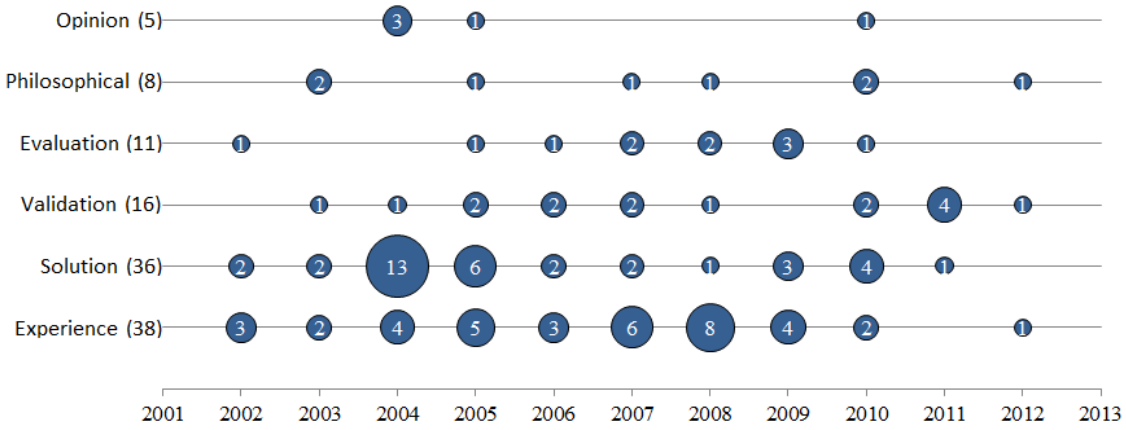

Figure 6: Beakdown of research type by year.

as a whole to foster more crossover between these two conferences, as used to be the case with XP/Agile Universe.

Figure 6, on the other hand, shows that, while Experience reports have traditionally been common in research publications in the past, they have not been popular in recent years. This could be a result of the Agile conference's recent introduction of different presentation formats – not all of which require a paper submission – resulting in fewer of these publications. Additionally, the number of Solution papers spiked in 2004 and 2005, but declined markedly recently. Similarly, the number of Evaluation papers is quite low, and none were in our paperset from 2011 onwards. This is worrying on two fronts, given that Solution papers pave the way for further exploration of new topics and Evaluation papers provide proper evidence for the effectiveness of a given technique. As these papers indicate both new and proven testing techniques, it's important to the future of the field of agile testing to encourage more of these publications.

When comparing these results against the results from last year, we can again see some distinct differences between these conferences and the field as a whole. The biggest difference is that there were far more Evaluation papers in last year's paperset – 46, making it the most frequent papertype. Experience and Solution papers, on the other hand, occurred at a comparable frequency in both conferences. This is disturbing as it shows that, within these conferences, there is a lack of Evaluation papers. Since Evaluation papers present rigorously-evaluated results, we should, in the future, find ways of encouraging this kind of publication at the Agile and XP conferences.

C. *What problems with agile testing do publications mention?*

It's natural to assume that every publication will address a specific problem; however, this is not always the case. In the 110 publications we considered in this study, only 45 problems with testing were identified – even given that it was possible for a single paper to list more than one problem. In the future, it is extremely important that publications identify the problem they are addressing in their abstracts or titles in order to make it easier for both researchers and practitioners to understand if a given paper is relevant. Because of the relatively low number of keywords for this research question, we address it only by conference, not by year.

All of the keywords identified for this question are shown broken down by conference in Figure 7. From this figure, we can see that the most commonly-discussed problems at Agile are how to adopt agile testing followed by the difficulty of understanding tests. At XP, however, the main issues are repeatability of tests (as in, if I run the same test multiple times, the same result should occur every time) and test maintenance. Again, this seems to indicate that more cross-pollination between these two conferences should occur.

