PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Universidade Federal do Rio de Janeiro

Arnaldo Gil de Souza Sardella Nascimento

AMEAÇAS À VALIDADE DE ARTIGOS QUE EMPREGAM MÉTODOS EMPÍRICOS EM TESTES DE APLICAÇÕES CONSTRUÍDAS PARA AMBIENTE *WEB*.

DISSERTAÇÃO DE MESTRADO

Instituto de Matemática

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ARNALDO GIL DE SOUZA SARDELLA NASCIMENTO

AMEAÇAS À VALIDADE DE ARTIGOS QUE EMPREGAM MÉTODOS EMPÍRICOS EM TESTES DE APLICAÇÕES CONSTRUÍDAS PARA AMBIENTE *WEB*.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática e Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Informática.

Orientador: Professor Eber Assis Schmitz, Ph.D

Co-orientadora: Professora Priscila Machado Vieira Lima, Ph.D

Rio de Janeiro
2019

de Souza Sardella Nascimento, Arnaldo Gil
d244a AMEAÇAS À VALIDADE DE ARTIGOS QUE EMPREGAM MÉTODOS EMPÍRICOS EM TESTES DE APLICAÇÕES CONSTRUÍDAS PARA AMBIENTE WEB / Arnaldo Gil de Souza Sardella Nascimento. -- Rio de Janeiro, 2019.
101 f.

Orientador: Eber Assis Schmitz.
Coorientadora: Priscila Machado Vieira Lima.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Programa de Pós-Graduação em informática, 2019.

1. Ameaças à validade. 2. Estudos empíricos. 3. Engenharia de software. 4. Testes aplicações Web. 5. Experimentos. I. Assis Schmitz, Eber , orient. II. Machado Vieira Lima, Priscila , coorient. III. Titulo.

ARNALDO GIL DE SOUZA SARDELLA NASCIMENTO

AMEAÇAS À VALIDADE DE ARTIGOS QUE EMPREGAM MÉTODOS EMPÍRICOS EM TESTES DE APLICAÇÕES CONSTRUÍDAS PARA AMBIENTE *WEB*.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática e Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Informática.

Aprovada em 28 de novembro de 2019.

_____

Professor Eber Assis Schmitz, Ph.D, UFRJ

_____

Professora Priscila Machado Vieira Lima, Ph.D, UFRJ

_____

Professora Mônica Ferreira da Silva, D.Sc, UFRJ

_____

Professor Denis Silva da Silveira, D.Sc, UFPE

_____

Professor Toacy Cavalcante de Oliveira, D.Sc, UFRJ

Rio de Janeiro

2019

Agradecimentos

Ao amigo, Professor e mentor Doutor Eber Assis Schmitz, pelo incentivo e confiança no meu trabalho.

À Professora Doutora Priscila Machado Vieira Lima, pelas palavras fortes que me incentivaram a seguir em frente.

À Professora Mônica Ferreira da Silva, por sua alegria e necessários ensinamentos.

Ao Professor Doutor Antonio Juarez Sylvio Menezes de Alencar, sem o qual minha carreira em TI não teria sido a mesma.

Ao Professor Doutor Denis Silva da Silveira, por toda a ajuda, o que tornou possível uma melhora significativa neste trabalho.

Ao Professor Doutor Toacy Cavalcante de Oliveira, por suas observações sempre pertinentes.

A todos os Professores e colegas de mestrado que tornaram possível que eu vencesse esta importante etapa da minha vida.

À minha mulher, Doutora Simone Escobar, pela paciência, ajuda e incentivo.

“As pessoas preferem acreditar naquilo que elas preferem que se seja verdade.”

(Francis Bacon)

Resumo

NASCIMENTO, Arnaldo Gil de Souza Sardella, Ameaças à validade de artigos que empregam métodos empíricos em testes de aplicações construídas para ambiente *Web*. 2019. 100 folhas. Dissertação (Mestrado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

O progresso na Engenharia de Software requer (1) mais estudos de qualidade empírica, (2) maior foco na síntese de evidências, (3) mais teorias a serem construídas e testadas, e (4) a validade do experimento está diretamente relacionada ao nível de confiança no processo de pesquisa experimental. Este trabalho apresenta os resultados de uma classificação qualitativa e quantitativa de ameaças à validade de experimentos de Engenharia de Software, compreendendo um total de 193 artigos publicados no período de 2010-2019, cujo tema é: testes de software de aplicações *Web*. Nossos resultados mostram que 104 (54%) artigos analisados não mencionam qualquer ameaça à validade. Dos 89 (46%) trabalhos que mencionam algum tipo de ameaça à validade de seus experimentos, 72 (37%) artigos mencionam suas ameaças de uma maneira sumária e somente 17 (9%) artigos o fazem detalhadamente.

Palavras-chave: Ameaça, Aplicação, Conclusão, Construção, Empírico, Estudo, Experimento, Externa, Interna, Método, Qualitativo, Quantitativo, Teste, Validade, *Web*.

Abstract

NASCIMENTO, Arnaldo Gil de Souza Sardella. Threats to validity of articles employing empirical methods in testing of applications constructed for *Web* environment. 2019. 100 folhas. Dissertação (Mestrado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Progress in software engineering requires (1) more empirical quality studies, (2) greater focus on evidence synthesis, (3) more theories to be constructed and tested, and (4) the validity of the experiment is directly related to the level of confidence in the experimental research process. This work presents the results of a qualitative and quantitative classification of threats to the validity of software engineering experiments, comprising a total of 193 articles published in the period 2010-2019, whose theme are *Web* application software tests. Our results show that 104 (54%) of the analyzed articles do not mention any threat to the validity. In the remaining, 89 (46%) papers that mention some kind of threat to validity of their experiments, 72 (37%) do it briefly and only 17 (9%) articles do it judiciously.

Keywords: Application, Conclusion, Construction, Empirical, Experiment, External, Internal, Method, Qualitative, Quantitative, Study, Test, Threat, Validity, *Web*.

Lista de Figuras

Figura 1 – Processo de seleção dos estudos primários para investigação.....42

Lista de Tabelas

Tabela 1– Artigos que fazem e que não fazem referência as suas ameaças à validade 84
Tabela 2– Totais e percentagem de artigos que fazem e que não fazem referência as suas ameaças à validade 89
Tabela 3– Artigos que fazem e que não fazem referência as suas ameaças à validade – frequência anual 90
Tabela 4– Ameaças mais citadas 90
Tabela 5– Total de ameaças mais citadas 94
Tabela 6– Resumo da Quantidade de Artigos Encontrados, Baixados e Selecionados 94
Tabela 7– Tipos de Publicação por Ano de Publicação 95
Tabela 8– Resumo dos resultados encontrados 100

Sumário

1 Introdução 13
1.1 Descrição do problema e sua relevância 13
1.2 Objetivo geral 13
1.2.1 Objetivos específicos 14
1.3 Resultados esperados 14
1.4 Metodologia adotada 14
1.5 Classificação do estudo 15
1.6 Organização do trabalho 15
2 Referencial teórico 16
2.1 Arcabouço conceitual 16
2.1.1 Revisão sistemática da literatura 16
2.1.2 Estudos empíricos em Engenharia de Software 17
2.1.3 Métodos empíricos em Engenharia de Software 19
2.1.4 Ameaças à validade de estudos empíricos 25
2.1.5 Testes em softwares 35
3 Método de pesquisa 39
3.1 Planejamento 39
3.1.1 Formulação das questões de pesquisa 39
3.1.2 Formulação da *String* de busca 40
3.1.3 Definição da estratégia de busca 40
3.1.4 Seleção das fontes de pesquisa 41
3.1.5 Seleção dos estudos primários 41
3.1.6 Avaliação da qualidade dos estudos 43
3.2 Condução da pesquisa 45
3.2.1 Planejamento do projeto de pesquisa 45
4 Resultados do mapeamento 47
4.1 Respostas às questões de pesquisa 47
4.1.1 Resposta à questão geral 47
4.1.2 Respostas às questões específicas 47
4.2 Ameaças à validade deste estudo 53
4.2.1 Validade de conclusão 54
4.2.2 Validade interna 55
4.2.3 Validade de construção 57

4.2.4 Validade externa 59
4.2.5 Prioridade entre os tipos de ameaças à validade 59

5 Conclusão 61

Referências Bibliográficas 64

Referências de Estudos Teóricos 64
Referências de Artigos Primários 66

Apêndices 84

APÊNDICE A – TABELAS RESULTADO 84

1 Introdução

Esta pesquisa teve como resultado o artigo "*A Mapping Study of Scientific Merit of Papers, which Subject are Web Applications Test Techniques, Considering their Validity Threats*", de minha autoria e dos Professores: Doutor Eber Assis Schmitz, Doutora Priscila Machado Vieira Lima e Doutora Mônica Ferreira da Silva, publicado pelo *"Journal of Information Systems and Technology Management* – JISTEM" que está disponível no endereço eletrônico:
http://www.scielo.br/scielo.php?pid=S1807-17752018000100308&script=sci_arttext.

1.1 Descrição do problema e sua relevância

Segundo Sjøberg *et al*. (2007), para que haja evolução na área de Engenharia de Software são necessários mais estudos empíricos de qualidade e maior foco em sintetizar evidências e construir teorias. Precisa-se, portanto, de estudos nos quais possamos confiar, principalmente em relação à sua validade. Para a obtenção desta confiança, as informações colhidas devem ser mais focadas na observação e experimentação do que na lógica dedutiva ou matemática. Para isto, devemos utilizar como nas ciências que estudam fenômenos do mundo real, isto é, ciências empíricas, métodos empíricos.

De acordo com Wainer (2007) e Travassos, Gurov e Amaral (2002), para que um experimento seja válido, deve-se ter um alto nível de confiança em todo o processo de investigação experimental. Esta confiança deve permear todos os elementos envolvidos no processo, desde a base teórica adotada até os resultados obtidos e considerar as ameaças (externa, interna, de construção e de conclusão) à validade do experimento.

Pela relevância dessas afirmações, achou-se importante investigar se as pesquisas em Engenharia de Software são verdadeiramente encaradas como ciência e se as ameaças à validade dos estudos empíricos são observadas criteriosamente quando o estudo é realizado.

1.2 Objetivo geral

A partir de um estudo de mapeamento sistemático de literatura, essa pesquisa propõe selecionar artigos cujo tema seja "Testes em aplicações *Web* que empregam

métodos empíricos" e analisar a quantidade de artigos, em números e percentual, que levam em consideração suas ameaças à validade.

1.2.1 Objetivos específicos

a) Realizar uma revisão da literatura em busca de conhecimento de quais são as ameaças à validade de um estudo e como avaliá-las;
b) Encontrar artigos publicados na *Web*, referentes a testes em aplicações *Web*, em que conste pelo menos um experimento próprio, e que possam ser baixados sem custo, no período de 2010 a 2019.
c) Criar um método de inclusão e exclusão de artigos baixados da *Web*, visando atingir um número de trabalhos, que possam ser analisados no tempo disponível para este estudo.
d) Criar um método de avaliação dos artigos selecionados baseado em critérios de apresentação de suas ameaças à validade, que seja mais apropriado para o contexto desta investigação;
e) Apresentar a quantidade e o percentual dos artigos selecionados que seguem os preceitos estabelecidos quanto às suas ameaças à validade e compará-los com a quantidade de artigos selecionados;
f) Tirar conclusões baseadas nos números e percentuais encontrados.

1.3 Resultados esperados

Espera-se que os resultados deste trabalho possam, não apenas apresentar uma avaliação da quantidade e percentual de artigos que levam em consideração suas ameaças à validade, como também auxiliar pesquisadores a terem um maior rigor metodológico nos seus estudos, o que aumentará a confiabilidade dos referidos estudos e beneficiará a comunidade que publica trabalhos sobre testes de aplicações *Web*.

1.4 Metodologia adotada

Esta pesquisa foi concebida através de artigos publicados nos principais veículos de divulgação científica da comunidade de Engenharia de Software. Para encontrar os artigos alvo, foram utilizadas as principais máquinas de busca utilizadas na *Web*.

Assim, foram encontrados na *Web* 797 artigos, no período de 2010 a 2019. Destes baixados 336 e selecionados 193 artigos para estudo, através de critérios de escolha que serão discutidos oportunamente.

1.5 Classificação do estudo

A pesquisa apresentada nesta dissertação é baseada em dados de natureza quantitativa e qualitativa. Foi feita através de um estudo de mapeamento sistemático, que é um tipo de revisão sistemática da literatura, com viés quantitativo e que utiliza uma pesquisa bibliográfica com abordagem dedutiva para sua finalidade. Esta pesquisa também pode ser classificada como descritiva.

1.6 Organização do trabalho

Além dessa Introdução, esta dissertação está organizada da seguinte maneira:

Capítulo 2 (Referencial Teórico): apresenta o arcabouço conceitual, que contém a revisão sistemática da literatura que é a base teórica dos principais conceitos que fundamentam os estudos empíricos e descreve e classifica suas ameaças à validade.

Capítulo 3 (Método de Pesquisa): descreve a metodologia de pesquisa e a estratégia utilizada para a busca e seleção dos artigos utilizados neste estudo;

Capítulo 4 (Resultados do mapeamento): responde as questões de pesquisa formuladas apresentado os números e percentuais encontrados; Apresenta e descreve as ameaças à validade as quais este estudo está sujeito;

Capítulo 5 (Conclusão): neste capítulo final são apresentadas as conclusões deste estudo de mapeamento e propostos trabalhos futuros.

Referências Bibliográficas: serão apresentados os estudos teóricos que embasaram esta pesquisa e também os estudos primários que são as fontes de dados para a mesma.

2 Referencial teórico

2.1 Arcabouço conceitual

2.1.1 Revisão sistemática da literatura

Kitchenham *et al*. (2010) defendem a ideia de que pesquisadores empíricos em Engenharia de Software devem desenvolver seus estudos baseados em evidências, como o fazem os pesquisadores em medicina e sociologia, e que a evidência mais confiável vem de agregar todos os estudos empíricos sobre um determinado tópico.

Revisões sistemáticas da literatura são um meio de agregar conhecimento sobre um tópico de Engenharia de Software ou uma questão de pesquisa e tornam-se o mais imparciais possível por serem auditáveis e repetíveis (KITCHENHAM *et al*. 2010).

Ainda segundo Kitchenham *et al*. (2010), as revisões sistemáticas da literatura são classificadas em convencionais e mapeamento de estudos.

As convencionais agregam resultados relacionados a uma determinada questão de pesquisa, por exemplo: é a técnica de testes “A” mais eficaz na detecção de defeitos que a técnica “B”? Se existem suficientes estudos primários comparáveis com estimativas quantitativas da diferença entre os métodos, meta-análise pode ser usada para realizar uma agregação com base em estatística formal. No entanto, foi descoberto que meta-análise raramente é utilizável em revisões sistemáticas da literatura de Engenharia de Software, porque muitas vezes há estudos primários insuficientes.

As classificadas como mapeamento de estudos destinam-se a encontrar e classificar os estudos primários em uma área temática específica. Elas têm granulação grosseira em questões de pesquisa, tais como: “o que sabemos sobre tópico X?” e podem ser usadas para identificar a literatura disponível antes de realizar as revisões sistemáticas da literatura convencionais. Utilizam os mesmos métodos para a busca e extração de dados como revisões sistemáticas da literatura convencionais, mas confiam mais na tabulação dos estudos primários em categorias específicas. Além disso, alguns estudos de mapeamento estão mais preocupados sobre como os acadêmicos realizam pesquisas em Engenharia de Software e não o que se sabe sobre um determinado tema de Engenharia de Software.

O estudo relatado neste trabalho é um estudo de mapeamento.

2.1.2 Estudos empíricos em Engenharia de Software

De acordo com Zelkowitz *et al*. (1998), experimentos em Engenharia de Software ainda não possuem a maturidade observada em outras áreas da ciência, pois muitos trabalhos ainda não seguem uma metodologia específica e não utilizam técnicas de validação apropriadas.

Para alavancar o método científico na área, foi criada a Engenharia de Software Experimental (ESE). Segundo Boehm *et al*. (2005), atribui-se a Victor R. Basili as primeiras experiências para aumentar o rigor da pesquisa em Engenharia de Software e a criação de ferramentas para aprimorar a qualidade dos trabalhos gerando evidências confiáveis.

Miller (2000) nos diz que a confiabilidade das evidências depende do tipo de estudo realizado e também da qualidade com que seus resultados são replicados, pois dificilmente os resultados obtidos de um único experimento são confiáveis. Para que uma conclusão confiável seja obtida é necessária à combinação dos resultados de diversos experimentos relacionados e atrelados a hipóteses com variáveis comuns ou a mesma hipótese (replicações).

Dentre os motivos que trazem dúvidas à confiabilidade das evidências dos estudos, estão: as falhas nas formulações das hipóteses, o erro no projeto do estudo e o erro na execução do estudo.

Conforme a afirmação de Basili *et al*. (1999), estudo experimental é uma ação realizada com a finalidade de desvendar algo desconhecido ou de provar uma hipótese, com coleta de dados e execução de análises feitas por um investigador, que deve determinar o que os dados significam.

Todo estudo experimental deve descrever seus objetivos, seu método, os resultados e conclusões.

É definido como objetivo, a hipótese de pesquisa que deve ser estabelecida com clareza. O método de pesquisa estabelece o rigor das técnicas e procedimentos adotados e a validade das evidências obtidas do estudo. Os resultados, além deles mesmos, devem discutir ameaças à validade, limitações do método utilizado e das variáveis do estudo. A conclusão confirma se a hipótese descrita no objetivo é verdadeira ou não e aponta implicações futuras.

Como características de um estudo experimental, podemos definir o seguinte conjunto de atributos:

- Hipótese - é a proposição, teoria ou suposição que pode explicar um determinado comportamento, independente do fato de ser verdadeira ou falsa e que seja isenta de intenção humana.
- População - o conceito de população, em conformidade com os achados de Juristo *et al.* (2001), para Engenharia de Software, é a de um estudo experimental que se divide em pessoas, produtos, problemas e processos. Para a realização do estudo uma amostra dessa população é escolhida. Cada elemento escolhido é uma unidade de estudo. Um experimento unitário constitui-se na aplicação de uma combinação de intervenções/tratamentos, em uma unidade de estudo, pelo aplicador do estudo. A seguir veremos as especificações das divisões de um estudo experimental:
    a) Pessoas: desenvolvedores, usuários e demais interessados no projeto de software;
    b) Produtos: software e artefatos que o compõem;
    c) Problema: necessidade do usuário que motivou a execução do projeto;
    d) Processo: atividades e métodos utilizados no desenvolvimento do software;
    e) Tratamento - ação a ser aplicada em um objeto do estudo experimental. É especificado como um conjunto de variáveis controladas de acordo com o planejamento do experimento (fatores). Para tal, observam-se níveis nessa variação. Esses níveis são os valores possíveis de um fator durante o estudo. O tratamento define os fatores e os níveis que serão aplicados para um elemento da amostra da população.
- Aplicador - pessoa que realiza a aplicação de um método na unidade experimental.
- Variáveis - características controladas e medidas em um estudo experimental.

2.1.2.1 Qualidade

A avaliação da qualidade do experimento, através das suas evidências, também deve ser mensurada. Para isso, podemos utilizar os seguintes critérios:

a) Força: caracterizada por três elementos:
    1. Nível – utilizado como indicador de como o tendenciamento foi eliminado do projeto do experimento. Está relacionado à escolha do projeto.
    2. Qualidade – método de investigação utilizado para reduzir o tendenciamento.
    3. Precisão estatística – corresponde ao intervalo de confiança.
b) Tamanho do efeito: distância para o valor nulo do efeito estimado do tratamento e a inclusão de efeitos importantes no intervalo de confiança.
c) Relevância: quão apropriados são os resultados medidos, na prática.

Portanto, seguindo esses critérios, chegamos à questão da garantia da qualidade dos estudos empíricos.

2.1.2.1.1 Garantia da qualidade

Kitchenham *et al*. (2010) relatam que o objetivo da Engenharia de Software experimental é prover os meios para que as melhores evidências científicas sejam combinadas com experiências práticas e valores humanos no processo de decisão relacionado ao desenvolvimento de software.

