



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EUDIS OLIVEIRA TEIXEIRA

PRIORITTVS: um processo que se destina a apoiar os pesquisadores a priorizar ameaças à validade e suas ações de mitigação ao planejar experimentos da área de engenharia de software

Recife
2020

EUDIS OLIVEIRA TEIXEIRA

PRIORITTVS: um processo que se destina a apoiar os pesquisadores a priorizar ameaças à validade e suas ações de mitigação ao planejar experimentos da área de engenharia de software

Tese apresentada ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciências da Computação.

Área de concentração: Engenharia de Software.

Orientador: Prof. Dr. Sergio Castelo Branco Soares.

Co-orientadora: Profa. Dra. Liliane Sheyla da Silva Fonseca.

Recife

2020

Catálogo na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

T266p Teixeira, Eudis Oliveira
PRIORITTVS: um processo que se destina a apoiar os pesquisadores a priorizar ameaças à validade e suas ações de mitigação ao planejar experimentos da área de engenharia de software / Eudis Oliveira Teixeira. – 2020.
228 f.:il., fig, tab.

Orientador: Sergio Castelo Branco Soares.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2020.

Inclui referências, apêndices e anexos.

1. Engenharia de software. 2. Ações de controle. I. Soares, Sergio Castelo Branco (orientador). II. Título.

005.1 CDD (23. ed.) UFPE - CCEN 2022-192

EUDIS OLIVEIRA TEIXEIRA

PRIORITTVS: um processo que se destina a apoiar os pesquisadores a priorizar ameaças à validade e suas ações de mitigação ao planejar experimentos da área de engenharia de software

Tese apresentada ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciências da Computação.

Aprovada em: 20/01/2020.

BANCA EXAMINADORA

Prof^a. Dr. André Luís de Medeiros Santos
UFPE | Centro de Informática

Prof. Dr. Juliano Manabu Iyoda
UFPE | Centro de Informática

Prof. Dr. Leopoldo Motta Teixeira
UFPE | Centro de Informática

Prof. Dr. Bruno Falcão de Souza Cartaxo
IFPE | Instituto Federal de Pernambuco

Prof. Dr. Eduardo Henrique da Silva Aranha
UFRN | Departamento de Informática e Matemática Aplicada

AGRADECIMENTOS

Gratidão a Deus, por iluminar minha trajetória e me dar forças para seguir sempre em frente. À Dayany Teixeira, minha esposa, pelo carinho, paciência e suporte à família. Aos meus pequenos notáveis, Alice (9), Clara (5) e Pedro (2), por me ensinarem o que é o amor incondicional. Aos meus pais e irmãos, mesmo distantes, porém, sempre apoiando minha qualificação e crescimento profissional.

Aos meus orientadores, Sérgio Soares e Liliane Fonseca, muito obrigado pela oportunidade, confiança, flexibilidade, e, principalmente, pelo empenho durante o doutorado. Ao Instituto Federal do Sertão Pernambucano (IF-Sertão), ao Centro de Informática da UFPE e também o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela brilhante parceria. Muito obrigado!

“Tentar e falhar é, pelo menos, aprender. Não chegar a tentar é sofrer a inestimável perda do que poderia ter sido” (EUSTÁQUIO, 2019, p. 1).

RESUMO

Em geral, os pesquisadores argumentam que as ameaças à validade é um dos componentes mais complexos de um plano experimental em Engenharia de Software (ES). Um dos motivos da dificuldade se deve ao fato do grande número de possibilidades de problemas que podem transcorrer desde o planejamento até a divulgação dos resultados da avaliação experimental, bem como, o tratamento da uma ameaça pode resultar no surgimento de novas ameaças à validade, não identificadas inicialmente. Embora existam estudos que abordam esta temática, a literatura não fornece apoio ao experimentador na priorização dos riscos mais relevantes ao propósito de seu estudo experimental. Isto resulta na dificuldade em elaborar ações de controle assertivas para mitigar os riscos. O principal objetivo desta pesquisa é desenvolver um processo estruturado, a fim de apoiar experimentadores, principalmente iniciantes, na identificação e também priorização de ameaças à validade e suas correspondentes ações de controle no contexto de um planejamento experimental em ES. Inicialmente, foi realizado um survey com 115 experimentadores experientes para compreender principais procedimentos para o controle das ameaças à validade, verificar suas percepções em relação às práticas reais de tratamento destas ameaças e coletar as referências mais utilizadas na literatura por estes especialistas. Em seguida, processo foi proposto e avaliado através de diferentes objetivos, tais como, completude, facilidade de uso, percepção de satisfação e aprendizagem, além de ser avaliado por abordagens distintas como experimento controlado, questionários e entrevistas. Como resultado desta pesquisa foi desenvolvido o PRIORITTVS, processo de suporte para priorizar ameaças à validade e sugerir ações de mitigação no planejamento de experimentos da área de ES, como também foi desenvolvido uma ferramenta web para viabilizar o processo de forma semiautomática aos experimentadores. Tanto o processo quanto a ferramenta foram avaliados a partir dos estudos experimentais com diferentes objetivos e perspectivas, por 26 participantes iniciantes e experientes. Os resultados das avaliações realizadas trazem evidências que o processo e ferramenta propostos são viáveis e possuem um bom nível de completude, facilidade de uso, percepção de satisfação e aprendizagem. No entanto, faz-se necessária a realização de mais estudos de avaliação, para que resultados possam ser generalizados. A estratégia possibilita, principalmente aos experimentadores menos experientes, acesso ao processo estruturado, que possa