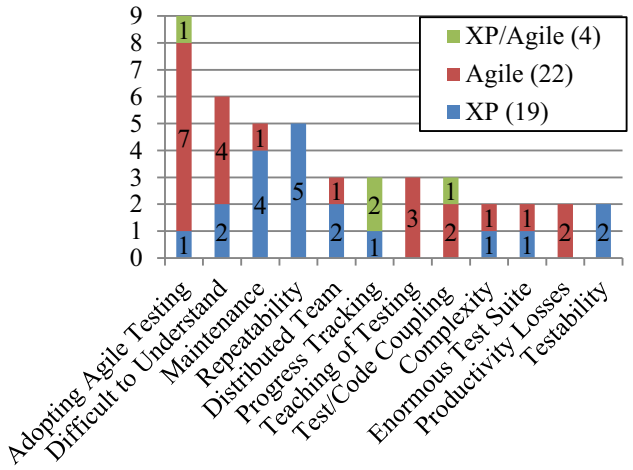

Figure 7: Breakdown of problems encountered with testing by conference.

D. *What benefits do publications propose their research will produce?*

The same issue as in the previous section was discovered in this section: the titles and abstracts of publications often did not list the benefits of their work. Out of the 110 papers included in our set, only 28 benefits were listed. The 7 keywords relating to benefits are broken down by conference in Figure 8. Even given the low number of keywords listed, we can infer some differences between the various conferences. For example, the initial XP/Agile conference focused mostly on productivity increases from testing and TDD. XP, on the other hand, focuses on topics like defect localization, test clarity, and the catch-all of quality. Agile also focused on quality to the exclusion of all else. This is disturbing because saying that a given technique or tool increases quality is very ambiguous. These publications could be discussing many different aspects of quality, but this will not be obvious from their abstracts.

This lack of discussion of what benefits are being offered by different research is especially confusing from the point of view of someone with a problem searching for a solution in research publications. Given that someone wanted to know,

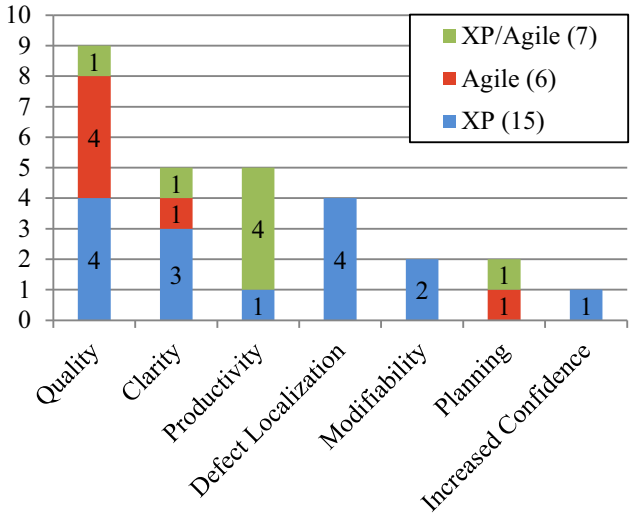

Figure 8: Breakdown of benefits offered by publications by conference.

for example, how testing can be used to help with project planning in an agile context, a search through the proceedings of these conferences would turn up a paltry two results. This is extremely concerning given that these papers were written in order to share ideas on how to solve problems. To address this, future publications should be extremely clear on what they can offer to people who would invest the time in reading them.

V. Secondary Research Questions

Given the issues mentioned above with titles and abstracts not adequately describing research, we also felt it would be strongly beneficial to provide an overview of topics that may help people new to the field to get oriented more quickly: who are the leading authors; which countries produce the most research; which tools come up most frequently; and is the research coming from industry or academia? This section investigates each of these topics in turn and, where possible, compares them against last year's study.

A. Which authors lead the field of agile testing?

In order to make it easier to find information on the topic of agile testing, we compiled a list of the authors who published most frequently in our paperset. The hope is that, for example, a new student trying to get oriented to this field would have a list of authors to contact for guidance.