Mesmo que a execução de estudos experimentais seja a mais cuidadosa possível, existem riscos aos seus resultados. Por isso, as ameaças à validade dos mesmos devem ser identificadas e relatadas. A confiança atribuída às evidências obtidas será maior quanto maior for o rigor do estudo, mesmo assim nunca é 100%. Para reduzirmos a um mínimo aceitável a probabilidade de fracasso e decidir se o resultado de um experimento em Engenharia de Software é aceitável é necessário julgamento humano fundamentado. Esta fundamentação detectará as razões do fracasso, o que fará com que em uma situação futura, haja maior probabilidade de sucesso.

2.1.3 Métodos empíricos em Engenharia de Software

Por causa da *string* de busca utilizada nesta pesquisa, que procura por experimentos, os únicos tipos de métodos empíricos encontrados nos estudos primários foram: estudo de

caso, experimento e *survey*. Por essa razão, somente esses três métodos serão contemplados a seguir.

2.1.3.1 Estudo de Caso

De acordo com Yin (2009), o método empírico denominado estudo de caso é uma pesquisa empírica que investiga um fenômeno contemporâneo dentro de um contexto real. Essa afirmação ganha mais força ainda quando as fronteiras entre o fenômeno e o contexto não estão totalmente evidentes.

Segundo Flyvbjerg (2006), estudos de caso oferecem uma profunda compreensão de como e por que certos fenômenos ocorrem e revelam os mecanismos de como relações causa-efeito acontecem. Pode-se dizer ainda que estudos de caso exploratórios são utilizados para derivar novas hipóteses e criar teorias a partir de investigações iniciais e que estudos de casos confirmatórios são usados para testar teorias existentes. Os dois tipos de estudos citados são muito importantes para refutar teorias, pois um estudo de caso detalhado de uma situação real, na qual a teoria falha, pode ser mais convincente que experimentos que falham no laboratório. Os conhecimentos obtidos a partir de estudos de casos confirmatórios são frequentemente utilizados quando se quer escolher entre teorias rivais.

Conforme Easterbrook *et al*. (2008), uma questão de pesquisa precisa, discriminando como ou por que certos fenômenos ocorrem, é pré-condição para a realização de um estudo de caso. Esta pré-condição é utilizada para derivar uma proposição (do estudo) que afirme precisamente o que o estudo tem a intenção de mostrar e com isto, guiar a seleção de casos e tipos de dados a serem coletados.

Ainda de acordo com Easterbrook *et al*. (2008), um passo importantíssimo na pesquisa de estudos de caso é a seleção dos casos, pois a pesquisa do estudo de caso usa amostragem intencional em vez de amostragem aleatória, com a intenção de selecionar os casos que são mais relevantes para a proposta do estudo, pois às vezes, um único caso pode ser suficiente se o caso for crítico para testar uma teoria bem formulada. Se a teoria valer para este caso, poderá valer para outros ou porventura, se o caso for extremo ou único, ele poderá gerar percepções interessantes sobre o que acontece em condições extremas, como uma crise por exemplo. Entretanto, para que se obtenha maior validade é melhor projetar diversos casos.

Diferentes casos são mais bem entendidos como replicações, em vez de partes de uma amostra. No caso de estudos de caso confirmatórios, estes podem ser escolhidos como replicações literais, onde se espera que cada caso mostre os mesmos resultados, ou como replicações teóricas, onde se espera que os casos mostrem resultados contrastantes por razões previsíveis. Um estudo de caso múltiplo pode incluir especialistas e novatos, a fim de confirmar que a teoria explica adequadamente ambos (Easterbrook *et al*., 2008).

Segundo Easterbrook *et al*. (2008), estudos de caso utilizam diversas fontes de dados. O papel central da pesquisa é desempenhado pelos dados qualitativos, incluindo-se neste caso entrevistas e observação, pois oferecem informações detalhadas sobre o caso. A coleta de dados é realizada levando-se em consideração uma bem definida unidade de análise. Para a Engenharia de Software, uma unidade de análise pode ser: uma empresa, um projeto, uma equipe, um desenvolvedor, um episódio particular ou evento, um produto específico, etc. Para garantir-se que o estudo se concentre no fenômeno pretendido é importante a escolha de uma unidade de análise apropriada.

Easterbrook *et al*. (2008), sugere que o estudo de caso é mais apropriado quando o reducionismo de experimentos controlados é inapropriado. Neste caso, situações em que se espera que o contexto desempenhe um papel no fenômeno ou quando os efeitos esperados são amplos ou demorados. A principal fraqueza dos estudos de caso é que a coleta e a análise de dados estão mais abertas à interpretação e ao viés do pesquisador, por isso, uma estrutura explícita é necessária para selecionar casos e coletar dados. Embora um estudo de caso individual revele profundos conhecimentos, a validade dos resultados depende de uma estrutura mais ampla de indução empírica. Por exemplo, em estudos de casos confirmatórios, as evidências se confirmam quando estudos de caso subsequentes também apoiam a teoria e / ou falham em apoiar teorias rivais.

Finalizando, embora diferentes posturas afetem a maneira como os casos são selecionados e a análise de dados é realizada, os estudos de caso podem ser aplicados a diferentes posições filosóficas, por exemplo, estudos de casos confirmatórios baseiam-se na perspectiva positivista da pesquisa conduzida pela teoria, mas os positivistas também usam estudos de caso exploratórios para desenvolver novas teorias. Os construtivistas usam estudos de caso exploratórios para investigar as diferenças de cultura e perspectiva em vários contextos. Os teóricos críticos usam os dois tipos de estudo de caso para chamar a

atenção para situações que são consideradas problemáticas, selecionando casos que são politicamente importantes ou pelos quais os próprios participantes podem se beneficiar mais.

2.1.3.2 Experimento

Segundo Travassos *et al*. (2002), experimentos verificam as teorias. Eles são o centro dos processos científicos e exploram fatores críticos, iluminando o fenômeno novo, para que as teorias sejam reformuladas e corrigidas. A atividade humana pode ser avaliada pelo método sistemático, disciplinado, computável e controlado da experimentação.

Travassos *et al*. (2002) questiona se a Engenharia de Software pode ser considerada como ciência ou engenharia. Se for considerado o processo de criação do produto (software), a Engenharia de Software pode ser considerada engenharia, pois possui características de produção. Se forem considerados os aspectos de melhoria contínua e sequencial da qualidade do processo e produto, a Engenharia de Software pode ser considerada ciência. Portanto, para estabelecer-se uma base de engenharia e ciência para a Engenharia de Software, metodologias específicas são necessárias.

Conforme Wohlin *et al*. (2012), existem quatro métodos relevantes para condução de experimentos na Engenharia de Software:

- Científico: através de observação, sugere o modelo ou a teoria de comportamento. Mede, analisa e verifica as hipóteses do modelo ou teoria. É uma abordagem para construção de modelos.
- De Engenharia: através de observação das soluções existentes, propõe as mais adequadas. Desenvolve, mede, analisa e repete até que nenhuma melhoria adicional seja possível. É uma abordagem orientada à melhoria evolutiva.
- Experimental: inicia com a sugestão de um modelo novo, não necessariamente baseado em um modelo existente. Desenvolve o método qualitativo e/ou quantitativo, aplica um experimento, mede, analisa e avalia o modelo, repetindo o processo. É uma abordagem orientada à melhoria revolucionária.
- Analítico ou matemático: propõe uma teoria formal, desenvolve a teoria, deriva resultados, comparando-os com observações empíricas. É uma abordagem dedutiva que não necessita de um projeto experimental no sentido estatístico, mas disponibiliza uma base analítica para o desenvolvimento de modelos.

Os experimentos verificam a precisão teórica contra a realidade, mas não oferecem provas com certeza absoluta.

2.1.3.3 Survey

Segundo Kitchenham *et al*. (2002), pesquisa de opinião (do inglês, *survey research*) é usada para identificar as características de uma ampla população de indivíduos. O uso de questionários, para coleta de dados, é frequentemente associado a este tipo de pesquisa. No entanto, pesquisas de opinião também podem ser conduzidas usando entrevistas estruturadas ou técnicas de registro de dados. A característica principal que define a pesquisa de opinião é a seleção de uma amostra representativa de uma população bem definida e as técnicas de análise de dados usadas para "generalizar" a partir dessa amostra para a população.

Ainda segundo Kitchenham *et al*. (2002), uma pré-condição para a realização de pesquisas de opinião é uma pergunta clara que questiona a natureza de uma determinada população-alvo. Como geralmente é inviável (e desnecessário) entrevistar todos os membros dessa população, a pesquisa de opinião identifica primeiramente um subconjunto representativo como amostra e determina como alcançar esse subconjunto para fazer a coleta de dados. A identificação da unidade de análise é importante para determinar uma técnica de amostragem apropriada. Por exemplo, se a pergunta da pesquisa for sobre empresas de software, então a amostragem sobre desenvolvedores individuais poderá resultar em uma amostra tendenciosa, com algumas empresas sendo super-representadas, pois vários desenvolvedores da mesma empresa foram incluídos. Além disso, a amostragem aleatória simples da população também pode ser inadequada. A título de exemplo, se nossa unidade de análise for "desenvolvedores individuais", uma amostragem aleatória pode acabar com a maioria ou todos os entrevistados trabalhando em uma única empresa dominante. Nesse caso, técnicas de amostragem estratificadas devem ser usadas para identificar subgrupos dentro da população, para que possamos ter amostras dentro de cada subgrupo.

Conforme Kitchenham *et al*. (2002), um grande desafio na pesquisa de opinião é controlar o viés de amostragem, pois ele causa problemas na generalização dos resultados da pesquisa, porque os entrevistados da pesquisa podem não ser representativos da população-alvo. Baixas taxas de resposta aumentam o risco de viés. Em geral, é difícil obter

altas taxas de resposta, a menos que incentivos significativos possam ser oferecidos para a participação, embora às vezes seja possível entrar em contato com não entrevistados para avaliar se ocorreu um viés de resposta sistemática.

De acordo com Kitchenham *et al*. (2002), um desafio maior ainda é garantir que as perguntas sejam formuladas de maneira a produzir dados úteis e válidos. Pode ser difícil formular as perguntas de modo que todos os participantes as compreendam da mesma forma, especialmente se a população-alvo for diversificada. Além disso, é possível que o que as pessoas dizem que fazem em resposta às perguntas da pesquisa não tenha relação com o que elas realmente fazem, porque elas são incapazes de fazer uma introspecção confiável em suas práticas de trabalho.

Em harmonia com Kitchenham *et al*. (2002), é instrutivo comparar pesquisa de opinião com outros métodos empíricos. Se o pesquisador estiver preocupado em estabelecer a verdade para um determinado grupo em geral ou obter uma compreensão mais profunda sobre como este grupo se comporta, é melhor que o pesquisador faça um estudo de caso. Isso sacrificaria alegações de representatividade (porque estudos de caso não usam amostragem representativa) em troca de conhecimentos mais profundos sobre o que acontece em um pequeno número de casos selecionados. Por outro lado, se o pesquisador estiver mais interessado em saber como, por exemplo, um determinado fenômeno altera a forma como o grupo compartilha informações, ele pode criar um experimento ou um quase experimento para testar um relacionamento causal.

Ainda conforme Kitchenham *et al*. (2002), pesquisa de opinião se encaixa quase exclusivamente na tradição positivista. O desejo de caracterizar uma população inteira por meio de técnicas de amostragem requer uma crença no reducionismo e uma preocupação com teorias generalizáveis. Se o pesquisador estiver mais interessado em entender a cultura do compartilhamento de informações dentro da população alvo, ele pode adotar uma postura construtivista e usar a etnografia ou a pesquisa-ação.

Segundo Dias e Silva (2010), realiza-se uma pesquisa de opinião quando se tem a necessidade de fazer uma pesquisa que envolva grandes parcelas de dados quantitativos. O projeto da pesquisa de opinião, que deverá ser desenvolvido após a definição do objetivo e da metodologia que serão utilizadas na pesquisa, será composto das atividades: (i) definir regras para localizar os possíveis respondentes; (ii) projetar as amostras apropriadas; (iii)

estabelecer o método de coleta de dados; (iv) desenhar e pré-testar o questionário a ser utilizado nas entrevistas que serão realizadas; (v) estabelecer o plano de análise dos dados coletados.

2.1.4 Ameaças à validade de estudos empíricos

PERRY *et al*. (2000) define ameaças à validade de um estudo empírico como sendo as influências que podem limitar nossa capacidade de interpretar ou tirar conclusões a partir de dados do estudo.

De acordo com Wohlin *et al*. (2012), uma questão fundamental a respeito dos resultados de um experimento é se estes resultados são realmente válidos. A validade dos resultados deve ser levada em consideração na fase de planejamento do estudo, pois esta questão pode influenciar sobremaneira a validade do resultado. As amostras que serão utilizadas no estudo são retiradas de uma população de interesse, e para que seu resultado seja adequado ele tem de ser válido para essa população. Posteriormente, os resultados do estudo podem ser generalizados para uma população mais ampla. Os resultados terão validade adequada e mais confiável, se forem válidos para uma população generalizada.

Se um experimento for realizado em um âmbito restrito (dentro de uma organização, por exemplo) e seu projeto for feito para responder questões inerentes ao âmbito em questão, seus resultados são válidos naquele âmbito. Nesse caso, validade adequada não significa validade generalizada. Mas se conclusões gerais serão tiradas do experimento, sua validade terá que abranger um âmbito mais geral.

De acordo Campbell e Stanley (1963) e Cook e Campbell (1979), existem quatro tipos de ameaças à validade de um experimento:

1. Validade de conclusão: possibilidade de tirar conclusões imprecisas das observações;
2. Validade Interna: ameaças que podem ter afetado os resultados e não foram devidamente levadas em conta;
3. Validade de construção: ameaças sobre a relação entre a teoria e a observação;
4. Validade externa: ameaças que afetam a generalização dos resultados.

2.1.4.1 Descrição detalhada das ameaças à validade de estudos empíricos

A lista de ameaças à validade abaixo foi feita com base nos experimentos discutidos por Cook e Campbell (1979) e será utilizada como uma lista de verificação para as ameaças à validade deste estudo.

2.1.4.1.1 Validade de conclusão

Ameaças à validade de conclusão dizem respeito à capacidade de tirar a conclusão correta sobre as relações entre o tratamento e o resultado de um experimento.

A validade de conclusão é subdividida em: baixo poder estatístico, suposições violadas dos testes estatísticos, pesca e taxa de erro, confiabilidade de medidas, confiabilidade da implementação do tratamento, irrelevâncias aleatórias na configuração experimental e heterogeneidade aleatória de assuntos.

a) Baixo poder estatístico - um teste estatístico poderoso revela um padrão verdadeiro através dos dados. Se o poder é baixo existe alto risco de se tirar uma conclusão errada ou rejeitar uma hipótese errada;
b) Suposições violadas dos testes estatísticos - certos testes supõem amostras independentes e normalmente distribuídas, por exemplo. Violar as suposições pode levar a conclusões erradas;
c) Pesca e taxa de erro - é composta por duas partes:
    1. Procurar (pescar) um resultado específico é uma ameaça, pois há influência dos pesquisadores nos resultados, que estão procurando por um resultado específico, fazendo com que as análises não sejam independentes;
    2. Taxa de erro, que está relacionada ao nível verdadeiro de significância. Exemplo: se realizarmos três investigações com um nível de significância de 0:05, significa que o nível de significância total é $1 – (1 – 0,05)^3 = 0,14$. Por isso, a taxa de erro (nível de significância) deve ser ajustada quando se realiza múltiplas análises.
d) Confiabilidade de medidas - depende de fatores diferentes como: formulação pobre de questões, má instrumentação ou instrumentação com *layout* ruim. O princípio básico diz que se medirmos um fenômeno duas vezes, o resultado deve ser o mesmo. Isto significa que medidas objetivas, que podem ser

repetidas com o mesmo resultado, são mais confiáveis que medidas subjetivas;

e) Confiabilidade da implementação do tratamento - existe o risco de que diferentes pessoas aplicando um mesmo tratamento a um assunto o façam de maneira diferente. Esse risco ocorre também se o tratamento for aplicado em diferentes ocasiões. A implementação deve ser o mais padrão possível se aplicada por pessoas diferentes ou em ocasiões diferentes;

f) Irrelevâncias aleatórias na configuração experimental - os resultados podem ser perturbados por elementos fora da configuração experimental, tal como uma interrupção súbita da experiência;

g) Heterogeneidade aleatória de assuntos - em um grupo de estudos há sempre heterogeneidade. Sendo assim, existe o risco que variações devidas a diferenças entre indivíduos sejam maiores que o tratamento. Ao mesmo tempo, se a escolha for por grupos mais homogêneos pode afetar a validade externa.

2.1.4.1.2 Validade interna

São influências de causa-efeito que podem afetar a variável independente sem o conhecimento do pesquisador, por isso ameaçam a conclusão tirada do experimento.

A validade interna é subdividida em Ameaças de grupo único, Ameaças de múltiplos grupos e Ameaças sociais.

- Ameaças de grupo único: história, maturação, teste, instrumentação, regressão estatística, seleção, mortalidade, ambiguidade sobre a direção da influência causal.
- Ameaças de múltiplos grupos: interação com a seleção.
- Ameaças sociais: difusão ou imitação de tratamentos, equalização compensatória de tratamentos, rivalidade compensatória, ressentimento desmoralizado.

a) Ameaças de grupo único - aplicam-se a experiências feitas com um único grupo, pois neste caso, não existe um grupo de controle ao qual não é aplicado o tratamento. Assim, há problemas para se determinar se o tratamento ou outro fator causou efeito observado;

b) História - em uma experiência, diferentes tratamentos podem ser aplicados para o mesmo objeto em momentos diferentes. Como as circunstâncias não são as mesmas em ambas as ocasiões existe o risco de a história afetar os resultados experimentais;
c) Maturação - este efeito surge quando os indivíduos reagem de forma diferente conforme o tempo passa. Exemplo: os indivíduos são afetados negativamente (cansados ou entediados) ou positivamente (aprendizagem) durante o decurso da experiência;
d) Teste - se o teste é repetido os envolvidos podem responder diferentemente diferentes vezes, pois eles sabem como o teste é realizado. Se existe uma necessidade de familiarização com os testes, é importante que os resultados do teste não sejam mostrados para os envolvidos para não criar aprendizagem não intencional;
e) Instrumentação - este é o efeito provocado pelos artefatos usados para a execução da experiência, pois se mal projetados, o experimento é afetado negativamente. Pode-se citar como exemplos: formulários de coleta de dados, documento a ser verificado em uma experiência de inspeção, etc;
f) Regressão estatística - é uma ameaça quando os objetos do estudo fazem parte de um grupo de experimentos ou estudo de caso e foram classificados anteriormente. Por exemplo, os estudos foram classificados entre os 10 melhores ou 10 piores. Os 10 piores estudos classificados previamente em uma experiência, devido à variação aleatória, provavelmente não constarão como os 10 piores em uma nova lista de classificação de uma nova experiência. Então eles só poderão ser classificados em melhores posições do que estavam anteriormente. Neste caso houve melhoria sem nenhum tratamento aplicado aos estudos;
g) Seleção - é o efeito da variação natural do desempenho humano. Dependendo de como os envolvidos são selecionados de um grupo maior, os efeitos da seleção podem influenciar a experiência. Se incluirmos voluntários na experiência, os resultados serão ainda mais influenciados, visto que os voluntários são geralmente mais motivados e mais adequados para uma nova

tarefa do que o antigo grupo. Dito isto, o grupo selecionado não representa toda a população;

h) Mortalidade - este efeito está relacionado à representatividade do abandono da experiência por tipos diferentes de envolvidos. Se, por exemplo, em um experimento de inspeção, pessoas de maior experiência o abandonam, a validade do experimento é altamente afetada. Assim sendo, é importante caracterizar o abandono, a fim de verificar quão representativo ele é do total da amostra;

i) Ambiguidade da direção da influência causal - esta é a questão de saber se A causa B ou B causa A ou mesmo X causa A e B. Por exemplo: se a relação entre a complexidade de um programa e a taxa de erro encontrada é observada. A questão é se a alta complexidade de um programa causa alta taxa de erros ou vice-versa, ou ainda se a alta complexidade do problema a ser resolvido causa os dois.