ajudar a fornecer suporte à identificação e priorização de ameaças a validade e suas respectivas ações de mitigação. Por fim, pretende-se estender a abordagem para gerenciar as ameaças específicas para os diferentes contextos de experimentos em ES.

Palavras-chave: ameaças à validade; ações de controle; engenharia de software experimental; experimentos controlados.

ABSTRACT

Researchers argue that a critical component of one experimental plan in Software Engineering (SE) is to identify, analyze, and mitigate threats to validity. One reason for the difficulty is due to the number of potential problems that can occur from planning to reporting the results of an experimental evaluation. Likewise, when mitigating a threat can result in new threats not initially identified. Although there are studies that address this theme, the literature does not provide support to the experimenter in prioritizing the most critical risks in the context of his experimental study. These facts result in the difficulty of applying assertive control actions to mitigate risks. The main aim of this research is to develop a structured process to support mainly novice experimenters in identifying and prioritizing threats to validity and their corresponding mitigation actions when planning controlled experiments in SE. Initially, a survey of 115 experienced researchers was conducted to understand the main procedures for controlling threats to validity, verifying their perceptions of actual threat management practices, and collecting the references most commonly used in the literature by these experts. Then, a process was proposed and evaluated through different objectives, such as correctness, ease of use, perception of satisfaction and learning in addition to being assessed through different research approaches such as controlled experiment, questionnaires and interviews. As a result of this research, PRIORITTVS, a support process to prioritize threats to validity and suggest mitigation actions in the context of one experimental plan in SE. A software tool was also designed to make the process semi-automatic to the experimenters. Both the process and the tool were evaluated through empirical studies with different objectives and perspectives by 26 novice and experienced participants. The results of the evaluations provide evidence that the proposed process and tool are viable and have a good level of correctness, ease of use, perception of satisfaction and learning. However, further evaluation studies are necessary for the results to be generalized. We believe both novice and expert experimenters can benefit from the process we have propose for addressing threats to validity when conducting SE experiments. Finally, we intend to extend our approach to managing specific threats for different SE experiment context.

Keywords: threats to validity; control actions; experimental software engineering; controlled experiments.

LISTA DE FIGURAS

Figura 1 –	Classificação das ameaças à validade dos experimentos.....	35
Figura 2 –	Por que identificar as ameaças à validade de um experimento?	47
Figura 3 –	Melhor momento para identificar as ameaças à validade.....	47
Figura 4 –	Estratégias para identificar e controlar as ameaças à validade.....	48
Figura 5 –	Há uma ordem de importância entre tipos de validade?.....	49
Figura 6 –	Qual a ordem de importância entre os tipos de validade?.....	49
Figura 7 –	Preocupação com relacionamentos entre ameaças e ações de controle.....	50
Figura 8 –	PrioriTTVs: processo para identificar e priorizar ameaças à validade.....	59
Figura 9 –	Identificar as características do experimento.....	69
Figura 10 –	Classificar as ameaças preliminares.....	70
Figura 11 –	Classificar as ações de controle.....	71
Figura 12 –	Visão Geral dos Casos de Uso da ValidEPlan.....	72
Figura 13 –	Dashboard da Ferramenta ValidEPlan.....	74
Figura 14 –	Produzir um novo plano experimental.....	75
Figura 15 –	Revisar um Plano Experimental.....	76
Figura 16 –	Identificar informações do experimento.....	77
Figura 17 –	Classificar ameaças à validade sugeridas para o experimento.....	78
Figura 18 –	Sexo dos participantes da pesquisa (1ª Etapa)	86
Figura 19 –	Instituição de ensino ou trabalho do participante.....	87
Figura 20 –	Posição atual dos participantes relacionada à educação.....	87
Figura 21 –	Total de experimentos que o especialista participou do planejamento.....	88
Figura 22 –	Total de experimentos que o especialista participou da execução.....	89
Figura 23 –	Nível de importância que o pesquisador oferece à ferramenta de planejamento e gestão de ameaças para experimentos.....	90
Figura 24 –	Modelo de avaliação das ameaças sugeridas.....	100
Figura 25 –	Escala de magnitude.....	101
Figura 26 –	Modelo de avaliação das ações de controle do experimento.....	102