Figure 9 presents an overview of the number of authors with a given number of publications (note: there are no authors in this set with 6-15 publications). In our analysis, we found that 199 distinct authors contributed publications to these conferences between 1 and 16 times. However, the vast majority of authors (172) were involved in precisely one paper. This is specifically worrying given that most papers in our set were either Solution or Experience. What this means is that, statistically, authors present either a novel (but not well-evaluated) idea or a summary of an experience in industry, but never return to publish more rigorous results. This is disturbing because it implies that there are a large number of techniques in the literature that have never been rigorously evaluated. (or, if they have been evaluated, have not been reported to the community).

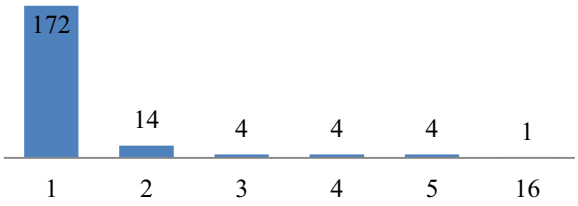

Figure 9: Number of authors with the given number of publications.

Table 1 shows the top 9 authors in the paperset, as well as any of the top 10 uses for agile testing (from PRQ1) associated with any of their publications. It is provided in the hope that it will help newcomers to this field get in touch with the leading authors in this field. Of the 9 authors mentioned this year, 4 (out of a possible 6) were also included in last year's paper. However, the remaining two were not identified in the present study – their publications were submitted to different conferences entirely. While it's an encouraging sign that we identified so many of the most important authors, it's somewhat discouraging that some of the leaders in the field do not choose to publish at Agile or XP.

B. In which countries do the leading authors work?

In addition to each author's name, we also kept track of each author's location at time of publication. This is based on the institution with which the author was associated at time of publication. In future work, it might be worthwhile to track the authors' affiliation at the time the study was conducted; however, this has not been done for the current paperset. A country was counted once per author of a paper for each paper in the set.

The authors included in our paperset published from 30 different nations; however, the top 10 countries published the vast majority of the papers, and, of these, two stand out extremely strongly: Canada and the United States. Figure 10 gives more details on where authors have published from over time. Out of 263 author instances, 83 came from Canada and 55 came from the United States. For comparison, all other non-top-10 countries combined came to 50 instances. Clearly, more needs to be done to encourage agile testers from other

Table 1: Top Authors in Agile Testing

Author	Publications	Keywords
F. Maurer	16	TDD, Automated Testing, Acceptance Testing, ATDD, Specification, Refactoring, GUI Testing, Automated Testing, Regression Testing
M. Smith	5	TDD, Automated Testing, Refactoring, Web Testing,
G. Melnik	5	TDD, Acceptance Testing, ATDD, Automated Testing, Specification,
J. Miller	5	TDD, Automated Testing, Refactoring, Web Testing,
A. Geras	5	TDD, Acceptance Testing, Automated Testing, Web Testing,
S. Park	4	TDD, Acceptance Testing, ATDD, Regression Testing, Automated Testing
J. Andrea	4	Acceptance Testing, Web Testing, Regression Testing, Automated Testing, Refactoring
Y. Ghanam	4	TDD, Refactoring, Acceptance Testing, Specification, ATDD, Regression Testing,
R. Mugridge	4	TDD, Automated Testing, Acceptance Testing, Web Testing