   A maioria das ameaças à validade interna pode ser tratada através do projeto de experimentação. Introduzindo um grupo de controle muitas das ameaças internas podem ser controladas. Por outro lado, vários grupos de ameaças podem ser introduzidos também;

j) Ameaças de vários grupos - diferentes grupos são estudados em uma experiência com múltiplos grupos. A ameaça a esses estudos é que o grupo de controle e os grupos selecionados para o experimento podem ser afetados de forma diferente pelas ameaças de grupo único, como definido anteriormente. Assim, há interações com a seleção;

k) Interações com a seleção - as interações com a seleção são devido a diferentes comportamentos em diferentes grupos. Isto significa que diferentes grupos amadurecem em velocidades diferentes. Por exemplo, se dois grupos aplicam um novo método e um grupo aprende o método novo mais rápido que o outro, eles amadurecem diferentemente;

l) Ameaças sociais à validade interna - são aplicáveis a um único grupo e a múltiplos grupos de experiências, por exemplo: em de uma experiência de

inspeção em que um novo método (leitura baseada em perspectiva) é comparado a um velho método (leitura à base de lista de verificação);

m) Difusão ou imitação de tratamentos - este efeito ocorre quando um grupo de controle aprende sobre o tratamento a partir do grupo de estudo experimental, ou tenta imitar o comportamento do grupo de estudo. Por exemplo, se um grupo de controle utiliza um método de inspeção baseado em uma lista de verificação e o grupo de experimentação usa métodos baseados em perspectiva, o primeiro grupo pode ouvir sobre o método baseado em perspectiva e realizar suas inspeções influenciadas pela sua própria perspectiva;

n) Equalização compensatória de tratamentos - se a um grupo de controle é dada alguma compensação por ser um grupo de controle, como um substituto porque eles fazem tratamentos, isto pode afetar o resultado da experiência. Se ao grupo de controle é ensinado outro novo método como uma compensação, por não ter sido ensinado o método baseado em perspectiva, seu desempenho pode ser afetado por esse novo método.

o) Rivalidade compensatória - um indivíduo que recebeu tratamentos menos desejáveis pode ser motivado a reduzir ou reverter o resultado esperado do experimento. O grupo que utiliza o método tradicional pode fazer o seu melhor para mostrar que o método antigo é competitivo;

p) Ressentimento desmoralizado - um indivíduo que recebe tratamentos menos desejáveis pode desistir e não ter um desempenho tão bom quanto ele geralmente tem. O grupo que utiliza o método tradicional não está motivado a fazer um bom trabalho, todavia, aprender algo novo inspira o grupo que usa o novo método.

2.1.4.1.3 Validade de construção

Validade de construção diz respeito à generalização do resultado do experimento com o conceito ou teoria por trás do experimento. Algumas ameaças referem-se ao projeto do experimento, outras a fatores sociais.

a) Ameaças de projeto - as ameaças de projeto à validade de construção abrangem questões que estão relacionadas ao projeto da experiência e a sua capacidade de refletir a construção a ser estudada;
b) Explicação pré-operacional inadequada de construções - significa que as construções não estão suficientemente definidas, antes de serem traduzidas em medidas ou tratamentos. A teoria não é clara suficiente, e, por conseguinte, a experiência não é suficientemente clara. Por exemplo, se dois métodos de inspeção são comparados e não é claro o suficiente o que ser "Melhor" significa? Será que isso significa encontrar mais falhas? Mais falhas por hora ou falhas mais graves?
c) Tendência à mono-operação - se o experimento inclui uma única variável independente, caso, assunto ou tratamento, o experimento pode sub-representar a construção e assim não dar a imagem completa da teoria. Por exemplo, se uma experiência de controle é realizada com um único documento como objeto, a causa da construção é sub-representada;
d) Tendência à monométodo - se utilizarmos um único tipo de medida ou observação existe o risco desta medida ou observação ser tendenciosa, com isto o experimento será enganoso. Quando envolvemos diferentes tipos de medidas e observações elas podem ser cruzadas entre si. Por exemplo, se em um experimento de inspeção o número de falhas encontradas é medido e a classificação do erro é baseada em julgamento subjetivo, as relações podem não ser suficientemente explicadas. Então, o experimentador pode tendenciar as medidas;
e) Confundindo construções e níveis de construções - em algumas relações não é essencial à presença ou ausência de uma construção. O importante para o resultado é o nível da construção. O efeito da presença da construção é confundido com o efeito do nível da construção. Por exemplo, a presença ou ausência de conhecimento prévio em uma linguagem de programação pode não explicar as causas de um experimento, mas a diferença pode depender se os programadores têm 1, 3 ou 5 anos de experiência na linguagem;

f) Interação de diferentes tratamentos - se uma pessoa está envolvida em mais de um estudo, os tratamentos a partir dos diferentes estudos podem interagir. Então você não pode concluir se o efeito é devido a qualquer dos tratamentos ou de uma combinação de tratamentos;

g) Interação de testes e tratamento - o próprio teste, isto é, a aplicação de tratamentos, pode tornar os indivíduos mais sensíveis ou receptivos ao tratamento. Então o teste é uma parte do tratamento. Por exemplo, se o teste envolve a medição do número de erros cometidos na codificação, então, os programadores serão mais conscientes desses erros, e, assim, tentaram reduzi-los;

h) Generalização restrita através de construções - o tratamento pode afetar a construção do estudo positivamente, mas não intencionalmente afetar negativamente outras construções. Esta ameaça torna o resultado difícil de generalizar para outros resultados potenciais. Por exemplo, um estudo comparativo conclui que se consegue melhorar a produtividade com um novo método. Por outro lado, observa-se que reduz a manutenabilidade, o que é um efeito colateral indesejado. Se a manutenabilidade não é medida ou observada, existe o risco de que as conclusões sejam tiradas com base no atributo de produtividade, ignorando a manutenabilidade;

i) Ameaças sociais à validade de construção - o comportamento dos envolvidos e experimentadores pode mudar baseado no fato de fazerem parte de um experimento, o que dá falsos resultados na experiência;

j) Hipótese de adivinhação - quando as pessoas tomam parte em uma experiência, elas podem tentar descobrir quais são o propósito e o resultado pretendido da experiência. Então, elas são susceptíveis de basear seu comportamento em suas suposições sobre as hipóteses, seja positiva ou negativamente, dependendo de sua atitude para com a hipótese prevista.

k) Apreensão de avaliação - algumas pessoas têm medo de ser avaliadas. A tendência humana é tentar parecer melhor quando está sendo avaliada, o que é confundido com os resultados da experiência. Por exemplo, se diferentes modelos de estimação são comparados, as pessoas não reportam

seus verdadeiros desvios entre estimativa e resultado, mas falsos, ocasionando melhores valores;

l) Expectativas do experimentador - os experimentadores podem influenciar os resultados de um estudo consciente e inconscientemente com base no que eles esperam da experiência. A ameaça pode ser reduzida envolvendo diferentes pessoas que não têm ou têm expectativas diferentes para o experimento. Por exemplo, questões podem ser criadas de maneiras diferentes, a fim de fornecer as respostas que você quer.

2.1.4.1.4 Validade externa

Ameaças à validade externa são condições que limitam a capacidade de generalizar os resultados da experiência com a prática industrial. Existem três tipos de interações com o tratamento: pessoas, local e horário.

a) Interação entre seleção e tratamento - esta ameaça representa ter pessoas erradas participando do experimento, isto é, utilizar no experimento uma população não representativa da população que queremos generalizar. Por exemplo, em um experimento de inspeção da qual participam programadores, testadores e engenheiros de sistemas, selecionar somente programadores;
b) Interação entre configuração e tratamento - este efeito representa lugar ou ambiente errado, isto é, não ter a configuração experimental ou o material representativo, por exemplo, em uma prática industrial. Exemplo: utilizar ferramentas ultrapassadas em um experimento quando ferramentas atualizadas são comuns na indústria;
c) Interação entre história e tratamento - este efeito é observado quando o experimento é realizado em um momento especial ou em um dia que afeta os resultados. Se, por exemplo, um questionário é realizado em sistemas críticos de segurança poucos dias depois de um grande problema de software, as pessoas tendem a responder de forma diferente do que alguns dias antes do problema, ou algumas semanas ou meses após o problema.

Ameaças à validade externa são reduzidas tornando o ambiente experimental o mais realista possível. Por outro lado, a realidade não é homogênea. O mais importante é

caracterizar e relatar as características do meio ambiente como: experiência da equipe, ferramentas e métodos, a fim de avaliar a aplicabilidade em um contexto específico.

2.1.4.1.5 Prioridade entre os tipos de ameaças à validade

Pode haver um conflito entre os quatro tipos de ameaças à validade (interna, externa, conclusão e construção), pois se aumentarmos um tipo o outro pode decrescer. Por isso, devemos otimizar a priorização das ameaças à validade dependendo do propósito do experimento. Por exemplo, se em um experimento de inspeção forem utilizados estudantes de graduação, isto irá, provavelmente, aumentar o tamanho do grupo de estudos, reduzindo a heterogeneidade e dando tratamento fiável a implementação. Isto aumenta a validade de conclusão, enquanto a validade externa é reduzida, já que a seleção não é representativa, se quisermos generalizar os resultados para a indústria de software.

Cook e Campbell (1979) propõem as seguintes prioridades para testes de teoria e pesquisa aplicada:

a) Teste de teoria - neste teste é mais importante mostrar que existe um relacionamento casual (validade interna) e que as variáveis no experimento representam a construção da teoria (validade de construção). Se somarmos à experiência, “tamanho” pode-se resolver os problemas de significância estatística (validade conclusão).

   Teorias são raramente relacionadas a configurações específicas, população ou vezes em que os resultados devem ser generalizados. Portanto, há pouca necessidade de questões de validade externa. As prioridades para experimentos em testes de teoria são, em ordem decrescente: interna, construção, conclusão e externa.

b) A pesquisa aplicada - as prioridades são diferentes na pesquisa aplicada, pois está é a área alvo para a maior parte dos experimentos de Engenharia de Software. Por ser o objetivo principal de um experimento estudar as relações entre causas e efeitos, as relações em estudo são de mais alta prioridade (validade interna). Em pesquisa aplicada a generalização (do contexto em que a experiência é realizada para um contexto mais amplo) também é de alta prioridade (validade externa).

> Para um pesquisador é interessante mostrar que o resultado da pesquisa é válido para empresas de um determinado tamanho ou domínio de aplicação e não que o resultado da pesquisa é válido para uma determinada empresa. Um pesquisador aplicado tem menos interesse em saber qual componente de um tratamento complexo causou o efeito (validade de construção). O interesse principal é o próprio efeito. Por ser difícil, em ambientes práticos, conseguir uma quantidade grande de conjunto de dados, as conclusões estatísticas podem ser tiradas com menos significância (validade de conclusão).
>
> As prioridades para experimentos em pesquisa aplicada são, em ordem decrescente: interna, externa, construção e conclusão.
>
> Pode-se concluir que, durante o planejamento de um experimento, ameaças à validade relacionadas aos resultados experimentais devem ser avaliadas e balanceadas. Diferentes prioridades devem ser dadas a diferentes tipos de validade, dependendo do propósito da experiência.

2.1.5 Testes em softwares

O tema desta dissertação, por si só, define um limite de escopo para este estudo. Porém, é interessante ressaltar que não foi objetivo desta pesquisa esmiuçar em detalhes as questões relacionadas aos testes para aplicações construídas para ambiente *Web*. No entanto, essas questões estão devidamente apresentadas e referenciadas, dando aos leitores interessados a possibilidade de um aprofundamento.

De acordo com Glass *et al*. (2006), a preocupação com a qualidade dos softwares produzidos pelas empresas de Tecnologia da Informação (TI) tem aumentado. Mas, muitas não dão a atenção devida a uma parte importante de um projeto de software, os testes. Diante desta afirmação, vem à pergunta: será que os testes de software praticados pelas empresas de TI efetivamente atendem suas necessidades?

A resposta a esta questão estaria intimamente ligada a um exaustivo estudo feito nas indústrias de software, que levasse em consideração suas respectivas necessidades e especificidades. Entretanto, segundo Tassey (2002), é seguro afirmar que testes mal conduzidos podem ter um impacto real na situação econômica de uma empresa.

Em muitos casos, os testes são realizados manualmente pelos responsáveis pela implementação do software, o que provoca um ciclo vicioso, pois o implementador está

testando seu próprio código, normalmente aplicando casos de testes genéricos que não refletem o comportamento dos usuários finais, que na maioria das vezes não tem vivência com a interface do software implementado (Reason, 1990).

Porém, nem sempre os testes acontecem de acordo com o descrito por Reason (1990), pois muitas empresas têm equipes dedicadas a testes de software, que tem como função primordial encontrar manualmente erros nos softwares de forma que eles não cheguem não conformes ao ambiente de produção. Mas será mesmo necessário que as empresas tenham equipes de testes de software?

Segundo Reason (1990), o avanço das tecnologias e dos métodos formais tem propiciado as indústrias de software dispensar os processos manuais de testes e optar por processos automatizados. Mas, para que se alcance este objetivo é necessário que o projeto tenha uma eficaz especificação formal que garanta a qualidade da própria especificação, pois somente desta forma é possível avançar em paralelo com o desenvolvimento do software e realizar testes automáticos.

2.1.5.1 Testes em aplicações para Web

Segundo Ginige *et al*. (2001), pode-se dizer que o desenvolvimento de aplicações *Web* é caracterizado pelo rápido crescimento de requisitos, conteúdo, funcionalidades e constantes alterações durante o ciclo de vida do software. Esta é uma atividade contínua e sem releases específicos, como ocorre no desenvolvimento de software convencional.

Segundo Pressman (1992), o objetivo do teste de aplicações para *Web* é exercitar cada uma das várias dimensões de qualidade da aplicação com intenção de encontrar erros ou descobrir tópicos que podem levar à falhas de qualidade.

Falsa sensação de segurança; olhar viciado e concentração polarizada (os testes são concentrados em somente um ponto, não abrangendo os demais) são dificuldades encontradas ao testar uma aplicação. Isto sem contar a máxima: se está funcionando, por que testar? A resposta é simples: em algum momento a aplicação vai falhar, é certo.

2.1.5.2 Estratégias e planejamento de testes para aplicações Web

De acordo com Pressman (1992), como estratégia para planejar testes, deve-se ter em mente três itens:

a) Estabelecer o que é crítico;

b) Fazer um mapeamento das partes críticas do software e listar as consequências caso elas não funcionem;
c) Levar sempre em consideração que aplicações para *Web* envolvem vários ambientes (browsers, por exemplo).

2.1.5.3 Testes automatizados

Pressman (1992) afirma que automatizar testes, consiste em escrever um software que prepare um cenário de teste, controle a execução do pedaço do software que contém aquele cenário, compare resultados e forneça relatórios. Com o avanço do campo de testes de softwares, foram criados para o mercado diversos *frameworks* que auxiliam em muito a tarefa de testar.

Os testes automatizados facilitam a tarefa de rodar os testes repetidamente e de forma mais rápida, além da formalização e estruturação dos casos de teste.

A dinâmica desse tipo de teste deve seguir o fluxo: especificação, plano de testes e testes, não podendo ser diferente desta ordem.

2.1.5.4 Tipos de teste

De acordo com Pressman (1992), entre os vários tipos de teste para ambiente *Web* pode-se listar testes para interface, conteúdo, segurança, desempenho, etc. Como exemplo, podem-se listar os seguintes tipos de teste: de conteúdo; de interface; de semântica de interface; configuração; segurança; desempenho; banco de dados; navegação; usabilidade; de compatibilidade.

2.1.5.5 Estilos de teste

Pressman (1992) cita os seguintes estilos de teste:

- Desenvolvimento orientado por comportamento (do inglês, *Behaviour-Driven Development*) ou BDD, que é utilizado em desenvolvimento ágil e visa integrar as regras de negócio com linguagem de programação. Durante o levantamento de requisitos a linguagem de negócio utilizada no BDD é extraída das especificações fornecidas pelos usuários;
- Desenvolvimento orientado por testes (do inglês, *Test-Driven Development*) ou TDD, que é utilizado para testar o software antes do mesmo ser construído. Esta forma faz com que o desenvolvedor programe pensando na interface das classes

e não na implementação das mesmas, o que gera um ganho de qualidade. Neste estilo de teste, o desenvolvedor deverá primeiramente escrever um teste que falhe. Em seguida escrever um código para passar no teste e por fim eliminar a redundância;

- Desenvolvimento orientado a domínio (do inglês, *Domain-Driven Design*) ou DDD, que é uma abordagem de projeto de software de forma disciplinada, utilizando uma série de conceitos e técnicas focadas no domínio do software. O cerne do DDD é criar um modelo do domínio do negócio do software, utilizando um conjunto de boas práticas de desenvolvimento de software. Estas boas práticas são as da literatura: o software deverá ser extensível, escalável, confiável, seguro, fácil de manter, fácil de utilizar, flexível e deverá atender aos requisitos demandados.

Pode-se dizer que a evolução tecnológica das aplicações *Web* aumentou a demanda por técnicas e ferramentas que abordam o problema da garantia de qualidade dessas aplicações e que a existência de diferentes métodos e tipos de aplicações que envolvem a *Web* aumenta a dificuldade da seleção de técnicas de teste para um projeto de software *Web*.

3 Método de pesquisa

3.1 Planejamento

3.1.1 Formulação das questões de pesquisa

De acordo com o problema relatado, esse estudo tem o objetivo de responder às questões de pesquisa apresentadas nas seções subsequentes:

3.1.1.1 Questão geral

Até que ponto os experimentos em Engenharia de Software, cujo tema é "Testes em aplicativos desenvolvidos para ambiente *Web*", estão cumprindo os rigores científicos necessários para serem considerados válidos?

3.1.1.2 Questões específicas

a) Qual a quantidade de artigos encontrados, baixados e selecionados para esta dissertação?

Dentro do universo de artigos selecionados:

b) Quantos e qual percentual de artigos referenciam as suas ameaças à validade e quantos não o fazem?
c) Qual a frequência anual de artigos que faz referência às suas ameaças à validade e quantos não fazem?
d) Qual a quantidade e percentual de artigos divididos entre os três critérios da abordagem simples (critério não atendido, critério parcialmente atendido e critério totalmente atendido)?
e) Para artigos que mencionam suas ameaças à validade, quais são as ameaças mais citadas?
f) Quantos e qual percentual de artigos (total e por ano de publicação) mencionam suas ameaças à validade de forma sumarizada e de forma detalhada?
g) Qual a quantidade e percentual de artigos que classificam suas ameaças formalmente e Informalmente (total e por ano de publicação)?

h) Qual a quantidade e percentual de artigos por tipo de publicação?
i) Qual a quantidade e percentagem de artigos que cumprem integralmente os rigores científicos necessários para serem considerados válidos, de acordo com a literatura especializada?

3.1.2 Formulação da *String* de busca

3.1.2.1 String de busca de artigos

Kitchenhan *et al.* (2007) salienta que as *strings* de busca devem ser construídas a fim de selecionar estudos que sirvam de base para responder as questões da pesquisa e que em alguns casos, são necessárias algumas adaptações nas mesmas para atender as particularidades do mecanismo de pesquisa das máquinas de busca que serão utilizadas. Fazem parte da composição das *strings* de busca operadores booleanos AND ou OR e particularidades de cada máquina de busca. No entanto, deverá ser criada uma *string* de busca genérica que servirá de base para a pesquisa.

Para ter acesso aos artigos primários utilizados neste estudo, foi feita uma busca automatizada na *Web*.

Nesta busca foi utilizada a *string*: "*Web* application testing" OR "*Web* applications tests" OR "test of *Web* applications" OR "testing *Web* applications" OR "tests for *Web* applications" AND experiment.

Não foram inseridos na *string* de busca os complementos: "validity threats" ou "threats to validity", pois como este é um estudo quantitativo causaria tendência, trazendo artigos que obrigatoriamente fizessem referência as suas ameaças à validade.

3.1.3 Definição da estratégia de busca

3.1.3.1 Máquinas de busca

Finalizadas as *strings* de busca, deve-se selecionar as máquinas de busca (bases de dados) que serão utilizadas para a obtenção dos estudos primários que serão a base da pesquisa.