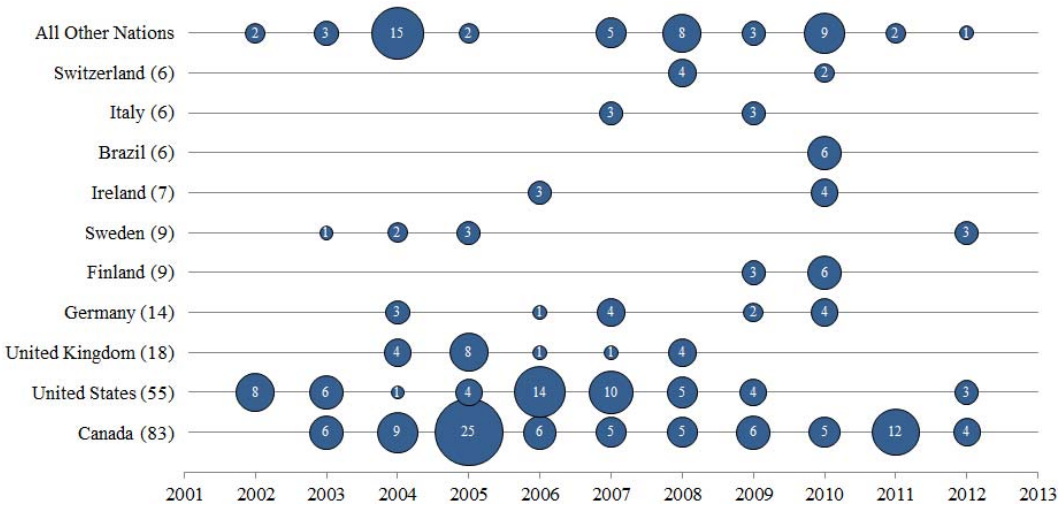

Figure 10: Location of authors in our paperset at time of publication.

countries to contribute to the Agile and XP conferences in order to keep the topic from being entirely dominated by voices from Canada and the United States.

For the field of agile testing as a whole, though, it's worth noting that regional conferences (for example Agile Brazil) are becoming more prominent. It's possible that much of the work that would traditionally have been submitted to Agile or XP is being directed towards these conferences instead. A useful extension of this paper would be to extend the paperset to include those conferences and workshops; however, this is left as future work.

It's interesting to note that, while the United States was pretty dominant in previous years, its influence over the field of agile testing has decreased somewhat since 2009. In 2010, this may have contributed to a general increase in the presence of other countries, but, since then, it may just be contributing to the general decrease in the number of publications in this field. The United States isn't unique in this regard; other countries, such as the United Kingdom, were major contributors in years past, but have had little or no presence in this field in recent years.

C. *What tools have been used for agile testing?*

Not every paper mentions a tool, but we did keep track of the tools that authors have made use of or have created where these tools were mentioned in the titles and abstracts of papers. Overall, tools were mentioned 47 times over 28 unique tools. Figure 11 shows a tag cloud wherein the color and size of tool names indicate how often those tools occurred in our paperset.

Out of these, Fit and jUnit were mentioned most frequently. This fits with the rest of the study given that acceptance testing, unit testing, TDD, ATDD, and automated testing are all keywords identified during PRQ1 and that these tools can be used in support of that style of testing. Web testing tools like Selenium were also fairly common. Again, this makes sense given that web and GUI testing were identified as keywords in PRQ1. When compared against the tools from last year's study, however, it's apparent that the tools from this study were less diverse. Last year's set included tools for advanced testing techniques – such as mutation testing – that are completely absent from this year's study. It seems as though it might be prudent to encourage some exploratory use of these kind of tools within the agile community as a first step towards broadening the types of testing that are discussed at the Agile and XP conferences.

D. *Is the field led by practitioners, academics, or collaborations?*

As with last year's study, we also sought to find out whether this year's publications were coming from academic groups, industry groups, or collaborations. In order to do this, we coded the author institutions of papers as one of: industry, academic, or both. This was done based on author affiliations at time of publication. The "both" keyword was used for papers with at least one author from each type of institution. We feel that this is an important distinction as collaborations have the potential to combine the rigor of academic evaluations with the practicality of industry concerns. The distribution of papers across these three keywords can be seen in Figure 12.