A seleção das máquinas de busca levou em consideração seu mecanismo de busca através de palavras chave (*strings* de busca) para encontrar e posteriormente baixar arquivos da *Web*.

Na *Web* existem documentos cujo download é público e documentos, que para serem baixados, é necessário que se pague por eles. Nesta pesquisa só foram baixados documentos cujo *download* é público (gratuito), pois se deseja que todas as pessoas que queiram reproduzir este estudo possam fazê-lo, e o custo do *download* de artigos pode ser um impedimento.

As máquinas de busca utilizadas nesta pesquisa foram:

a) http://ieeexplore.ieee.org
b) http://dl.acm.org
c) http://scholar.google.com
d) http://academic.research.microsoft.com
e) http://citeseerx.ist.psu.edu
f) http://www.sciencedirect.com

3.1.4 Busca das fontes de pesquisa

As fontes de pesquisa para este estudo são publicações sobre testes de software construídos para ambiente *Web*, que possuam um ou mais experimentos próprios.

3.1.5 Busca dos estudos primários

A coleta de dados que será documental indireta, tanto dos dados que servirão de fonte primária como os de fonte secundária, dar-se-á através das principais máquinas de busca (bases de dados) existentes na *Web*, por meio de "*strings* de busca" específica para cada máquina de busca.

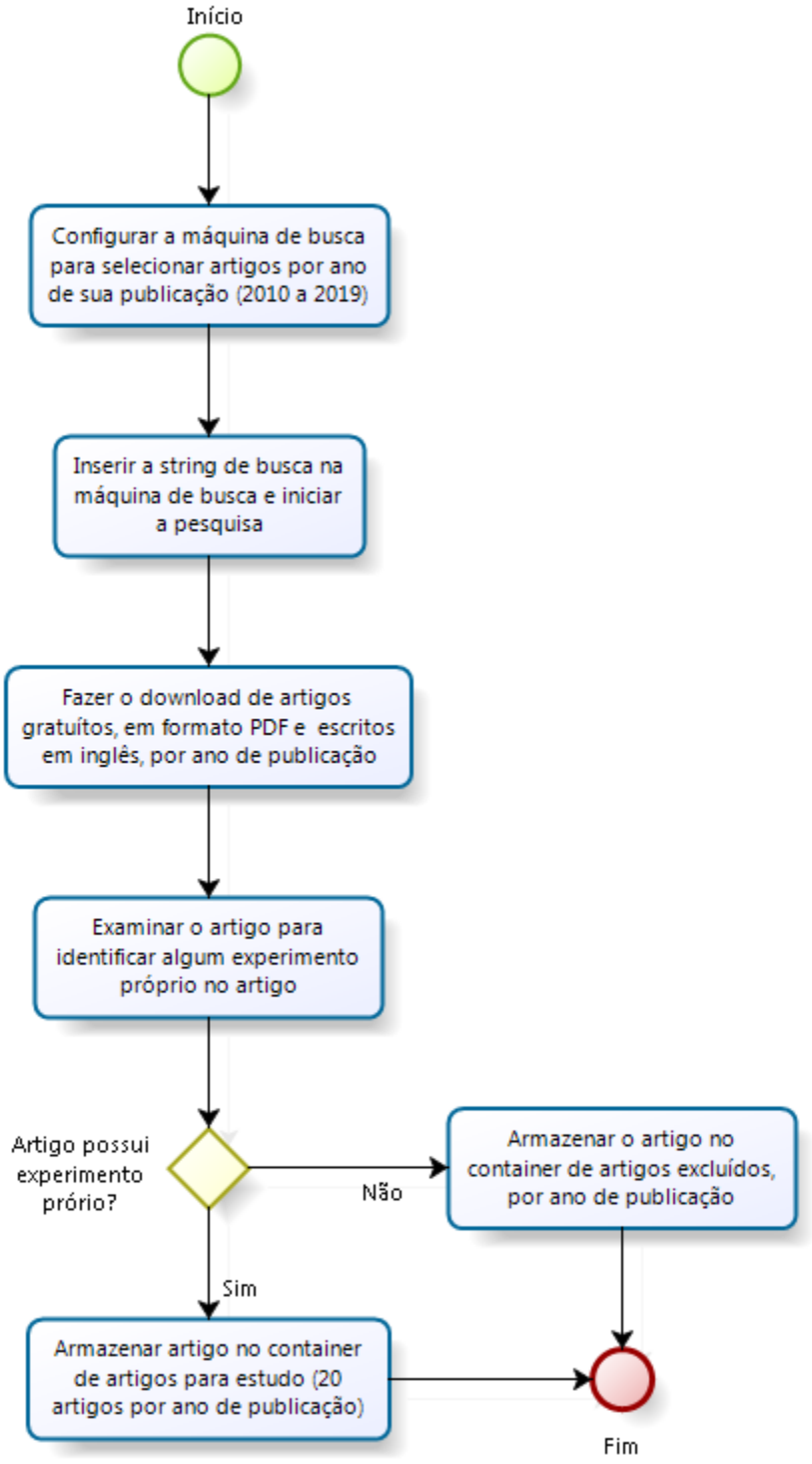
3.1.5.1 Processo de download e seleção de artigos para estudo

A regra principal para a inclusão de um artigo neste trabalho foi sua relevância para as questões de pesquisa definidas no item 3.1.1. Porém, os seguintes critérios também foram levados em consideração para a escolha dos artigos:

a) O artigo deveria estar em formato **.PDF** (acrônimo do termo em inglês, *Portable Document Format*) e, conforme já mencionado, disponível publicamente (gratuitamente) na *Web* para *download*.
b) O período de publicação do artigo para seleção foi de 2010 a 2019.
c) O artigo deveria ter sido escrito em língua inglesa.

d) O tema do artigo deveria ser “Testes para aplicações *Web*” e no artigo deveria constar pelo menos um experimento próprio.

Figura 1 – Processo de seleção dos estudos primários para investigação



Observações:

1. A “Figura 1 – Processo de seleção dos estudos primários para investigação” apresenta o procedimento de seleção de artigos.
2. Existem artigos que referenciam experimentos, mas não criam/fazem nenhum experimento próprio. Neste caso, estes artigos foram descartados.
3. Obtiveram-se vinte artigos de cada ano (2010 a 2018), com exceção do ano de 2019 em que foram selecionados somente 13 artigos até julho do referido ano, data em que a pesquisa de artigos na *Web* foi feita.
4. A Tabela 6 - Resumo da Quantidade de Artigos Encontrados, Baixados e Selecionados, que está no APÊNDICE A – TABELAS DE RESULTADO, apresenta a quantidade de artigos encontrados, baixados e selecionados.

3.1.6 Avaliação da qualidade dos estudos

Neste estudo não se pretende avaliar a qualidade dos estudos primários quanto a sua classificação de publicação. Não importa para a conclusão desta dissertação se o artigo foi publicado num veículo de comunicação com fator de impacto A1, A2, B1, etc., pois o desejado foi investigar um amplo espectro de publicações e somente examinar os artigos quanto as suas ameaças à validade.

3.1.6.1 Avaliação da qualidade dos estudos primários quanto às suas ameaças à validade

Para avaliar a qualidade da descrição das ameaças à validade dos estudos empíricos contidos nos trabalhos escolhidos, que formarão a fonte primária de dados para este estudo, procurou-se na literatura algo que fosse pertinente a este tema. Foram encontrados alguns conceitos e definições gerais que serão utilizados para a avaliação.

De acordo com Higgins *et al*. (2008) e Khan *et al*. (2001), podemos destacar três tipos de estimativas de avaliação de qualidade para estudos:

a) Abordagens Simples: critérios de validade que utilizam uma escala classificatória (critério totalmente atendido, critério parcialmente atendido, critério não foi atendido).
b) Listas de Verificação: compostas por questões relacionadas à qualidade com respostas sim ou não.

c) Escalas de Qualidade: avaliação numérica de uma série de itens de qualidade, que fornecem uma estimativa quantitativa da qualidade do estudo. Atribui-se pesos aos itens avaliados, dependendo de sua importância. Multiplica-se o valor atribuído ao item por seu peso e somam-se todos os itens utilizados para cada critério.

Observações:

1. O método utilizado para a busca dos quatro tipos de ameaças à validade nos artigos selecionados para investigação foi a leitura do texto destas publicações. Com isto, foi identificado se os artigos citavam ou não suas ameaças à validade: interna, externa, de construção e de conclusão.
2. Foi levada em consideração a qualidade da descrição das ameaças citadas nos textos dos artigos, com a finalidade de classifica-las de acordo com o critério de estimativa de avaliação de qualidade para estudos, que será definido oportunamente.

O método utilizado nas observações 1 e 2 pode ser classificado como um método qualitativo.

3.1.6.2 Definição da maneira como os resultados serão interpretados

Entre os três tipos de estimativas de avaliação de qualidade para estudos citados acima, foi escolhida a abordagem simples:

1. Por ser mais próxima do que se quer avaliar em relação às ameaças à validade de um estudo empírico.
2. Neste trabalho não se quer avaliar a qualidade de um estudo com o intuito de excluí-lo e sim descobrir:
   - Se o estudo não cita suas ameaças à validade (critério não foi atendido);
   - Se o estudo cita suas ameaças de maneira resumida, ou não cita os 4 principais tipos de ameaça, ou ainda não cita suas ameaças de maneira formal (critério parcialmente atendido) ou;
   - Se o estudo cita seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).

3.2 Condução da pesquisa

3.2.1 Planejamento do projeto de pesquisa

3.2.1.1 Escolha do tema;

O tema escolhido para esta pesquisa trata das publicações que retratam testes em aplicações desenvolvidas para ambiente *Web* e possuem um experimento próprio. A partir destes artigos, qual a percentagem destes trabalhos leva em consideração suas ameaças à validade e, diante desta percentagem, até que ponto estes experimentos estão cumprindo os rigores científicos para que estes artigos tenham sua validade assegurada de acordo com a literatura especializada.

3.2.1.2 Identificação do problema

Este problema foi identificado quando lendo artigos sobre testes em aplicações construídas para ambiente *Web*, notou-se que alguns explicitavam suas ameaças à validade e outros não. Por quê?

3.2.1.3 Busca literária

Para encontrar as fontes de pesquisa que basearam a metodologia adotada nesta dissertação, foi executada uma revisão sistemática da literatura que buscou os principais autores nos campos de mapeamento sistemático da literatura, estudos e métodos experimentais em engenharia de software, ameaças à validade de experimentos e testes em softwares.

A população estudada (fontes primárias de pesquisa) foram artigos publicados nos principais veículos de divulgação científica, da comunidade de engenharia de software, que tratavam de testes para aplicações *Web* e possuíssem um ou mais experimentos próprios.

A coleta de dados foi feita através das principais máquinas de busca (bases de dados) existentes na *Web*, por meio de "*strings* de busca" específicas para cada máquina de busca.

3.2.1.4 Planejamento da pesquisa

A partir dos dados coletados, foi feito um mapeamento numérico da quantidade de ameaças à validade que cada artigo referenciava e que foi disponibilizada em uma planilha do seguinte modo:

- Artigos que possuíam experimentos próprios;

- Artigos que não explicitavam nenhuma ameaça à sua validade;
- Artigos que explicitavam ameaças à sua validade de maneira resumida ou não citavam os quatro principais tipos de ameaça, ou ainda não citavam suas ameaças de maneira formal;
- Artigos que explicitavam ameaças à sua validade e citavam seus quatro principais tipos de ameaças de maneira formal e minuciosamente;

Os dados numéricos foram computados e contribuíram para responder as perguntas feitas neste trabalho, assim como foram à base para a conclusão desta dissertação.

4 Resultados do mapeamento

4.1 Respostas às questões de pesquisa

4.1.1 Resposta à questão geral

Até que ponto os experimentos em Engenharia de Software, cujo tema é “Testes em aplicativos desenvolvidos para ambiente *Web*”, estão cumprindo os rigores científicos necessários para serem considerados válidos?

Esta resposta será apresentada na Conclusão desta dissertação.

4.1.2 Respostas às questões específicas

Observação: A Tabela 8 – Resumo dos resultados encontrados, que está no APÊNDICE A – TABELAS DE RESULTADO, apresenta uma visão geral de todos os resultados encontrados.

a) Qual a quantidade de artigos encontrados, baixados e selecionados para esta dissertação?

   De acordo com a Tabela 6 – Resumo da Quantidade de Artigos Encontrados, Baixados e Selecionados, no período de 2010 a 2019, foram encontrados 797 artigos, baixados 336 artigos (42,1%) e selecionados 20 artigos por ano (2010 – 2018) e 13 artigos no ano de 2019, totalizando 193 artigos (24,2%).

Dentro do universo de artigos selecionados:

b) Quantos e qual percentual de artigos referenciam as suas ameaças à validade e quantos não o fazem?

   Esta resposta tem por base a Tabela 2 – Totais e percentagem de artigos que fazem e que não fazem referência as suas ameaças à validade, apresentadas no item APÊNDICE A – TABELAS DE RESULTADO.

   De um total de 193 artigos estudados, 89 (46,1%) citaram, formal e informalmente, suas ameaças à validade e 104 (53,9%) não citaram suas ameaças à validade.

Observação:

1. Existem artigos (total de 32) que embora, em algum lugar do texto, citem as palavras “Ameaças à validade”, na descrição das ameaças eles não referenciam nenhuma das quatro ameaças formais consideradas pela literatura específica. Estes artigos são: A31*, A32*, A35*, A52*, A61*, A65*, A67*, A68*, A75*, A81*, A83*, A86*, A102*, A103*, A106*, A107*, A111*, A116*, A122*, A124*, A127*, A141*, A148*, A150*, A154*, A159*, A162*, A173*, A179*, A181*, A182* e A190*.
2. Os artigos marcados com * encontram-se na Tabela 4 – Ameaças mais citadas, do APÊNDICE A – TABELAS DE RESULTADO.

c) Qual a frequência anual de artigos que faz referência as suas ameaças à validade e quantos não fazem?

Esta resposta tem por base a Tabela 3 – Artigos que fazem e que não fazem referência as suas ameaças à validade – frequência anual, apresentado no item APÊNDICE A – TABELAS DE RESULTADO.

- No ano de 2010, 6 artigos (30%) fizeram referência e 14 artigos (70%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 14 artigos (70%) não citaram suas ameaças à validade (critério não foi atendido), 6 artigos (30%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e nenhum artigo citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).
- No ano de 2011, 11 artigos (55%) fizeram referência e 9 artigos (45%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 9 artigos (45%) não citaram suas ameaças à validade (critério não foi atendido), 10 artigos (50%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e 1 artigo (5%) citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).

- No ano de 2012, 7 artigos (35%) fizeram referência e 13 artigos (65%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 13 artigos (65%) não citaram suas ameaças à validade (critério não foi atendido), 7 artigos (35%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e nenhum artigo citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).
- No ano de 2013, 13 artigos (65%) fizeram referência e 7 artigos (35%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 7 artigos (35%) não citaram suas ameaças à validade (critério não foi atendido), 12 artigos (60%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e 1 artigo (5%) citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).
- No ano de 2014, 10 artigos (50%) fizeram referência e 10 artigos (50%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 10 artigos (50%) não citaram suas ameaças à validade (critério não foi atendido), 9 artigos (45%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e 1 artigo (5%) citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).
- No ano de 2015, 13 artigos (65%) fizeram referência e 7 artigos (35%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 13 artigos (65%) não citaram suas ameaças à validade (critério não foi atendido), 7 artigos (35%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e nenhum artigo citou

seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).

- No ano de 2016, 8 artigos (40%) fizeram referência e 12 artigos (60%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 12 artigos (60%) não citaram suas ameaças à validade (critério não foi atendido), 7 artigos (35%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e 1 artigo (5%) citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).
- No ano de 2017, 10 artigos (50%) fizeram referência e 10 artigos (50%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 10 artigos (50%) não citaram suas ameaças à validade (critério não foi atendido), 9 artigos (45%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e 1 artigo (5%) citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).
- No ano de 2018, 5 artigos (25%) fizeram referência e 15 artigos (75%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 15 artigos (75%) não citaram suas ameaças à validade (critério não foi atendido), 5 artigos (25%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido) e nenhum artigo citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).
- No ano de 2019, 6 artigos (46,2%) fizeram referência e 7 artigos (53,8%) não fizeram referência as suas ameaças à validade. Pelo critério de classificação, 7 artigos (53,8%) não citaram suas ameaças à validade (critério não foi atendido), 6 artigos (46,2%) citaram suas ameaças de maneira resumida ou não citaram os 4 principais tipos de ameaça, ou ainda não citaram suas

ameaças de maneira formal (critério parcialmente atendido) e nenhum artigo citou seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).

Observação:

a. No ano de 2019, foram investigados 13 artigos.

d) Qual a quantidade e percentual de artigos divididos entre os três critérios da abordagem simples (critério não atendido, critério parcialmente atendido e critério totalmente atendido)?

Esta resposta tem por base a Tabela 2 – Totais e percentagem de artigos que fazem e que não fazem referência as suas ameaças à validade, apresentados no item APÊNDICE A – TABELAS DE RESULTADO.

- De um total de 193 artigos estudados, 104 artigos (53,9%) não citaram suas ameaças à validade (critério não foi atendido).
- De um total de 193 artigos estudados, 84 artigos (43,5%) citaram suas ameaças de maneira resumida ou não citam os 4 principais tipos de ameaça, ou ainda não citaram suas ameaças de maneira formal (critério parcialmente atendido).
- De um total de 193 artigos estudados, 5 artigos (2,6%) citaram seus 4 principais tipos de ameaças de maneira formal e minuciosamente (critério totalmente atendido).

e) Para artigos que mencionam suas ameaças a validade, quais são as ameaças mais citadas?

Esta resposta tem por base a Tabela 4 – Ameaças mais citadas e a tabela 5 – Total de ameaças mais citadas, apresentados no item APÊNDICE A – TABELAS DE RESULTADO.

- Por ordem decrescente de quantidade de ameaças citadas, a ameaça externa aparece em primeiro lugar com 55 ocorrências, seguido da ameaça interna com 41 ocorrências. Em terceiro lugar está à ameaça de construção com 32 ocorrências e em último lugar, a ameaça de conclusão com 15 ocorrências.

f) Quantos e qual percentual de artigos (total e por ano de publicação) mencionam suas ameaças à validade de forma sumarizada e de forma detalhada?

De acordo com a Tabela 5 – Total de ameaças mais citadas, apresentados no item APÊNDICE A – TABELAS DE RESULTADO, a resposta a essa questão é: 72 artigos citaram suas ameaças à validade de forma Sumarizada e 17 citaram suas ameaças à validade de forma detalhada.

- No ano de 2010, 6 artigos descreveram suas ameaças à validade de forma sumarizada e nenhum artigo descreveu suas ameaças de forma detalhada.
- No ano de 2011, 7 artigos descreveram suas ameaças à validade de forma sumarizada e 4 artigos descreveram suas ameaças de forma detalhada.
- No ano de 2012, 7 artigos descreveram suas ameaças à validade de forma sumarizada e nenhum artigo descreveu suas ameaças de forma detalhada.
- No ano de 2013, 11 artigos descreveram suas ameaças à validade de forma sumarizada e 2 artigos descreveram suas ameaças de forma detalhada.
- No ano de 2014, 8 artigos descreveram suas ameaças à validade de forma sumarizada e 2 artigos descreveram suas ameaças de forma detalhada.
- No ano de 2015, 10 artigos descreveram suas ameaças à validade de forma sumarizada e 3 artigos descreveram suas ameaças de forma detalhada.
- No ano de 2016, 6 artigos descreveram suas ameaças à validade de forma sumarizada e 2 artigos descreveram suas ameaças de forma detalhada.
- No ano de 2017, 8 artigos descreveram suas ameaças à validade de forma sumarizada e 2 artigos descreveram suas ameaças de forma detalhada.
- No ano de 2018, 4 artigos descreveram suas ameaças à validade de forma sumarizada e 1 artigo descreveu suas ameaças de forma detalhada.
- No ano de 2019, 5 artigos descreveram suas ameaças à validade de forma sumarizada e 1 artigo descreveu suas ameaças de forma detalhada.