Figure 12 provides an interesting look at the composition of each conference. It seems that the original XP/Agile Universe conference was slanted towards industry authors, the XP conference is slanted towards academic authors, and the Agile conference has a roughly even mix of both groups. XP seems to have the strongest showing for collaborations between academics and practitioners even though it is somewhat weak in industry authors. While the Agile conference is strong in both academic and industry contributions, it should strive in the future to find ways of encouraging collaborative submissions.

Compared to the findings of last year's study, however, the Agile and XP/Agile Universe publications are quite diverse. The distribution for the field of agile testing as sampled last year is closer to the distribution for XP: heavily slanted towards academic publications. While the Agile conference should strive to encourage collaborations, XP

needs to find ways of encouraging more industry authors and collaborations to investigate the topic of agile testing.

VI. Limitations

In the process of performing this systematic mapping, we encountered a variety of issues that are worth mentioning. Some are important given that this is a follow-up study; some are important given that they impact not only this study, but can apply to systematic mappings and systematic literature reviews generally.

A. Issues with This Study

Out of 9 authors, only 2 participated in both studies (Hellmann and Maurer). This means that the way keywords were applied to papers might not have been consistent across years. In order to minimize this risk, one of the authors (Hellmann) collaborated on each step of the analysis with at least one other author. This was done in order to increase the consistency of our results across years. However, this solution, in turn, could also have biased the analysis towards his opinions. A different way of approaching this problem would have been to measure inter-rater reliability using a statistical method such as kappa statistics. This could be measured by, for example, a pilot evaluation in which authors from the current year's study re-keyword papers from the previous year. This method would have the advantage of allowing us to specify a confidence value for the level of agreement that could be expected of different researchers in different studies. This method would be promising for future work, but was not used in this study.

Second, there is a risk that, by restricting our sources to three specific, hand-picked conferences, we are unreasonably biasing our results towards the idiosyncrasies of those conferences. We attempt to minimize this limitation by comparing our results to the results from last year's study where possible. This also allows us to compare the general set of papers from last year with the restricted set of papers from this year – especially given that the overlap between the two sets was so small.

B. Issues with Systematic Mappings Generally

The fact that the overlap between this year's and last year's papersets is small is another limitation of this study. However, given that most systematic studies collect their initial papersets by querying online databases, this is by no means a limitation specific to this study. There seem to be two fundamental issues with systematic studies: keywording papers reliably given that different people will view papers in different ways (or even that different authors will use different terms or use terms in different ways); and finding a consistent initial paper set. If we search for papers manually, as in this study, the validity of our initial paperset suffers; however, if we search for papers using an automated system, we run a strong risk of missing relevant papers (as the difference between this year's study and last year's study strongly shows). Additionally, when performing a manual search on specific conferences, there is an additional problem: how do researchers know which proceedings to search? This is a serious issue given that automated searches seem to be prone to missing relevant work.

Figure 11: Tools used for agile testing.

Even ignoring the issues in the previous paragraph, a systematic mapping study is still less reliable than a systematic review. Our results could be different if we were considering the full text of each paper rather than the abstracts and titles alone. However, from the perspective of a person new to the field, a systematic mapping is still a realistic approach given that people will make the decision to read a paper or not based on its title and abstract.

Unfortunately, it's again worth noting that abstracts of papers frequently did not include information crucial to understanding what the work was about. As was made clear in PRQ3 and PRQ4, basic information about the problem a publication addresses and the benefits it offers is frequently missing. This makes it difficult to perform systematic mapping studies in general.

VII. Future Work

Again, an obvious direction for future work would be to extend this paper into a full systematic review. This would enable us to look into each source in more depth to make sure

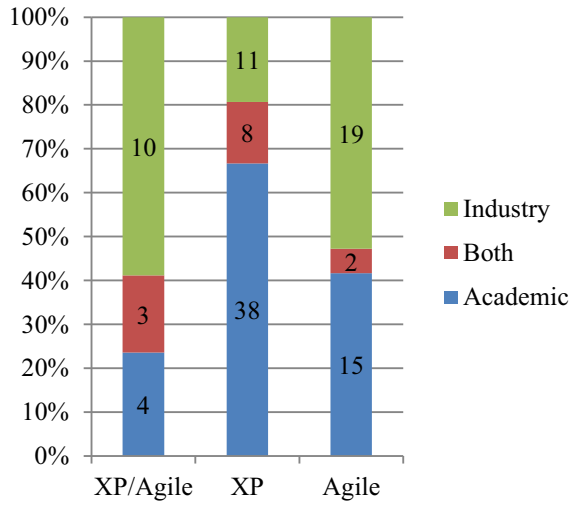

Figure 12: Distribution of papers between industry, academic, and collaborative sources.

that we have gathered all appropriate keywords for each paper. However, the obvious problem with this is that we will still be selecting papers for the review based on their title and abstract and would still run into issues with the initial paperset we select.

One way of getting around the paper selection issue would be to follow the same approach used for this study in reverse: manually sort through the proceedings of leading testing conferences in search of papers relating to agile. This would allow us to extend the set of papers we consider while still constraining the set of papers we need to manually read through such that the study would be feasible. This would allow us to approach this topic again from a new angle while still gathering results in such a way that we would be able to compare that study to the results of this study and of last year's study.

Again, as we mentioned last year, it would be interesting to perform a survey of existing testing tools in order to better understand what options are available to developers. This could be done both through both search engines and through searching open-source repositories (like CodePlex and SourceForge). Given the limitations we mentioned above, this has the potential to miss relevant tools, but it would still be useful for understanding what sort of tools exist that might be able to address some of the issues identified through this study.

Finally, it would be worthwhile to look into the top authors in greater depth – to provide details about the kind of work they are involved in outside of a strictly agile testing context, for example. This would allow us to get a clearer picture of how their work in agile testing fits into the greater context of their research as a whole. However, this is left for future work.

VIII. Conclusion

In this paper we present the findings of a systematic mapping of the field of agile testing wherein we manually searched the proceedings of the XP, Agile, and XP/Agile Universe conferences for relevant papers. This allowed us to address a variety of research questions and provide insight into the distribution of topics, authors, etc. within this field. We were also able, in many cases, to compare this data with the data collected in the previous iteration of this study, providing insight into the reliability of both studies and, to an extent, allowing us to compare the results of a mapping study constrained to just a few conferences to the results of a mapping study which searched through millions of papers.

What has agile testing been used for? We found an emphasis on types of testing that are already closely associated with agile, including acceptance testing, automated testing, TDD, and ATDD. However, we also found that few publications address difficult questions in the field of testing that are important in agile contexts, such as how to perform GUI testing.

What types of papers on agile testing have been published? Contrary to last year's study, we found that this year's paperset tended to focus on Solution and Experience papers. These papers tend to introduce new ideas, but the paper types that would indicate in-depth, follow-up work (like Evaluation papers) are much lower in this paperset than in last year's. In a way this is troubling. While the conferences considered in this study appear to be good at fostering new ideas, more effort should be placed on encouraging the publication of follow-up studies.

What problems with agile testing do publications mention? The analysis of this question showed a marked difference between the conferences considered in this study. While the Agile conference focuses on issues like adopting and teaching agile testing, XP focuses on more research-oriented topics such as the repeatability and maintainability of tests.

What benefits do publications propose their research will produce? As with the previous question, this analysis showed an interesting distinction between the two ongoing major agile conferences: Agile papers focused almost exclusively on "quality" but did not provide much additional detail while XP focused on quality, clarity, and defect localization. However the main finding for this question remains that the abstracts for these papers were not created in such a way that they can be properly analyzed using a systematic mapping study.

Which authors lead the field of agile testing? We enumerate the top 9 authors for the field of agile testing along with a list of topics within agile testing on which they have published. Out of the 6 top authors mentioned last year, 4 appear in this list as well. This section may be of particular interest to those just getting started in the field of agile testing.