Observação:

a. Neste caso os 32 artigos citados no item “a” foram levados em consideração, pois de alguma maneira, mesmo que não formal, descreveram de forma sumarizada algum tipo de ameaça à validade em seus textos.
b. Os totais computados nas colunas de “Descrição (Sumarizada e Detalhada)” correspondem à quantidade de artigos.

g) Qual a quantidade e percentual de artigos que classificam suas ameaças formalmente e Informalmente (total e por ano de publicação)?

   De acordo com a Tabela 5 – Total de ameaças mais citadas, apresentada no item APÊNDICE A – TABELAS DE RESULTADO, 32 artigos (16,6%), que citaram alguma ameaça de maneira não formal, foram encontrados, o que nos deixa um total de 57 artigos (29,5%) que citaram alguma ameaça de maneira formal.

   Observação:

   a) Citar uma ameaça de maneira formal é nomeá-la como: ameaça externa, ameaça interna, ameaça de construção ou ameaça de conclusão. Isto não significa que no artigo em questão, as quatro ameaças formais tenham sido citadas.
   b) Citar uma ameaça de maneira não formal é escrever um texto sobre a ameaça, mas não nomeá-la.

h) Qual a quantidade e percentual de artigos por Tipo de Publicação?

   De acordo com a Tabela 7 – Tipos de Publicação por Ano de Publicação, de um total de 193 artigos, 172 artigos (89,1%) foram classificados como experimento, 19 artigos (9,9%) foram classificados como estudo de caso e 2 artigos (1,0%) foram classificados como *survey*.

i) Qual a quantidade e percentagem de artigos que cumprem integralmente os rigores científicos necessários para serem considerados válidos, de acordo com a literatura especializada?

   De acordo com a Tabela 2 – Totais e percentagem de artigos que fazem e que não fazem referência as suas ameaças à validade, apresentada no item APÊNDICE A – TABELAS DE RESULTADO, de 193 artigos baixados, somente 5 artigos (2,6%) cumprem os rigores científicos necessários para serem considerados válidos. Os artigos são: A27, A72, A91, A125 e A146.

4.2 Ameaças à validade deste estudo

Mesmo este trabalho sendo classificado como um estudo de mapeamento sistemático da literatura, ele sofre ameaças à sua validade.

Para a descrição das ameaças à validade deste estudo, será utilizada a lista feita com base nos experimentos discutidos por (COOK E CAMPBELL, 1979).

4.2.1 Validade de conclusão

Neste trabalho nos concentramos em ameaças à validade de estudos que possuem experimentos próprios em testes de aplicativos para *Web*. Descobrimos que de um total de 193 artigos, 104 (53,9%) estudos não fazem referência a qualquer ameaça à validade. Além disso, nos 89 artigos que referenciam suas ameaças (46,1%), apenas 5 artigos (2,6% do total de artigos pesquisados ou 5,6% dos artigos que fazem referência as suas ameaças à validade) o fazem de maneira criteriosa. A grande disparidade entre os dois conjuntos não deixa espaço para inferências não realistas sobre a validade de conclusão deste estudo.

a) Baixo poder estatístico.

No caso deste estudo, o teste estatístico revela um padrão através dos dados, pois a utilização da estatística foi simplesmente o cálculo percentual feito a partir das quantidades de artigos que mencionavam ou não suas ameaças à validade.

b) Suposições violadas dos testes estatísticos.

Neste estudo não foram feitas suposições. A estatística utilizada foi simples, sendo aplicada sobre quantidades. Então, este item não representa uma ameaça real.

c) Pesca e taxa de erro.

Como este é um estudo de mapeamento, os resultados vieram da quantificação de artigos que mencionam ou não suas ameaças à validade. Existiram critérios para baixar os artigos, já mencionados, mas a escolha dos artigos foi randômica. Este item também não foi considerado uma ameaça real.

d) Confiabilidade de medidas.

Não é uma ameaça. Este estudo foi feito cinco vezes, em épocas distintas, e com quantidade de artigos diferentes. Em todas elas os resultados foram muito parecidos, ficando em torno de 70% dos artigos que não mencionam suas ameaças à validade, exatamente como nesta dissertação – 57 artigos (29,5%) citam suas ameaças à validade e 136 artigos (70,5%) não citam suas ameaças à validade.

e) Confiabilidade da implementação do tratamento.

Não é uma ameaça. No caso deste estudo, ele foi feito por uma única pessoa.

f) Irrelevâncias aleatórias na configuração experimental.

Não é uma ameaça. Este estudo foi feito por uma única pessoa e não houve interrupção da pesquisa.

g) Heterogeneidade aleatória de assuntos.

Este item não foi considerado uma ameaça, pois esta pesquisa foi feita por somente uma pessoa.

4.2.2 Validade interna

É possível que alguns artigos (referências primárias) relevantes não tenham sido escolhidos durante o processo de pesquisa, pois foram escolhidos, por ano de publicação, os vinte primeiros que atendiam aos critérios de inclusão de artigos. Mitigamos essa ameaça, na medida do possível, usando uma *string* de busca abrangente e sem tendências, para obter um universo maior de opções e escolher artigos de tópicos variados.

Outro item relevante a ser considerado para avaliar a ameaça interna desta dissertação é a quantidade de artigos (104 de um total de 193) cujas ameaças à validade dos experimentos não foram mencionadas ou documentadas:

1. Autores que em seus artigos não as mencionaram por erro ou esquecimento;
2. Autores que consideraram óbvia a validade dos seus experimentos, e não exibiram as suas ameaças em seus artigos;
3. Artigos em que mesmo os experimentos descritos, tendo seguido os ritos científicos adequados, sendo eles válidos e sem ameaças, não mencionaram suas ameaças à validade.

Nenhum dos itens citados acima foi considerado uma real ameaça interna a este estudo, pois a literatura especializada afirma que as ameaças à validade de estudos que possuam experimentos devem ser explicitadas nas publicações (Sjøberg, 2007; Wainer, 2007; Travassos, 2002), mesmo se não forem ameaças reais.

Existe também uma ameaça à validade interna deste trabalho no que diz respeito à busca dos quatro tipos de ameaças à validade nos artigos (fontes primárias). Nesta pesquisa não houve um método automático de busca das ameaças como, por exemplo, a utilização de palavras chave para encontrar nos textos os quatro tipos de ameaças e sim foi feita a leitura cuidadosa dos artigos e a identificação ou não, por interpretação, de ameaças.

Esta ameaça interna não foi considerada relevante, pois este critério (qualitativo) reforça a qualidade da busca das ameaças à validade nos textos dos artigos.

Ameaças de grupo único.

Neste caso específico deve-se levar em consideração este tipo de ameaça, pois o estudo foi feito por uma só pessoa.

a) História.

Não é uma ameaça. Não foram aplicados tratamentos diferentes para o objeto de estudo em momentos diferentes.

c) Maturação.

Não é uma ameaça. Este estudo foi feito por uma única pessoa e foi mantido o mesmo comportamento durante todo o decurso do estudo.

d) Teste.

Não é uma ameaça, pois a dinâmica do estudo que foi aplicada era simples. Foram escolhidos artigos por ano estudado. Depois verificado se os artigos possuíam experimentos. Os primeiros vintes artigos (por ano de publicação) que possuíam experimentos eram investigados quanto às ameaças à validade relatadas nos artigos.

e) Instrumentação.

Não é uma ameaça real, pois a inspeção dos artigos, embora trabalhosa, foi simples. Era preciso somente ler o artigo para descobrir se dentro do mesmo havia um experimento. Se sim, investigar as ameaças à validade constantes no artigo.

f) Regressão estatística.

Não é uma ameaça. A cada vez que o estudo era feito, uma nova leva de artigos era baixada para investigação.

g) Seleção.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

h) Mortalidade.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

i) Ambiguidade da direção da influência causal.

Não é uma ameaça. A questão era simples: os artigos que contém experimentos relatam suas ameaças à validade? Neste caso A não causa B, B não causa A ou mesmo X não causa A e B.

j) Ameaças de vários grupos.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

k) Interações com a seleção.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

l) Ameaças sociais à validade interna.

Não é uma ameaça. Este estudo foi feito por uma única pessoa. Um só método foi utilizado.

m) Difusão ou imitação de tratamentos.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

n) Equalização compensatória de tratamentos.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

o) Rivalidade compensatória.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

p) Ressentimento desmoralizado.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

4.2.3 Validade de construção

Um dos detalhes da validade da construção é a ameaça que os experimentadores podem influenciar os resultados de um estudo, consciente e inconscientemente, com base no que esperam da experiência.

Neste estudo, não vemos a ameaça da validade de construção como uma ameaça real, pois os artigos foram escolhidos de maneira abrangente e captados os números encontrados no seu conteúdo. Na verdade, o resultado foi surpreendente.

a) Ameaças de projeto.

Não é uma ameaça. O projeto da experiência é muito simples e reflete a construção estudada.

b) Explicação pré-operacional inadequada de construções.

Não é uma ameaça. As construções foram definidas e a teoria simples. Somente foi necessário saber se o artigo relatava uma experiência e se sim, se as ameaças à validade destes artigos foram mencionadas.

b) Tendência à mono-operação.

Não é uma ameaça. Este estudo foi feito com 193 artigos diferentes.

c) Tendência à monométodo.

A tendência à monométodo poderia ser levada em consideração como uma ameaça à construção da pesquisa, pois foi utilizado um único tipo de medida: a quantificação

numérica de artigos que mencionam ou não suas ameaças à validade, mas neste caso específico, por ser um procedimento muito simples (contagem de artigos), a tendência à monométodo não foi considerada como uma ameaça real.

d) Confundindo construções e níveis de construções.

Não é uma ameaça. Este estudo foi feito por uma única pessoa e o procedimento utilizado foi simples.

e) Interação de diferentes tratamentos.

Não é uma ameaça. Este estudo foi feito por uma única pessoa.

f) Interação de testes e tratamento.

Não é uma ameaça. Não houve interação entre testes e tratamento, pois a necessidade do estudo foi a quantificação de artigos com e sem ameaças relatadas. Para achar as ameaças à validade nos artigos, bastou somente ler os artigos e identificá-las.

g) Generalização restrita através de construções.

Não se aplica a este tipo de estudo.

h) Ameaças sociais à validade de construção.

As ameaças sociais à validade de construção não foram consideradas uma ameaça, pois o comportamento do envolvido não mudou por fazer parte do experimento. O estudo foi simples e quantitativo.

i) Hipótese de adivinhação.

Não é uma ameaça. Em momento nenhum houve tentativa de adivinhar qual o resultado do estudo. Os artigos foram lidos e o resultado obtido quanto às ameaças à sua validade foram colocados em uma planilha e só no final os números computados.

j) Apreensão de avaliação.

Não é uma ameaça. Neste caso, pelo tipo de estudo, a apreensão de avaliação não se aplica.

k) Expectativas do experimentador.

Não é uma ameaça. Não se tinham expectativas quanto aos resultados do estudo, somente se quantificaram os resultados.

4.2.4 Validade externa

Ameaças à validade externa são condições que limitam a capacidade de generalizar os resultados da experiência com a prática industrial. Existem três tipos de interações com o tratamento: pessoas, local e horário.

Com base no que foi apresentado por Wohlin *et al*. (2012) se um experimento é realizado em um escopo limitado e seu design é feito para abordar questões relacionadas ao escopo em questão, seus resultados são válidos nesse contexto. Neste caso, validade adequada não significa validade geral, mas se conclusões gerais forem tiradas do experimento, sua validade terá que abranger um contexto mais amplo.

Concluímos que para o universo dos 193 estudos utilizados, nossos resultados são válidos, mas se considerarmos todo o universo de artigos relativos a testes de aplicações *Web*, a quantidade selecionada de artigos é reduzida em relação ao tamanho do universo. Então, se escolhermos um número maior de estudos, talvez as conclusões tiradas sejam diferentes. Portanto, achamos que existe uma ameaça real à validade externa deste estudo, porque não podemos generalizá-lo para todos os artigos do universo.

a) Interação entre seleção e tratamento.

Não se aplica, pois nesta pesquisa somente uma pessoa participou.

b) Interação entre configuração e tratamento.

Não se aplica, pois neste tipo de pesquisa (quantitativa) a ferramenta utilizada foi somente uma planilha eletrônica de versão atualizada.

c) Interação entre história e tratamento.

Não se aplica, pois a pesquisa não foi feita em nenhum momento específico.

4.2.5 Prioridade entre os tipos de ameaças à validade

Pela natureza deste estudo ser classificada como pesquisa aplicada, a prioridade definida entre os quatro tipos de ameaças à validade foi: interna, externa, construção e conclusão.

Como nesta dissertação o objetivo maior foi estudar as relações entre causa e efeito, a prioridade mais alta foi dada à validade interna.

Seguindo-se à validade interna foi escolhida a validade externa, pois a pesquisa foi generalizada de um contexto menor para um contexto mais amplo.

Como, em pesquisa aplicada, existe maior interesse no próprio efeito e menor interesse no componente do tratamento que causou o efeito, à validade de construção ficou na terceira posição.

Em quarta e última posição foi elencada à validade de conclusão, pois apesar do conjunto de dados encontrado, foi dada menor significância as conclusões estatísticas.

5 Conclusão

Segundo Sjøberg *et al*. (2007), o progresso na Engenharia de Software requer (1) mais estudos empíricos de qualidade, (2) maior foco na síntese de evidências e (3) mais teorias devem ser construídas e testadas. Portanto, precisamos de mais estudos em que possamos confiar, especialmente em relação à sua validade.

Para obter essa confiança, as informações coletadas devem ser mais focadas em observação e experimentação do que em lógica dedutiva ou matemática. Para isso, devemos usar os mesmos métodos empíricos usados pelas ciências que estudam fenômenos do mundo real - ciências empíricas.

Segundo Wainer (2007) e Travassos *et al*. (2002), para um experimento ser considerado válido, deve fornecer um alto nível de confiança no processo de investigação experimental. Esse nível de confiança deve permear todos os elementos envolvidos no processo, desde a base teórica adotada, até os resultados finais. Também é importante que o experimento leve em conta todas as ameaças à sua validade: externa, interna, construção e conclusão.

Para ter uma ampla gama de estudos que nos permitissem analisar com mais precisão as ameaças à validade dos experimentos de Engenharia de Software em testes de aplicativos da *Web*, selecionamos artigos com temas variados que vão desde algoritmos a metodologias para testes de aplicações *Web*.

A análise dessa ampla gama de artigos (193) mostra que 104 artigos (53,9%) não mencionam nenhuma ameaça à validade. Por outro lado, dos 89 artigos restantes (46,1%), apenas 5 artigos (2,6%) o fazem de forma criteriosa.

Do conjunto dos 89 artigos (46,1%) que mencionam ameaças a sua validade 84 artigos (94,4% do total de 84 artigos), não fornecem classificação completa de suas ameaças à validade.

Um total de 72 artigos (37%) descrevem suas ameaças à validade de uma maneira sumária e 17 (9%) o fazem detalhadamente.

Os artigos publicados nos anos de 2013 e 2015 foram os que mais citaram suas ameaças à validade, cada um com 13 artigos, 65% do total de artigos de cada ano.

As ameaças à validade mais citadas são, respectivamente: Externa – 55 artigos citaram este tipo de ameaça; Interna – 41 artigos citaram este tipo de ameaça; Construção – 32 artigos citaram este tipo de ameaça e Conclusão – 15 artigos citaram este tipo de ameaça.

Esta pesquisa também sofre ameaças à sua validade. A primeira, validade de conclusão, não é crítica, pois, como este é um estudo de mapeamento, ele ficou restrito a encontrar a frequência relativa dos artigos que mencionam ameaças de validade (46,1%) contra aqueles que não mencionam ameaças (53,9%).

Considera-se que exista um pequeno risco deste estudo sofrer ameaças de validade interna, pois é possível que alguns artigos (referências primárias) relevantes não tenham sido escolhidos durante o processo de pesquisa. Mitigamos essa ameaça, na medida do possível, usando uma *string* de pesquisa abrangente e sem tendências que retornasse um universo maior de opções com artigos de tópicos variados.

Outra ameaça interna não levada em consideração foi o fato de que podem ter existido autores que não explicitaram as ameaças de seus estudos por erro, ou esquecimento, ou considerarem óbvia a validade dos seus experimentos. Além desta prática ser improvável no meio acadêmico, a literatura especializada afirma que as ameaças à validade de estudos que possuam experimentos próprios devem ser evidenciadas nas publicações (Sjøberg, 2007; Wainer, 2007; Travassos, 2002), mesmo se não forem ameaças reais.

A validade de construção deste estudo de mapeamento sistemático não foi considerada uma ameaça real, porque não haviam ideias preconcebidas sobre as descobertas. O estudo consistiu simplesmente em coletar e contabilizar informações.

Finalmente, pode-se crer que a validade externa seja a maior ameaça deste estudo. Embora, os resultados numéricos dos 193 artigos investigados sejam confiáveis, existe uma chance não negligenciável, caso o universo de artigos investigados fosse maior, de apurarmos números diferentes.

Reiterando nossas observações anteriores, e respondendo a questão geral formulada no item 3.1.1.1, embora este estudo não esteja totalmente imune à ameaças de validade, os números apurados nos levam à conclusão que os trabalhos estudados no período analisado, que tiveram como alvo artigos que possuíam experimentos próprios em testes de aplicações

construídas para ambiente *Web*, carecem de rigor científico dos seus autores, pois apenas cinco artigos, de um total de 193, cumpriram integramente os ritos científicos.

Como trabalhos futuros, se deseja:

- Fazer o levantamento do Qualis dos artigos citados na dissertação, como fontes primárias, para relacionar e classificar os veículos utilizados na divulgação destes e discutir à relevância desta classificação.
- Ampliar esta pesquisa para o universo dos diversos campos da Engenharia de Software e comparar os resultados obtidos com outros campos das ciências biomédicas e sociais.

Referências Bibliográficas

Referências de Estudos Teóricos

BASILI, V. R., SHULL, F., and LANUBILE, F (1999). Building knowledge through families of experiments. *IEEE Transactions of Software Engineering*, **25.4**: 456-474.

BOEHM, BARRY; ROMBACH, HANS DIETER; ZELKOWITZ, MARVIN V. (2005). *Foundations of Empirical Software Engineering: The Legacy of Victor R. Basili*. 83. Germany, Springer-Verlag.

CAMPBELL, D.T., STANLEY, J.C. (1963). *Experimental and Quasi-experimental Designs for Research*. Boston, Houghton Mifflin Company.

CASSIANI, SILVIA HELENA DE BORTOLI. Buscando significado para o trabalho: o aperfeiçoamento profissional sob a perspectiva de enfermeiras. 1994.

COOK, T.D., CAMPBELL, D.T. (1979). *Quasi-experimentation – Design and Analysis Issues for Field Settings*. Boston, Houghton Mifflin Company.

DAVISON, R.; MARTINSONS, M. G.; KOCK, N. (2004). Principles of canonical action research. Information systems Journal, v. 14, n. 1, p. 65–86.

DIAS, DONALDO DE SOUZA. SILVA, MÔNICA FERREIRA DA. Como escrever uma monografia: manual de elaboração com exemplos e exercícios. São Paulo: Atlas, 2010.

DICK, B. Grounded theory: a thumbnail sketch, (2002). Disponível em: http://www.scu.edu.au/schools/gcm/ar/arp/grounded.html, acesso em 09 jun 2005.

EASTERBROOK, S., SINGER, J., STOREY, M. A., & DAMIAN, D. (2008). Selecting empirical methods for software engineering research. Guide to advanced empirical software engineering. [S.l.]: Springer, p. 285–311.

FLYVBJERG, B. (2006). Five misunderstandings about case-study research. Qualitative inquiry, v. 12, n. 2, p. 219–245.

GLASER, B.G. and STRAUSS, A. (1967) Discovery of Grounded Theory: Strategies for Qualitative Research. Sociology Press, Mill Valley, CA

GLASS, Robert L. *et al.* Software testing and industry needs. IEEE Software, v. 23, n. 4, p. 55-57, 2006.

GINIGE, A., MURUGESAN, S. (2001) "*Web* Engineering: an Introduction, IEEE Multimedia", Vol. 8, Issue: 1, p. 14 -18.

HIGGINS, J. P.; GREEN, S.; COLLABORATION, C. (2008). Cochrane handbook for

systematic reviews of interventions. [S.l.]: Wiley Online Library, v. 5.