In which countries do the leading authors work? In an attempt to understand where, globally, publications were being produced, we also looked into the country of affiliation of authors. Despite the fact that XP is usually located in Europe and could be expected to attract more European authors, our paperset was dominated by Canada and the United States.

What tools have been used for agile testing? We found that a variety of tools were mentioned across papers with tools that can be used for automated testing, acceptance testing, unit testing, TDD, and ATDD in the lead. Unlike last year, no tools for advanced testing concepts, like mutation testing, were found.

Is the field led by practitioners, academics, or collaborations? Overall, we found a good mix of academic and industry authors, though collaboration between the two groups could stand to be increased. We were also able to analyze the conferences individually to find that XP has a stronger academic showing while Agile has a good mix of industry and academic authors but far fewer collaborations. This knowledge could help authors decide which of the two major agile conferences to submit their work to.

Overall, this paper provides a detailed overview of the field of agile testing with respect to the XP/Agile Universe, Agile, and XP conferences. For newcomers, this study serves as an introduction to the field of agile testing. For authors who have already published work on agile testing, this paper serves as a guide towards what could be done to improve the field as a whole. Finally, and perhaps most importantly, this paper provides details on the process of performing this sort of systematic mapping study that will be useful for making sure that future studies are more comprehensive.

WORKS CITED

[1] Glenn Vanderburg, "A Simple Model of Agile Software Practices - or - Extreme Programming Annealed," in *Object-Oriented Programming, Systems, Languages, and Applications*, New York, 2005, pp. 539-545.

[2] S. MacDonnel, M. Shepperd, B. Kitchenham, and E. Mendes, "How Reliable Are Systematic Reviews in Empirial Software Engineering?," *IEEE Transactions on Software Engineering*, vol. 36, no. 5, pp. 676-687, Sep. - Oct. 2010.

[3] Theodore D. Hellmann, Abhishek Sharma, Jennifer Ferreira, and Frank Maurer, "Agile Testing: Past, Present, and Future - Charting a Systematic Map of Testing in Agile Software Development," in *AGILE 2012*, Dallas, Texas, 2012, pp. 55-63.

[4] Tore Dybå and Torgeir Dingsøyr, "Empirical Studies of Agile Software Development: A Systematic Review," *Information and Software Technology*, vol. 50, no. 9-10, pp. 833-859 , January 2008.

[5] F. Shull et al., "What Do We Know about Test-Driven Development?," *IEEE Software*, vol. 27, no. 6, pp. 16-19, Nov. - Dec. 2010.

[6] Adnan Causevic, Daniel Sundmark, and Sasikumar Punnekkat, "Factors Limiting Industrial Adoption of Test Driven Development: A Systematic Review," in *International Conference on Software Testing, Verification, and Validation*, Berlin, 2011, pp. 337-346.

[7] Ron Jeffries and Grigori Melnik, "Guest Editors' Introduction: TDD - The Art of Fearless Programming," *IEEE Software*, vol. 24, no. 3, pp. 24-30, May-June 2007.

[8] Barbara Kitchenham, Pearl Brereton, Li Zhi, David Budgen, and Andrew Burn, "Repeatability of Systematic Literature Reviews," in *Evaluation & Assessment in Software Engineering*, Durham, UK, 2011, pp. 46-55.

[9] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson, "Systematic Mapping Studies in Software Engineering," in *12th International Conference on Evaluation and Assessment in Software Engineering*, Bari, Italy, 2008, pp. 71-80.

[10] Barney G. Glaser and Anselm L. Strauss, *The Discovery of Grounded Theory*. Chicago, United States of America: Aldine, 1967.

[11] Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland, "Requirements Engineering Paper Classification and Evaluation Criteria: A Proposal and a Discussion," *Journal of Requirements Engineering*, vol. 1, no. 11, pp. 102-107, 2005.