JURISTO, NATALIA; MORENO, ANA M (2001). *Basics of Software Engineering Experimentation*. Kluwer Academic Publishers.

KHAN, K. S., TER RIET, G., GLANVILLE, J., SOWDEN, A. J., & Kleijnen, J. (2001). Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews. [S.l.]: NHS Centre for Reviews and Dissemination.

KITCHENHAM, BARBARA *et al.* (2010). Systematic literature reviews in software engineering–a tertiary study. *Information and Software Technology* **52.8**: 792-805.

KITCHENHAM, Barbara; CHARTERS, Stuart. Guidelines for performing systematic literature reviews in software engineering. 2007.

KITCHENHAM, B. A., PFLEEGER, S. L., PICKARD, L. M., JONES, P. W., HOAGLIN, D. C., EL EMAM, K., & ROSENBERG, J. (2002). Preliminary guidelines for empirical research in software engineering. IEEE Transactions on software engineering, 28(8), 721-734.

LAU, F. Toward a Framework for Action Research in Information Systems Studies, Information Technology & People, 12(2): 148–175, 1999

MILLER, JAMES (2000). Applying meta-analytical procedures to software engineering experiments. *Journal of Systems and Software*, **54.1**: 29-39.

NASCIMENTO, Arnaldo Gil Sardella. A Mapping Study of Scientific Merit of Papers, which Subject are *Web* Applications Test Techniques, Considering their Validity Threats. JISTEM-Journal of Information Systems and Technology Management, v. 15, 2018.

PERRY, DEWAYNE E., ADAM A. PORTER, and LAWRENCE G. VOTTA (2000). Empirical studies of software engineering: a roadmap. *Proceedings of the conference on The future of Software engineering. ACM.*

PRESSMAN, R., Software Engineering - A Practioner's Approach, Third Edition, McGraw Hill International Edition, 1992.

REASON, James. Human error. Cambridge university press, 1990.

ROBINSON, H.; SEGAL, J.; SHARP, H. (2007). Ethnographically-informed empirical studies of software practice. Information and Software Technology, v. 49, n. 6, p. 540–551.

RONKKO, Kari; LINDEBERG, Olle; DITTRICH, Yvonne. 'Bad practice 'or' Bad methods' are software engineering and ethnographic discourses incompatible?. In: Proceedings International Symposium on Empirical Software Engineering. IEEE, 2002. p. 204-210.

SCALET, D. (1995). Avaliação da Qualidade do Produto de Sotfware, Workshop da

Qualidade e Produtividade em Software e IX SBES/SBC, Recife, Outubro.

SJØBERG, D. I., DYBÅ, T., and JØRGENSEN, M. (2007). The Future of Empirical Methods in Software Engineering Research. *In 2007 Future of Software Engineering (May 23 - 25, 2007). International Conference on Software Engineering. IEEE Computer Society, Washington, DC, 358-378.*

TASSEY, Gregory. The economic impacts of inadequate infrastructure for software testing. National Institute of Standards and Technology, RTI Project, v. 7007, n. 011, p. 429-489, 2002.

TRAVASSOS, G. H., GUROV, D. E AMARAL, E. A. G. (2002): Introdução à Engenharia de Software Experimental. COPPE/UFRJ, Rio de Janeiro, Relatório técnico: RT-ES-590/02, http://cultura.ufpa.br/cdesouza/teaching/methods/6-ES-Experimental.pdf. Accessed 12-Dec-2014.

WAINER, J. (2007): Métodos de pesquisa quantitativa e qualitativa para a Ciência da Computação.
http://www.pucrs.br/famat/viali/mestrado/mqp/material/textos/Pesquisa.pdf. Accessed 01-Nov-2014.

WOHLIN, C., RUNESON, P., HÖST, M., OHLSSON, M. C., REGNELL, B., and WESSLÉN, A. (2012). *Experimentation in software engineering*. Berlin, Springer-Verlag Berlin Heidelberg.

YIN, R. K. (2009). Case study research: Design and methods. [S.l.]: Sage, v. 5.

ZELKOWITZ, MARVIN V., WALLACE, DOLORES R (1998). Experimental models for validating computer technology. *IEEE Computer* **31.5**: 23-31.

Referências de Artigos Primários

2010

A1 DOBOLYI, Kinga; WEIMER, Westley. Addressing high severity faults in *Web* application testing. In: The IASTED International Conference on Software Engineering. 2010.

A2 HUANG, Ying; LU, Lu. A methodology for test suit reduction in user-session-based testing. In: Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on. IEEE, 2010. p. 864-868.

A3 KOSINDRDECHA, Nicha; DAENGDEJ, Jirapun. A test case generation process and technique. J. Software Eng, v. 4, p. 265-287, 2010.

A4 CHEN, Xiang *et al.* Applying particle swarm optimization to pairwise testing. In: Computer Software and Applications Conference (COMPSAC), 2010 IEEE

34th Annual. IEEE, 2010. p. 107-116.

A5 ROBERTS-MORPETH, Paul; ELLMAN, Jeremy. Some security issues for *Web* based frameworks. 2010.

A6 ENGSTRÖM, E., PER RUNESON, and GREGER WIKSTRAND (2010). An empirical evaluation of regression testing based on fix-cache recommendations. Software Testing, Verification and Validation (ICST), Third International Conference.

A7 LI, NUO *et al.* (2010). Perturbation-based user-input-validation testing of *Web* applications. Journal of Systems and Software 83.11: 2263-2274.

A8 OFFUTT, JEFF, and YE WU (2010). Modeling presentation layers of *Web* applications for testing. Software & Systems Modeling 9.2: 257-280.

A9 PRAPHAMONTRIPONG, U., & OFFUTT, J. (2010, April). Applying mutation testing to *Web* applications. In Software Testing, Verification, and Validation Workshops (ICSTW), 2010 Third International Conference on (pp. 132-141). IEEE.

A10 ROEST, D., MESBAH, A., & VAN DEURSEN, A. (2010, April). Regression testing ajax applications: Coping with dynamism. In Software Testing, Verification and Validation (ICST), 2010 Third International Conference on (pp. 127-136). IEEE.

A11 ANDREWS, ANNELIESE A. *et al.* (2010). Scalability issues with using FSMWeb to test *Web* applications. Information and Software Technology 52.1: 52-66.

A12 CHOUDHARY, S. R., VERSEE, H., & ORSO, A. (2010, September). A cross-browser *Web* application testing tool. In Software Maintenance (ICSM), 2010 IEEE International Conference on (pp. 1-6). IEEE.

A13 GUPTA, S., & SHARMA, L. (2010). Performance analysis of internal vs. external security mechanism in *Web* applications. Int. J. Advan. Network Applic, 1(05), 314-317.

A14 HALLÉ, SYLVAIN *et al.* (2010). Eliminating navigation errors in *Web* applications via model checking and runtime enforcement of navigation state machines. Proceedings of the IEEE/ACM international conference on Automated software engineering. ACM.

A15 JURECZKO, M., & MLYNARSKI, M. (2010). Automated acceptance testing tools for *Web* applications using test-driven development. Electrotechnical Review, 86(09), 198-202.

A16 QIAN, Z. (2010). User session-based test case generation and optimization using genetic algorithm. Journal of Software Engineering and Applications, 3(06), 541.

A17 SAXENA, PRATEEK *et al.* (2010). A symbolic execution framework for javascript. Security and Privacy (SP), 2010 IEEE Symposium. IEEE.

A18 ZHANG, Hongyu; SHI, Bei; ZHANG, Lu. Automatic checking of license compliance. In: 2010 IEEE International Conference on Software Maintenance. IEEE, 2010. p. 1-3.

A19 KRISHNAMURTHY, Diwakar; SHAMS, Mahnaz; FAR, Behrouz H. A Model-Based Performance Testing Toolset for *Web* Applications. Engineering Letters, v. 18, n. 2, 2010.

A20 DING, Xiaoning *et al.* Splitter: a proxy-based approach for post-migration testing of *Web* applications. In: Proceedings of the 5th European conference on Computer systems. ACM, 2010. p. 97-110.

2011

A21 ASKARUNISA, A., & RAMARAJ, N. (2011). An algorithm for test data set reduction for *Web* application testing. Neural Network World, 21(1), 27.

A22 CAO, M., CAO, Z., & LI, H. Q. (2011). Support for development and test of *Web* application: A tree-oriented model. Journal of Shanghai University (English Edition), 15(5), 357.

A23 PENG, X., & LU, L. (2011). User-session-based automatic test case generation using GA. International Journal of Physical Sciences, 6(13), 3232-3245.

A24 ZHENG, YUNHUI, TAO BAO, and XIANGYU ZHANG (2011). Statically locating *Web* application bugs caused by asynchronous calls. Proceedings of the 20th international conference on World wide *Web*. ACM.

A25 ALSHAHWAN, N., & HARMAN, M. (2011, November). Automated *Web* application testing using search based software engineering. In Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering (pp. 3-12). IEEE Computer Society.

A26 ENGSTRÖM, EMELIE, PER RUNESON, and ANDREAS LJUNG (2011). Improving Regression Testing Transparency and Efficiency with History-Based Prioritization--An Industrial Case Study. Software Testing, Verification and Validation (ICST), 2011 IEEE Fourth International Conference. IEEE.

A27 MARCHETTO, ALESSANDRO *et al.* (2011). Crawlability metrics for automated *Web* testing. International journal on software tools for technology transfer 13.2: 131-149.

A28 MESBAH, ALI, and MUKUL R. PRASAD (2011). Automated cross-browser compatibility testing. Proceedings of the 33rd International Conference on

Software Engineering. ACM.

A29 DOBOLYI, Kinga; SOECHTING, Elizabeth; WEIMER, Westley. Automating regression testing using *Web*-based application similarities. International journal on software tools for technology transfer, v. 13, n. 2, p. 111-129, 2011.

A30 JIA, Yue; HARMAN, Mark. An analysis and *survey* of the development of mutation testing. IEEE transactions on software engineering, v. 37, n. 5, p. 649-678, 2011.

A31 BRYCE, Renée C. *et al.* Test suite prioritization by cost-based combinatorial interaction coverage. International Journal of System Assurance Engineering and Management, v. 2, n. 2, p. 126-134, 2011.

A32 SPRENKLE, Sara; POLLOCK, Lori; SIMKO, Lucy. A study of usage-based navigation models and generated abstract test cases for *Web* applications. In: Software Testing, Verification and Validation (ICST), 2011 IEEE Fourth International Conference on. IEEE, 2011. p. 230-239.

A33 BOCHMANN, Gregor v; JOURDAN, Guy-Vincent; WAN, Bo. Improved usage model for *Web* application reliability testing. In: Testing Software and Systems. Springer, Berlin, Heidelberg, 2011. p. 15-31.

A34 HALFOND, William GJ; CHOUDHARY, Shauvik Roy; ORSO, Alessandro. Improving penetration testing through static and dynamic analysis. Software Testing, Verification and Reliability, v. 21, n. 3, p. 195-214, 2011.

A35 SEGURA, Sergio *et al.* Mutation testing on an object-oriented framework: An experience report. Information and Software Technology, v. 53, n. 10, p. 1124-1136, 2011.

A36 BARTOLINI, Cesare *et al.* Bringing white-box testing to service oriented architectures through a service oriented approach. Journal of Systems and Software, v. 84, n. 4, p. 655-668, 2011.

A37 BRYCE, Renee C.; SAMPATH, Sreedevi; MEMON, Atif M. Developing a single model and test prioritization strategies for event-driven software. IEEE Transactions on Software Engineering, v. 37, n. 1, p. 48-64, 2011.

A38 DESSIATNIKOFF, Anthony *et al.* A clustering approach for *Web* vulnerabilities detection. In: 17th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2011). IEEE Computer Society, 2011. p. 194-203.

A39 KHOURY, Nidal *et al.* An analysis of black-box *Web* application security scanners against stored SQL injection. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011. p. 1095-1101.

A40 SUN, Chang-ai *et al.* Metamorphic testing for *Web* services: Framework and a case study. In: *Web* Services (ICWS), 2011 IEEE International Conference on. IEEE, 2011. p. 283-290.

2012

A41 KUMAR, M. J. P., & YOGI, M. K. (2012). A *survey* on models and test strategies for event-driven software. Int. J. Comput. Eng. Res, 2, 1087-1091.

A42 DALLMEIER, VALENTIN *et al.* (2012). Automatically generating test cases for specification mining. Software Engineering, IEEE Transactions 38.2: 243-257.

A43 CHOUDHARY, SHAUVIK ROY, MUKUL R. PRASAD, and ALESSANDRO ORSO (2012). Crosscheck: Combining crawling and differencing to better detect cross-browser incompatibilities in *Web* applications. Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference. IEEE.

A44 ISLAM, Sadeka *et al.* Empirical prediction models for adaptive resource provisioning in the cloud. Future Generation Computer Systems, v. 28, n. 1, p. 155-162, 2012.

A45 BOLIS, F., GARGANTINI, A., GUARNIERI, M., & MAGRI, E. (2012, September). Evolutionary testing of PHP *Web* applications with WETT. In International Symposium on Search Based Software Engineering (pp. 285-291). Springer Berlin Heidelberg.

A46 SAMPATH, Sreedevi; BRYCE, Renée C. Improving the effectiveness of test suite reduction for user-session-based testing of *Web* applications. Information and Software Technology, v. 54, n. 7, p. 724-738, 2012.

A47 MESBAH, ALI, ARIE VAN DEURSEN, and DANNY ROEST (2012). Invariant-based automatic testing of modern *Web* applications. Software Engineering, IEEE Transactions 38.1: 35-53.

A48 MIRSHOKRAIE, SHABNAM, and ALI MESBAH (2012). JSART: JavaScript assertion-based regression testing. *Web* Engineering. 238-252. Springer Berlin Heidelberg.

A49 AL-AHMAD, A., ATA, B. A., & WAHBEH, A. (2012). Pen Testing for *Web* Applications. International Journal of Information Technology and *Web* Engineering (IJITWE), 7(3), 1-13.

A50 WANG, W., & LEI, Y. (2012). Zoomer: An Automated *Web* Application Change Localization Tool. Journal of Communication and Computer, 9(8), 913-919.

A51 BAI, Xiaoying; KENETT, Ron S.; YU, Wei. Risk assessment and adaptive group testing of semantic *Web* services. International Journal of Software Engineering and Knowledge Engineering, v. 22, n. 05, p. 595-620, 2012.

A52 MESBAH, Ali; VAN DEURSEN, Arie; LENSELINK, Stefan. Crawling Ajax-based *Web* applications through dynamic analysis of user interface state changes. ACM Transactions on the *Web* (TWEB), v. 6, n. 1, p. 3, 2012.

A53 AWAD, Mamoun A.; KHALIL, Issa. Prediction of user's *Web*-browsing behavior: Application of markov model. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), v. 42, n. 4, p. 1131-1142, 2012.

A54 JAYASINGHE, Deepal *et al.* Expertus: A generator approach to automate performance testing in IaaS clouds. In: Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on. IEEE, 2012. p. 115-122.

A55 ZHU, Hong; ZHANG, Yufeng. Collaborative testing of *Web* services. IEEE Transactions on Services Computing, v. 5, n. 1, p. 116-130, 2012.

A56 ALSHAHWAN, Nadia; HARMAN, Mark. State aware test case regeneration for improving *Web* application test suite coverage and fault detection. In: Proceedings of the 2012 International Symposium on Software Testing and Analysis. ACM, 2012. p. 45-55.

A57 MARIANI, Leonardo *et al.* Autoblacktest: Automatic black-box testing of interactive applications. In: Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on. IEEE, 2012. p. 81-90.

A58 ABBORS, Fredrik *et al.* MBPeT: a model-based performance testing tool. In: 2012 Fourth International Conference on Advances in System Testing and Validation Lifecycle. 2012.

A59 CHYNAŁ, Piotr; SZYMAŃSKI, Jerzy M.; SOBECKI, Janusz. Using eyetracking in a mobile applications usability testing. In: Asian Conference on Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, 2012. p. 178-186.

A60 STEPIEN, Bernard; PEYTON, Liam; XIONG, Pulei. Using TTCN-3 as a modeling language for *Web* penetration testing. In: Industrial Technology (ICIT), 2012 IEEE International Conference on. IEEE, 2012. p. 674-681.

2013

A61 DENG, Lin; OFFUTT, Jeff; LI, Nan. Empirical evaluation of the statement deletion mutation operator. In: Software Testing, Verification and Validation (ICST), 2013 IEEE Sixth International Conference on. IEEE, 2013. p. 84-93.

A62 AHMAD, Mian Asbat; ORIOL, Manuel. Automated discovery of failure domain. Lecture Notes on Software Engineering, v. 1, n. 3, p. 289, 2013.

A63 HUANG, S. K., LU, H. L., LEONG, W. M., & LIU, H. (2013, June). CraxWeb: Automatic *Web* application testing and attack generation. In Software Security and Reliability (SERE), 2013 IEEE 7th International Conference on (pp. 208-217). IEEE.

A64 SAKAMOTO, KAZUNORI *et al.* (2013). POGen: a test code generator based on template variable coverage in gray-box integration testing for *Web* applications. Fundamental Approaches to Software Engineering. 343-358. Berlin, Springer Berlin Heidelberg.

A65 SINGH, Kochaar Pavneet *et al.* Adoption of Software Testing in Open Source Projects A Preliminary Study on 50,000 Projects. In: 17th European Conference on Software Maintenance and Reengineering (CSMR 2013). 2013. p. 1-4.

A66 ALEMERIEN, Khalid Ali. Evaluation of Software Testing Coverage Tools: An Empirical Study. 2013.

A67 AFSHAN, Sheeva; MCMINN, Phil; STEVENSON, Mark. Evolving readable *string* test inputs using a natural language model to reduce human oracle cost. In: Software Testing, Verification and Validation (ICST), 2013 IEEE Sixth International Conference on. IEEE, 2013. p. 352-361.

A68 PASTORE, Fabrizio; MARIANI, Leonardo; FRASER, Gordon. Crowdoracles: Can the crowd solve the oracle problem. In: International Conference on Software Testing, Verification and Validation (ICST). 2013. p. 342-351.

A69 GLIGORIC, Milos *et al.* Comparing non-adequate test suites using coverage criteria. In: Proceedings of the 2013 International Symposium on Software Testing and Analysis. ACM, 2013. p. 302-313.

A70 QI, Yuhua *et al.* Using automated program repair for evaluating the effectiveness of fault localization techniques. In: Proceedings of the 2013 International Symposium on Software Testing and Analysis. ACM, 2013. p. 191-201.

A71 LAKHOTIA, Kiran; HARMAN, Mark; GROSS, Hamilton. AUSTIN: An open source tool for search based software testing of C programs. Information and Software Technology, v. 55, n. 1, p. 112-125, 2013.

A72 MÄNTYLÄ, Mika V.; ITKONEN, Juha. More testers–The effect of crowd size and time restriction in software testing. Information and Software Technology, v. 55, n. 6, p. 986-1003, 2013.

A73 CHEN, Tsong Yueh *et al.* Code coverage of adaptive random testing. IEEE

Transactions on Reliability, v. 62, n. 1, p. 226-237, 2013.

A74 MALIK, Haroon; HEMMATI, Hadi; HASSAN, Ahmed E. Automatic detection of performance deviations in the load testing of large scale systems. In: Proceedings of the 2013 International Conference on Software Engineering. IEEE Press, 2013. p. 1012-1021.

A75 JENSEN, Casper S.; PRASAD, Mukul R.; MØLLER, Anders. Automated testing with targeted event sequence generation. In: Proceedings of the 2013 International Symposium on Software Testing and Analysis. ACM, 2013. p. 67-77.

A76 SEMENENKO, Nataliia; DUMAS, Marlon; SAAR, Tõnis. Browserbite: Accurate cross-browser testing via machine learning over image features. In: Software Maintenance (ICSM), 2013 29th IEEE International Conference on. IEEE, 2013. p. 528-531.

A77 CONEJO, Ricardo *et al.* A *Web* based collaborative testing environment. Computers & Education, v. 68, p. 440-457, 2013.

A78 SHEA, Ryan; LIU, Jiangchuan. Performance of virtual machines under networked denial of service attacks: Experiments and analysis. IEEE systems journal, v. 7, n. 2, p. 335-345, 2013.

A79 NI, Tongguang *et al.* Real-time detection of application-layer DDoS attack using time series analysis. Journal of Control Science and Engineering, v. 2013, p. 4, 2013.

A80 SAMPATH, Sreedevi; BRYCE, Renee; MEMON, Atif M. A uniform representation of hybrid criteria for regression testing. IEEE transactions on software engineering, v. 39, n. 10, p. 1326-1344, 2013.

2014

A81 CHAUDHARY, NEHA, O. P. SANGWAN, and RICHA ARORA (2014). Event-Coverage and Weight based Method for Test Suite Prioritization. International Journal of Information Technology and Computer Science (IJITCS) 6.12: 61.

A82 EL YOUMI, M., & FALAH, B. (2014, April). Testing *Web* applications by unifying Fuzzy and All-Pairs techniques. In Multimedia Computing and Systems (ICMCS), 2014 International Conference on (pp. 547-551). IEEE.

A83 MIRZAAGHAEI, MEHDI, AND ALI MESBAH (2014). DOM-based test adequacy criteria for *Web* applications. Proceedings of the 2014 International Symposium on Software Testing and Analysis. ACM.

A84 ROY CHOUDHARY, S., PRASAD, M. R., & ORSO, A. (2014, July). X-PERT: a *Web* application testing tool for cross-browser inconsistency detection. In Proceedings of the 2014 International Symposium on Software Testing and Analysis (pp. 417-420). ACM.

A85 MUKHERJEE, J., WANG, M., & KRISHNAMURTHY, D. (2014, March). Performance testing *Web* applications on the cloud. In Software Testing, Verification and Validation Workshops (ICSTW), 2014 IEEE Seventh International Conference on (pp. 363-369). IEEE.

A86 TAPPENDEN, ANDREW F., and JAMES MILLER (2014). Automated cookie collection testing. ACM Transactions on Software Engineering and Methodology (TOSEM) 23.1: 3.

A87 SELAY, Elmin; ZHOU, Zhi Quan; ZOU, Jingjie. Adaptive random testing for image comparison in regression *Web* testing. In: Digital lmage Computing: Techniques and Applications (DlCTA), 2014 International Conference on. IEEE, 2014. p. 1-7.

A88 TUNG, Yuan-Hsin; LIN, Chen-Chiu; SHAN, Hwai-Ling. Test as a Service: A framework for *Web* security TaaS service in cloud environment. In: 2014 IEEE 8th International Symposium on Service Oriented System Engineering (SOSE). IEEE, 2014. p. 212-217.

A89 NGUYEN, Hung Viet; KÄSTNER, Christian; NGUYEN, Tien N. Exploring variability-aware execution for testing plugin-based *Web* applications. In: Proceedings of the 36th International Conference on Software Engineering. ACM, 2014. p. 907-918.

A90 YAN, Minzhi; SUN, Hailong; LIU, Xudong. iTest: testing software with mobile crowdsourcing. In: Proceedings of the 1st International Workshop on Crowd-based Software Development Methods and Technologies. ACM, 2014. p. 19-24.

A91 FELDERER, Michael; BEER, Armin; PEISCHL, Bernhard. On the role of defect taxonomy types for testing requirements: Results of a controlled experiment. In: Software Engineering and Advanced Applications (SEAA), 2014 40th EUROMICRO Conference on. IEEE, 2014. p. 377-384.

A92 APPELT, Dennis *et al.* Automated testing for SQL injection vulnerabilities: an input mutation approach. In: Proceedings of the 2014 International Symposium on Software Testing and Analysis. ACM, 2014. p. 259-269.

A93 MILANI FARD, AMIN, MEHDI MIRZAAGHAEI, and ALI MESBAH (2014). Leveraging existing tests in automated test generation for *Web* applications. Proceedings of the 29th ACM/IEEE international conference on Automated software engineering. ACM.

A94 LEOTTA, Maurizio *et al.* Visual vs. DOM-based *Web* locators: An empirical study. In: International Conference on *Web* Engineering. Springer, Cham, 2014. p. 322-340.

A95 TAMBURRELLI, Giordano; MARGARA, Alessandro. Towards automated A/B testing. In: International Symposium on Search Based Software Engineering. Springer, Cham, 2014. p. 184-198.

A96 HOSSAIN, Md; DO, Hyunsook; EDA, Ravi. Regression testing for *Web* applications using reusable constraint values. In: Software Testing, Verification and Validation Workshops (ICSTW), 2014 IEEE Seventh International Conference on. IEEE, 2014. p. 312-321.

A97 FONSECA, Jose; VIEIRA, Marco; MADEIRA, Henrique. Evaluation of *Web* security mechanisms using vulnerability and attack injection. IEEE Transactions on Dependable and Secure Computing, n. 1, p. 1, 2014.

A98 LI, Nan; OFFUTT, Jeff. An empirical analysis of test oracle strategies for model-based testing. In: Software Testing, Verification and Validation (ICST), 2014 IEEE Seventh International Conference on. IEEE, 2014. p. 363-372.

A99 KHARI, Manju; SINGH, Neha. *Web* Services Vulnerability Testing Using Open source Security Scanners: An experimental Study. International Journal of Advanced Engineering and Global Technology (IJAEGT), p. 790-799, 2014.

A100 MACHRA, Sumit; KHATRI, Narendra. Model Based Testing of Website. Int. J. Comput. Sci. Appl, v. 4, n. 1, p. 143-152, 2014.

2015

A101 LEE, T. J., TSENG, S. S., & KUNG, C. C. (2015). Factors Influencing the Performance of Human Computation: An Empirical Study in *Web* Application Testing. J. Inf. Sci. Eng., 31(2), 641-658.

A102 MAHAJAN, SONAL, and WILLIAM GJ HALFOND (2015). Detection and localization of html presentation failures using computer vision-based techniques. Software Testing, Verification and Validation (ICST), 2015 IEEE 8th International Conference. IEEE.

A103 BANITAAN, Shadi *et al.* Test Case Selection using Software Complexity and Volume Metrics. In: 24th International Conference on Software Engineering and Data Engineering, SEDE'15.

A104 PAPADAKIS, Mike; LE TRAON, Yves. Metallaxis-FL: mutation-based fault localization. Software Testing, Verification and Reliability, v. 25, n. 5-7, p. 605-628, 2015.

A105 KOCHHAR, Pavneet Singh *et al.* Understanding the test automation culture of app developers. 2015.

A106 LEOTTA, Maurizio *et al.* Automated generation of visual *Web* tests from DOM-based *Web* tests. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing. ACM, 2015. p. 775-782.

A107 GAO, Zebao *et al.* Making system user interactive tests repeatable: When and what should we control?. In: Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on. IEEE, 2015. p. 55-65.

A108 FARD, Amin Milani; MESBAH, Ali; WOHLSTADTER, Eric. Generating Fixtures for JavaScript Unit Testing (T). In: Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on. IEEE, 2015. p. 190-200.

A109 TANG, Xinye *et al.* Attention, Test Code is Low-quality!. 2015.

A110 MEI, Lijun *et al.* Preemptive regression testing of workflow-based *Web* services. IEEE Transactions on Services Computing, v. 8, n. 5, p. 740-754, 2015.

A111 SHEN, Du *et al.* Automating performance bottleneck detection using search-based application profiling. In: Proceedings of the 2015 International Symposium on Software Testing and Analysis. ACM, 2015. p. 270-281.

A112 VILLANES, Isabel Karina; COSTA, Erick Alexandre Bezerra; DIAS-NETO, Arilo Claudio. Automated mobile testing as a service (AM-TaaS). In: 2015 IEEE World Congress on Services (SERVICES). IEEE, 2015. p. 79-86.

A113 CHOUDHARY, Shauvik Roy; GORLA, Alessandra; ORSO, Alessandro. Automated test input generation for Android: Are we there yet?. arXiv preprint arXiv:1503.07217, 2015.

A114 SHAMSHIRI, Sina *et al.* Random or genetic algorithm search for object-oriented test suite generation?. In: Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation. ACM, 2015. p. 1367-1374.

A115 LEOTTA, Maurizio *et al.* Using multi-locators to increase the robustness of *Web* test cases. In: Software Testing, Verification and Validation (ICST), 2015 IEEE 8th International Conference on. IEEE, 2015. p. 1-10.

A116 WALSH, Thomas A.; MCMINN, Phil; KAPFHAMMER, Gregory M. Automatic detection of potential layout faults following changes to responsive *Web* pages (N). In: Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on. IEEE, 2015. p. 709-714.

A117 TORRANO GIMENEZ, Carmen *et al.* Combining expert knowledge with automatic feature extraction for reliable *Web* attack detection. Security and

Communication Networks, v. 8, n. 16, p. 2750-2767, 2015.

A118 HAJJAT, Mohammad *et al.* Application-specific configuration selection in the cloud: impact of provider policy and potential of systematic testing. In: Computer Communications (INFOCOM), 2015 IEEE Conference on. IEEE, 2015. p. 873-881.

A119 NÚÑEZ, Alberto; HIERONS, Robert M. A methodology for validating cloud models using metamorphic testing. annals of telecommunications-annales des télécommunications, v. 70, n. 3-4, p. 127-135, 2015.

A120 WU, Chi-Yun *et al.* Automated testing of *Web* applications with text input. In: Progress in Informatics and Computing (PIC), 2015 IEEE International Conference on. IEEE, 2015. p. 343-347.

2016

A121 MAO, Ke; HARMAN, Mark; JIA, Yue. Sapienz: multi-objective automated testing for Android applications. In: Proceedings of the 25th International Symposium on Software Testing and Analysis. ACM, 2016. p. 94-105.

A122 HIRZEL, Matthias; KLAEREN, Herbert. Graph-Walk-based Selective Regression Testing of *Web* Applications Created with Google *Web* Toolkit. In: Software Engineering (Workshops). 2016. p. 55-69.

A123 HÄSER, Florian; FELDERER, Michael; BREU, Ruth. An integrated tool environment for experimentation in domain specific language engineering. In: Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering. ACM, 2016. p. 20.

A124 SUNG, Chungha *et al.* Static DOM event dependency analysis for testing *Web* applications. In: Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering. ACM, 2016. p. 447-459.

A125 PANICHELLA, S., PANICHELLA, A., BELLER, M., ZAIDMAN, A., & GALL, H. C. (2016, May). The impact of test case summaries on bug fixing performance: An empirical investigation. In Proceedings of the 38th International Conference on Software Engineering (pp. 547-558). ACM.

A126 JOSHI, Chanchala; SINGH, Umesh Kumar. Performance evaluation of *Web* application security scanners for more effective defense. International Journal of Scientific and Research Publications (IJSRP), v. 6, n. 6, p. 660-667, 2016.

A127 LEOTTA, Maurizio *et al.* ROBULA+: An algorithm for generating robust XPath locators for *Web* testing. Journal of Software: Evolution and Process, v. 28, n. 3, p. 177-204, 2016.

A128 ESPARCIA-ALCÁZAR, Anna I. *et al.* Q-learning strategies for action selection in the TESTAR automated testing tool. In: Proceedings of META 2016 6th international conference on metaheuristics and nature inspired computing. 2016. p. 174-180.

A129 JOSHI, Chanchala; SINGH, Umesh Kumar. Security Testing and Assessment of Vulnerability Scanners in Quest of Current Information Security Landscape. International Journal of Computer Applications, v. 145, n. 2, p. 1-7, 2016.

A130 VÖGELE, Christian *et al.* WESSBAS: extraction of probabilistic workload specifications for load testing and performance prediction—a model-driven approach for session-based application systems. Software & Systems Modeling, p. 1-35, 2016.

A131 SCHERMANN, Gerald *et al.* Bifrost: supporting continuous deployment with automated enactment of multi-phase live testing strategies. In: Proceedings of the 17th International Middleware Conference. ACM, 2016. p. 12.

A132 ALMENAR, Francisco *et al.* Automated testing of *Web* applications with TESTAR. In: International Symposium on Search Based Software Engineering. Springer, Cham, 2016. p. 218-223.

A133 MARCULESCU, Bogdan *et al.* Tester interactivity makes a difference in search-based software testing: A controlled experiment. Information and Software Technology, v. 78, p. 66-82, 2016.

A134 SAAR, Tõnis *et al.* Browserbite: cross-browser testing via image processing. Software: Practice and Experience, v. 46, n. 11, p. 1459-1477, 2016.

A135 CASTILLA, Diana *et al.* Effect of *Web* navigation style in elderly users. Computers in Human Behavior, v. 55, p. 909-920, 2016.

A136 PIECZUL, Olgierd; FOLEY, Simon N. Runtime detection of zero-day vulnerability exploits in contemporary software systems. In: IFIP Annual Conference on Data and Applications Security and Privacy. Springer, Cham, 2016. p. 347-363.

A137 GARGANTINI, Angelo *et al.* Validation of constraints among configuration parameters using search-based combinatorial interaction testing. In: International Symposium on Search Based Software Engineering. Springer, Cham, 2016. p. 49-63.

A138 HE, Meimei *et al.* X-check: a novel cross-browser testing service based on record/replay. In: 2016 IEEE International Conference on *Web* Services (ICWS). IEEE, 2016. p. 123-130.

A139 MIRZAEI, Nariman *et al.* Reducing combinatorics in GUI testing of android applications. In: Software Engineering (ICSE), 2016 IEEE/ACM 38th

International Conference on. IEEE, 2016. p. 559-570.

A140 ERMUTH, Markus; PRADEL, Michael. Monkey see, monkey do: effective generation of GUI tests with inferred macro events. In: Proceedings of the 25th International Symposium on Software Testing and Analysis. ACM, 2016. p. 82-93.

2017

A141 LIN, Jun-Wei; WANG, Farn; CHU, Paul. Using semantic similarity in crawling-based *Web* application testing. In: Software Testing, Verification and Validation (ICST), 2017 IEEE International Conference on. IEEE, 2017. p. 138-148.

A142 APTE, Varsha *et al.* AutoPerf: Automated load testing and resource usage profiling of multi-tier internet applications. In: Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering. ACM, 2017. p. 115-126.

A143 MARASHDIH, Abdalla Wasef; ZAABA, Zarul Fitri; OMER, Herman Khalid. *Web* Security: Detection of Cross Site Scripting in PHP *Web* Application using Genetic Algorithm. International Journal of Advanced Computer Science and Applications (ijacsa), v. 8, n. 5, 2017.

A144 PETSIOS, Theofilos *et al.* Nezha: Efficient domain-independent differential testing. In: Security and Privacy (SP), 2017 IEEE Symposium on. IEEE, 2017. p. 615-632.

A145 APPELT, Dennis; PANICHELLA, Annibale; BRIAND, Lionel. Automatically repairing *Web* application firewalls based on successful SQL injection attacks. In: Software Reliability Engineering (ISSRE), 2017 IEEE 28th International Symposium on. IEEE, 2017. p. 339-350.

A146 MALAVOLTA, Ivano *et al.* Assessing the impact of service workers on the energy efficiency of progressive *Web* apps. In: Proceedings of the 4th International Conference on Mobile Software Engineering and Systems. IEEE Press, 2017. p. 35-45.

A147 PORTILLO DOMINGUEZ, A. Omar *et al.* PHOEBE: an automation framework for the effective usage of diagnosis tools in the performance testing of clustered systems. Software: Practice and Experience, v. 47, n. 11, p. 1837-1874, 2017.

A148 DENG, Lin *et al.* Mutation operators for testing Android apps. Information and Software Technology, v. 81, p. 154-168, 2017.

A149 ROSENFELD, Ariel; KARDASHOV, Odaya; ZANG, Orel. ACAT: A Novel Machine-Learning-Based Tool For Automating Android Application Testing. In: Haifa

Verification Conference. Springer, Cham, 2017. p. 213-216.

A150 ZECH, Philipp; FELDERER, Michael; BREU, Ruth. Knowledge-based security testing of *Web* applications by logic programming. International Journal on Software Tools for Technology Transfer, p. 1-26, 2017.

A151 GRIGERA, Julián *et al.* Automatic detection of usability smells in *Web* applications. International Journal of Human-Computer Studies, v. 97, p. 129-148, 2017.

A152 LUO, Qi *et al.* Forepost: Finding performance problems automatically with feedback-directed learning software testing. Empirical Software Engineering, v. 22, n. 1, p. 6-56, 2017.

A153 ZHOU, Yuchen; EVANS, David. SSOScan: Automated Testing of *Web* Applications for Single Sign-On Vulnerabilities. In: USENIX Security Symposium. 2014. p. 495-510.

A154 HALFOND, William GJ; ORSO, Alessandro. Improving test case generation for *Web* applications using automated interface discovery. In: Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering. ACM, 2007. p. 145-154.

A155 CAI, Yuhong; GRUNDY, John; HOSKING, John. Synthesizing client load models for performance engineering via *Web* crawling. In: Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering. ACM, 2007. p. 353-362.

A156 ANTUNES, Nuno *et al.* Effective detection of SQL/XPath injection vulnerabilities in *Web* services. In: Services Computing, 2009. SCC'09. IEEE International Conference on. IEEE, 2009. p. 260-267.

A157 MARBACK, Aaron; DO, Hyunsook; EHRESMANN, Nathan. An effective regression testing approach for php *Web* applications. In: Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on. IEEE, 2012. p. 221-230.

A158 ZHANG, Kehuan *et al.* Sidebuster: automated detection and quantification of side-channel leaks in *Web* application development. In: Proceedings of the 17th ACM conference on Computer and communications security. ACM, 2010. p. 595-606.

A159 AHMED, Tarek M. *et al.* Studying the effectiveness of application performance management (APM) tools for detecting performance regressions for *Web* applications: an experience report. In: Proceedings of the 13th International Conference on Mining Software Repositories. ACM, 2016. p. 1-12.

A160 BALDUZZI, Marco *et al.* A solution for the automated detection of clickjacking attacks. In: Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security. ACM, 2010. p. 135-144.

2018

A161 AHMAD, Tanwir; TRUSCAN, Dragos; PORRES, Ivan. Identifying worst-case user scenarios for performance testing of *Web* applications using Markov-chain workload models. Future Generation Computer Systems, 2018.

A162 DADEAU, Frederic *et al.* Contract-based testing for PHP with Praspel. Journal of Systems and Software, v. 136, p. 209-222, 2018.

A163 LI, Xiaosong *et al.* A Quantitative Approach in Heuristic Evaluation of E-commerce Websites. arXiv preprint arXiv:1801.04829, 2018.

A164 TIWARI, Saurabh *et al.* A case study on the application of case-based learning in software testing. In: Proceedings of the 11th Innovations in Software Engineering Conference. ACM, 2018. p. 11.

A165 GUO, Junxia *et al.* A Test Case Generation Method Based on State Importance of EFSM for *Web* Application.

A166 RJEIB, Hasanein d.; Al-sadawi, Haider alsharqi2 Basheer. Multi Mechanism Approach for Preventing sql Injection Attacks in Stored Procedures.

A167 AZARNOVA, T. V. *et al.* Development of dynamic Bayesian models for *Web* application test management. In: Journal of Physics: Conference Series. IOP Publishing, 2018. p. 012024.

A168 SELAY, Elmin *et al.* Adaptive Random Testing in Detecting Layout Faults of *Web* Applications.

A169 HANNA, Milad; ABOUTABL, Amal Elsayed; MOSTAFA, Mostafa-Sami M. Automated Software Testing Framework for *Web* Applications. International Journal of Applied Engineering Research, v. 13, n. 11, p. 9758-9767, 2018.

A170 KHANNA, Munish *et al.* A Novel Approach for Regression Testing of *Web* Applications. International Journal of Intelligent Systems and Applications, v. 10, n. 2, p. 55, 2018.

A171 KIRAÇ, M. Furkan; AKTEMUR, Barış; SÖZER, Hasan. VISOR: A fast image processing pipeline with scaling and translation invariance for test oracle automation of visual output systems. Journal of Systems and Software, v. 136, p. 266-277, 2018.

A172 BENITTI, Fabiane Barreto Vavassori. A Methodology to Define Learning Objects Granularity: A Case Study in Software Testing. Informatics in Education, v. 17, n. 1, 2018.

A173 AKBULUT, Akhan. VinJect: Toolkit for Penetration Testing and Vulnerability Scanning.

A174 LIU, Di *et al.* Generating descriptions for screenshots to assist crowdsourced testing. In: 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER). IEEE, 2018. p. 492-496.

A175 GUPTA, Shashank; GUPTA, B. B. RAJIVE: restricting the abuse of JavaScript injection vulnerabilities on cloud data centre by sensing the violation in expected workflow of *Web* applications. International Journal of Innovative Computing and Applications, v. 9, n. 1, p. 13-36, 2018.

A176 HARATY, Ramzi A.; MANSOUR, Nashat; ZEITUNLIAN, Hratch. Metaheuristic Algorithm for State-Based Software Testing. Applied Artificial Intelligence, v. 32, n. 2, p. 197-213, 2018.

A177 MUKHERJEE, Joydeep; KRISHNAMURTHY, Diwakar. Subscriber-Driven Cloud Interference Mitigation for Network Services.

A178 BLÄSER, Luc. Practical detection of concurrency issues at coding time. In: Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis. ACM, 2018. p. 221-231.

A179 BURES, Miroslav; FRAJTAK, Karel; AHMED, Bestoun S. Tapir: Automation Support of Exploratory Testing Using Model Reconstruction of the System Under Test. IEEE Transactions on Reliability, 2018.

A180 SAMAD, Hafiza *et al.* Performance Evaluation of *Web* Application Server based on Request Bit per Second and Transfer Rate Parameters. In: Journal of Physics: Conference Series. IOP Publishing, 2018. p. 012007.

2019

A181 NGUYEN, Hung Viet *et al.* Exploring output-based coverage for testing PHP *Web* applications. Automated Software Engineering, v. 26, n. 1, p. 59-85, 2019.

A182 NDIAYE, Youssou *et al.* Requirements for preventing logic flaws in the authentication procedure of *Web* applications. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. ACM, 2019. p. 1620-1628.

A183 DANG, Tran Tri; DANG, Tran Khanh. A visual model for *Web* applications

security monitoring. arXiv preprint arXiv:1904.03320, 2019.

A184 DUPPA, Gilang Intan Permatasari; SURANTHA, Nico. Evaluation of network security based on next generation intrusion prevention system. TELKOMNIKA, v. 17, n. 1, p. 39-48, 2019.

A185 KIRAÇ, M. Furkan *et al.* Automatically learning usage behavior and generating event sequences for black-box testing of reactive systems. Software Quality Journal, p. 1-23.

A186 DASHEVSKYI, Stanislav *et al.* TestREx: a framework for repeatable exploits. International Journal on Software Tools for Technology Transfer, v. 21, n. 1, p. 105-119, 2019.

A187 SALVA, Sébastien; REGAINIA, Loukmen. A catalogue associating security patterns and attack steps to design secure applications. Journal of Computer Security, n. Preprint, p. 1-26, 2019.

A188 SHARIFF, Shahnaz Mohammedi. Investigating Selenium Usage Challenges and Reducing the Performance Overhead of Selenium-based Load Tests. 2019. Tese de Doutorado.

A189 LIU, Xiao *et al.* Automatic Grading of Programming Assignments: An Approach Based on Formal Semantics. In: Proceedings-International Conference on Software Engineering. 2019.

A190 ALTHOMALI, Ibrahim; KAPFHAMMER, Gregory M.; MCMINN, Phil. Automatic Visual Verification of Layout Failures in Responsively Designed *Web* Pages. In: 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST). IEEE, 2019. p. 183-193.

A191 GAO, Pengfei *et al.* Model-based Automated Testing of JavaScript *Web* Applications via Longer Test Sequences. arXiv preprint arXiv:1905.07671, 2019.

A192 BAJAJ, Anu; SANGWAN, Om Prakash. Study the Impact of Parameter Settings and Operators Role for Genetic Algorithm Based Test Case Prioritization. Available at SSRN 3356318, 2019.

A193 ELSAYED, Marwa; ZULKERNINE, Mohammad. Offering security diagnosis as a service for cloud SaaS applications. Journal of information security and applications, v. 44, p. 32-48, 2019.

Apêndices

APÊNDICE A – TABELAS RESULTADO

Tabela 1– Artigos que fazem e que não fazem referência as suas ameaças à validade

2010	Possui ameaça?		Abordagem simples		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A1		X	X		
A2		X	X		
A3		X	X		
A4	X			X	
A5		X	X		
A6	X			X	
A7	X			X	
A8	X			X	
A9	X			X	
A10	X			X	
A11		X	X		
A12		X	X		
A13		X	X		
A14		X	X		
A15		X	X		
A16		X	X		
A17		X	X		
A18		X	X		
A19		X	X		
A20		X	X		
Totais ano 2010	**6**	**14**	**14**	**6**	**0**
2011	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A21		X	X		
A22		X	X		
A23		X	X		
A24		X	X		
A25	X			X	
A26	X			X	

A27	X				X
A28	X			X	
A29	X			X	
A30		X	X		
A31	X			X	
A32	X			X	
A33		X	X		
A34	X			X	
A35	X			X	
A36	X			X	
A37	X			X	
A38		X	X		
A39		X	X		
A40		X	X		
Totais ano 2011	**11**	**9**	**9**	**10**	**1**
2012	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A41	X			X	
A42	X			X	
A43		X	X		
A44		X	X		
A45		X	X		
A46	X			X	
A47	X			X	
A48		X	X		
A49		X	X		
A50		X	X		
A51		X	X		
A52	X			X	
A53		X	X		
A54		X	X		
A55		X	X		
A56	X			X	
A57	X			X	
A58		X	X		
A59		X	X		
A60		X	X		
Totais ano 2012	**7**	**13**	**13**	**7**	**0**
2013	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**

A61	X			X	
A62	X			X	
A63		X	X		
A64		X	X		
A65	X			X	
A66	X			X	
A67	X			X	
A68	X			X	
A69	X			X	
A70	X			X	
A71	X			X	
A72	X				X
A73	X			X	
A74	X			X	
A75	X			X	
A76		X	X		
A77		X	X		
A78		X	X		
A79		X	X		
A80		X	X		
Totais ano 2013	**13**	**7**	**7**	**12**	**1**
2014	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A81	X			X	
A82		X	X		
A83	X			X	
A84		X	X		
A85		X	X		
A86	X			X	
A87		X	X		
A88		X	X		
A89	X			X	
A90		X	X		
A91	X				X
A92	X			X	
A93	X			X	
A94	X			X	
A95		X	X		
A96	X			X	
A97		X	X		
A98	X			X	

A99		X	X		
A100		X	X		
Totais ano 2014	**10**	**10**	**10**	**9**	**1**
2015	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A101		X	X		
A102	X			X	
A103	X			X	
A104	X			X	
A105	X			X	
A106	X			X	
A107	X			X	
A108	X			X	
A109	X			X	
A110	X			X	
A111	X			X	
A112		X	X		
A113		X	X		
A114	X			X	
A115	X			X	
A116	X			X	
A117		X	X		
A118		X	X		
A119		X	X		
A120		X	X		
Totais ano 2015	**13**	**7**	**7**	**13**	**0**
2016	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A121	X			X	
A122	X			X	
A123		X	X		
A124	X			X	
A125	X				X
A126		X	X		
A127	X			X	
A128		X	X		
A129		X	X		
A130	X			X	
A131		X	X		
A132		X	X		

A133	X			X	
A134		X	X		
A135		X	X		
A136		X	X		
A137		X	X		
A138		X	X		
A139		X	X		
A140	X			X	
Totais ano 2016	**8**	**12**	**12**	**7**	**1**
2017	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A141	X			X	
A142		X	X		
A143		X	X		
A144		X	X		
A145	X			X	
A146	X				X
A147		X	X		
A148	X			X	
A149		X	X		
A150	X			X	
A151	X			X	
A152	X			X	
A153		X	X		
A154	X			X	
A155		X	X		
A156		X	X		
A157	X			X	
A158		X	X		
A159	X			X	
A160		X	X		
Totais ano 2017	**10**	**10**	**10**	**9**	**1**
2018	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A161		X	X		
A162	X			X	
A163		X	X		
A164		X	X		
A165		X	X		
A166		X	X		

A167		X	X		
A168		X	X		
A169		X	X		
A170		X	X		
A171	X			X	
A172	X			X	
A173	X			X	
A174		X	X		
A175		X	X		
A176		X	X		
A177		X	X		
A178		X	X		
A179	X			X	
A180		X	X		
Totais ano 2018	**5**	**15**	**15**	**5**	**0**
2019	**Possui ameaça?**		**Abordagem simples**		
Cód. Artigo	**SIM**	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
A181	X			X	
A182	X			X	
A183		X	X		
A184		X	X		
A185	X			X	
A186		X	X		
A187	X			X	
A188	X			X	
A189		X	X		
A190	X			X	
A191		X	X		
A192		X	X		
A193		X	X		
Totais ano 2019	**6**	**7**	**7**	**6**	**0**

Tabela 2– Totais e percentagem de artigos que fazem e que não fazem referência as suas ameaças à validade

Totais gerais	**Possui ameaça?**		**Abordagem simples**		
	SIM	**NÃO**	**Não Atendido**	**Parcialmente**	**Totalmente**
Qtd. Artigos	**89**	**104**	**104**	**84**	**5**
Percentagem	**46,1%**	**53,9%**	**53,9%**	**43,5%**	**2,6%**

Tabela 3– Artigos que fazem e que não fazem referência as suas ameaças à validade – frequência anual

Totais anuais	Possui ameaça?		Abordagem simples		
	SIM	NÃO	Não Atendido	Parcialmente	Totalmente
2010					
Qtd. Artigos	6	14	14	6	0
Percentagem	30,0%	70,0%	70,0%	30,0%	0,0%
2011					
Qtd. Artigos	11	9	9	10	1
Percentagem	55,0%	45,0%	45,0%	50,0%	5,0%
2012					
Qtd. Artigos	7	13	13	7	0
Percentagem	35,0%	65,0%	65,0%	35,0%	0,0%
2013					
Qtd. Artigos	13	7	7	12	1
Percentagem	65,0%	35,0%	35,0%	60,0%	5,0%
2014					
Qtd. Artigos	10	10	10	9	1
Percentagem	50,0%	50,0%	50,0%	45,0%	5,0%
2015					
Qtd. Artigos	13	7	13	7	0
Percentagem	65,0%	35,0%	65,0%	35,0%	0,0%
2016					
Qtd. Artigos	8	12	12	7	1
Percentagem	40,0%	60,0%	60,0%	35,0%	5,0%
2017					
Qtd. Artigos	10	10	10	9	1
Percentagem	50,0%	50,0%	50,0%	45,0%	5,0%
2018					
Qtd. Artigos	5	15	15	5	0
Percentagem	25,0%	75,0%	75,0%	25,0%	0,0%
2019					
Qtd. Artigos	6	7	7	6	0
Percentagem	46,2%	53,8%	53,8%	46,2%	0,0%

Tabela 4– Ameaças mais citadas

2010	Ameaça					Descrição	
Cód. Artigo	Conclusão	Construção	Externa	Interna	Não Formal	Sumarizada	Detalhada
A4	X		X	X		X	

A6	X	X	X	X		X	
A7			X	X		X	
A8			X			X	
A9		X	X	X		X	
A10			X	X		X	
Totais	**2**	**2**	**6**	**5**	**0**	**6**	**0**
2011	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A25			X	X		X	
A26		X	X	X			X
A27	X	X	X	X			X
A28			X	X		X	
A29			X	X		X	
A31*					X	X	
A32*					X	X	
A34		X	X	X			X
A35*					X	X	
A36	X		X			X	
A37		X	X				X
Totais	**2**	**4**	**8**	**6**	**3**	**7**	**4**
2012	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A41		X				X	
A42		X	X	X		X	
A46		X	X			X	
A47			X	X		X	
A52*					X	X	
A56		X	X	X		X	
A57		X	X			X	
Totais	**0**	**5**	**5**	**3**	**1**	**7**	**0**
2013	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A61*					X	X	
A62			X			X	
A65*					X	X	
A66		X	X	X		X	
A67*					X		X

A68*					X	X	
A69		X	X			X	
A70		X	X	X		X	
A71			X	X		X	
A72	X	X	X	X			X
A73		X	X	X		X	
A74	X	X	X	X		X	
A75*					X	X	
Totais	**2**	**6**	**8**	**6**	**5**	**11**	**2**
2014	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A81*					X	X	
A83*					X	X	
A86*					X	X	
A89		X	X			X	
A91	X	X	X	X			X
A92			X	X		X	
A93			X			X	
A94		X	X	X			X
A96			X	X		X	
A98		X	X	X		X	
Totais	**1**	**4**	**7**	**5**	**3**	**8**	**2**
2015	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A102*					X	X	
A103*					X	X	
A104		X	X	X			X
A105			X	X			X
A106*					X	X	
A107*					X	X	
A108			X			X	
A109			X	X		X	
A110		X	X			X	
A111*					X		X
A114			X	X		X	
A115				X		X	
A116*					X	X	
Totais	**0**	**2**	**6**	**5**	**6**	**10**	**3**

2016	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A121			X	X		X	
A122*					X	X	
A124*					X	X	
A125	X	X	X	X			X
A127*					X	X	
A130	X	X	X			X	
A133		X	X				X
A140	X		X			X	
Totais	**3**	**3**	**5**	**2**	**3**	**6**	**2**
2017	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A141*					X	X	
A145	X	X	X	X		X	
A146	X	X	X	X			X
A148*					X	X	
A150*					X	X	
A151	X	X	X	X		X	
A152		X	X	X			X
A154*					X	X	
A157			X	X		X	
A159*					X	X	
Totais	**3**	**4**	**5**	**5**	**5**	**8**	**2**
2018	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A162*					X	X	
A171	X	X	X	X		X	
A172			X	X		X	
A173*					X	X	
A179*					X		X
Totais	**1**	**1**	**2**	**2**	**3**	**4**	**1**
2019	**Ameaça**					**Descrição**	
Cód. Artigo	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
A181*					X	X	

A182*					X	X	
A185	X		X	X		X	
A187			X			X	
A188		X	X	X			X
A190*					X	X	
Totais	**1**	**1**	**3**	**2**	**3**	**5**	**1**

Tabela 5– Total de ameaças mais citadas

	Ameaça					**Descrição**	
Ano	**Conclusão**	**Construção**	**Externa**	**Interna**	**Não Formal**	**Sumarizada**	**Detalhada**
2010	2	2	6	5	0	6	0
2011	2	4	8	6	3	7	4
2012	0	5	5	3	1	7	0
2013	2	6	8	6	5	11	2
2014	1	4	7	5	3	8	2
2015	0	2	6	5	6	10	3
2016	3	3	5	2	3	6	2
2017	3	4	5	5	5	8	2
2018	1	1	2	2	3	4	1
2019	1	1	3	2	3	5	1
Totais	**15**	**32**	**55**	**41**	**32**	**72**	**17**

Tabela 6– Resumo da Quantidade de Artigos Encontrados, Baixados e Selecionados

Ano	**Qtd Artigos Encontrados**	**Qtd Artigos Baixados**	**Qtd Artigos Selecionados**
2010	80	35	20
2011	74	32	20
2012	66	28	20
2013	98	29	20
2014	85	39	20
2015	52	65	20
2016	84	36	20
2017	93	28	20
2018	95	27	20
2019	70	17	13
Totais	**797**	**336 (42,1%)**	**193 (24,2%)**

Tabela 7– Tipos de Publicação por Ano de Publicação

2010	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A1		X	
A2	X		
A3		X	
A4	X		
A5	X		
A6	X		
A7	X		
A8	X		
A9	X		
A10		X	
A11	X		
A12	X		
A13	X		
A14	X		
A15	X		
A16		X	
A17	X		
A18	X		
A19	X		
A20	X		
Totais ano 2010	**16**	**4**	**0**
2011	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A21		X	
A22		X	
A23	X		
A24		X	
A25	X		
A26		X	
A27	X		
A28	X		
A29		X	
A30	X		
A31	X		
A32	X		
A33	X		
A34	X		
A35	X		

A36	X		
A37	X		
A38	X		
A39		X	
A40	X		
Totais ano 2011	**14**	**6**	**0**
2012	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A41	X		
A42	X		
A43	X		
A44		X	
A45	X		
A46	X		
A47	X		
A48	X		
A49		X	
A50	X		
A51	X		
A52		X	
A53	X		
A54	X		
A55	X		
A56	X		
A57	X		
A58	X		
A59	X		
A60	X		
Totais ano 2012	**17**	**3**	**0**
2013	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A61	X		
A62	X		
A63		X	
A64	X		
A65		X	
A66	X		
A67	X		
A68	X		
A69	X		

A70	X		
A71	X		
A72			X
A73	X		
A74	X		
A75	X		
A76	X		
A77	X		
A78	X		
A79	X		
A80	X		
Totais ano 2013	**17**	**2**	**1**
2014	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A81	X		
A82	X		
A83	X		
A84	X		
A85		X	
A86	X		
A87			X
A88	X		
A89	X		
A90	X		
A91	X		
A92	X		
A93	X		
A94	X		
A95	X		
A96	X		
A97	X		
A98	X		
A99	X		
A100	X		
Totais ano 2014	**18**	**1**	**1**
2015	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A101	X		
A102	X		
A103	X		

A104	X		
A105	X		
A106	X		
A107	X		
A108	X		
A109	X		
A110	X		
A111	X		
A112	X		
A113	X		
A114	X		
A115	X		
A116	X		
A117	X		
A118	X		
A119	X		
A120	X		
Totais ano 2015	**20**	**0**	**0**
2016	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A121	X		
A122	X		
A123	X		
A124	X		
A125	X		
A126	X		
A127	X		
A128	X		
A129	X		
A130	X		
A131	X		
A132	X		
A133	X		
A134	X		
A135	X		
A136	X		
A137	X		
A138	X		
A139	X		
A140	X		
Totais ano 2016	**20**	**0**	**0**

2017	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A141	X		
A142	X		
A143	X		
A144	X		
A145	X		
A146	X		
A147	X		
A148	X		
A149	X		
A150	X		
A151	X		
A152	X		
A153	X		
A154	X		
A155	X		
A156	X		
A157	X		
A158	X		
A159		X	
A160	X		
Totais ano 2017	**19**	**1**	**0**
2018	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A161	X		
A162	X		
A163	X		
A164		X	
A165	X		
A166	X		
A167	X		
A168	X		
A169	X		
A170	X		
A171	X		
A172		X	
A173	X		
A174	X		
A175	X		

A176	X		
A177	X		
A178	X		
A179	X		
A180	X		
Totais ano 2018	**18**	**2**	**0**
2019	**Tipo de artigo**		
Cód. Artigo	**Experimento**	**Estudo de caso**	***Survey***
A181	X		
A182	X		
A183	X		
A184	X		
A185	X		
A186	X		
A187	X		
A188	X		
A189	X		
A190	X		
A191	X		
A192	X		
A193	X		
Totais ano 2019	**13**	**0**	**0**
Total geral (qtd)	**172**	**19**	**2**
Total geral (%)	**89,1%**	**9,9%**	**1**

Tabela 8– Resumo dos resultados encontrados

Artigos	**Quantidade**	**Percentual**
Investigados	797	100,0%
Baixados	336	42,1%
Estudados	193	24,2%
Que referenciaram suas ameaças à validade	89	46,1%
Que NÃO referenciaram suas ameaças à validade	104	53,9%
Que referenciaram suas ameaças de maneira NÃO formal	32	36,0%
Abordagem simples - critério NÃO atendido	104	53,9%
Abordagem simples - critério PARCIALMENTE atendido	72	37,3%
Abordagem simples - critério TOTALMENTE atendido	5	2,6%
Que citaram a ameaça EXTERNA	55	28,5%
Que citaram a ameaça INTERNA	41	21,2%
Que citaram a ameaça de CONSTRUÇÃO	32	16,6%
Que citaram a ameaça de CONCLUSÃO	15	7,8%
Que referenciaram suas ameaças de forma SUMARIZADA	72	37,3%

Que referenciaram suas ameaças de forma DETALHADA	17	8,8%
Classificados como EXPERIMENTO	172	89,1%
Classificados como ESTUDO DE CASO	19	9,9%
Classificados como *SURVEY*	2	1,0%
Que cumpriram integralmente os rigores científicos	5	2,6%