Figura 27 –	Percepção de utilidade da ferramenta ValidEPlan.....	105
Figura 28 –	Leitura e compreensão de textos em língua inglesa.....	117
Figura 29 –	Áreas de conhecimento em Computação.....	118
Figura 30 –	Conhecimento sobre Engenharia de Software Experimental.....	118
Figura 31 –	Nível de conhecimento com Planejamento de Experimentos.....	119
Figura 32 –	Nível de experiência com Planejamento de Experimentos.....	119
Figura 33 –	Proporção de respostas relacionadas aos tratamentos.....	133
Figura 34 –	Temáticas relacionadas aos tratamentos.....	134
Figura 35 –	Percentuais para a subescala Percepção de Aprendizagem.....	149
Figura 36 –	Percentuais para a subescala Percepção de Facilidade.....	150
Figura 37 –	Percentuais para a subescala Percepção de Satisfação.....	151

LISTA DE TABELAS

Tabela 1 –	Guidelines para conduzir e reportar experimentos em ES.....	28
Tabela 2 –	Estudos que reportaram experimentos.....	40
Tabela 3 –	Características dos participantes.....	45
Tabela 4 –	Importância das ameaças à validade.....	46
Tabela 5 –	Categorias de Análise.....	51
Tabela 6 –	Crítérios, pesos e intensidades de classificação de ameaças.....	60
Tabela 7 –	Escala e Categorias de Magnitude de Ameaças à Validade.....	61
Tabela 8 –	Visão geral dos estudos avaliados.....	80
Tabela 9 –	Etapas do Planejamento da Avaliação com Especialistas.....	82
Tabela 10 –	Classificação das ameaças à validade experimental identificadas por tipologia.....	94
Tabela 11 –	Lista de verificação por categoria em instrumento de coleta quanto à avaliação do processo.....	96
Tabela 12 –	Resumo quantitativo das respostas sobre aprendizado do sistema.....	107
Tabela 13 –	Etapas do Planejamento da Avaliação com Estudantes.....	115
Tabela 14 –	Temáticas abordadas no Modelo de Referência.....	124
Tabela 15 –	Organização das Réplicas no Quadrado Latino.....	126
Tabela 16 –	Aleatorização dos Quadrados Latinos.....	130
Tabela 17 –	Respostas obtidas a partir dos participantes.....	131
Tabela 18 –	Testes de Normalidade.....	135
Tabela 19 –	Respostas relacionadas aos tratamentos.....	135
Tabela 20 –	Respostas obtidas a partir dos participantes.....	137
Tabela 21 –	Testes de Normalidade.....	138
Tabela 22 –	Teste de hipótese estatística Wilcoxon.....	139
Tabela 23 –	Principais ameaças identificadas através do processo PrioriTTVs.....	141
Tabela 24 –	Possíveis relacionamentos entre Ações de Controle e Ameaças à Validade.....	144
Tabela 25 –	Estatística de Confiabilidade Alfa de Cronbach.....	147
Tabela 26 –	Análises e Escores para Aprendizagem.....	147

Tabela 27 – Análises e Escores para Facilidade.....	149
Tabela 28 – Análises e Escores para Satisfação.....	150
Tabela 29 – Frequência dos Temas relacionados à qualidade dos resultados.....	153
Tabela 30 – Frequência dos Temas relacionados a Identificação de Ameaças à Validade.....	155
Tabela 31 – Frequência dos Temas relacionados às percepções da ferramenta.....	157
Tabela 32 – Frequência dos Temas relacionados às percepções da ferramenta.....	160
Tabela 33 – Frequência dos Temas relacionados aos comentários e oportunidades.....	162

LISTA DE ABREVIATURAS E SIGLAS

CBSOFT	<i>Congresso Brasileiro de Software</i>
CIn	<i>Centro de Informática</i>
EASE	<i>International Conference on Evaluation and Assessment in Software Engineering</i>
EE	<i>Estudos Experimentais</i>
ES	<i>Engenharia de Software</i>
ESE	<i>Engenharia de Software Experimental</i>
ESEG	<i>Empirical Software Engineering Group</i>
IDOESE	<i>International Doctoral Symposium on Empirical Software Engineering</i>
IF Sertão-PE	<i>Instituto Federal do Sertão Pernambucano</i>
IFPE	<i>Instituto Federal de Pernambuco</i>
UFPB	<i>Universidade Federal da Paraíba</i>
UFPE	<i>Universidade Federal de Pernambuco</i>
UPE	<i>Universidade de Pernambuco</i>
ValidEPlan	<i>Validity-Driven Software Engineering Experiments Planning Tool</i>

SUMÁRIO

1	INTRODUÇÃO.....	18
1.1	PROBLEMA, MOTIVAÇÃO E JUSTIFICATIVA	18
1.2	OBJETIVOS DA PESQUISA.....	20
1.3	QUESTÕES DE PESQUISA	20
1.4	ESTRATÉGIA DE PESQUISA	21
1.5	CONTRIBUIÇÕES.....	22
1.6	CONQUISTAS RESULTADOS	23
1.7	ORGANIZAÇÃO DA TESE	24
2	FUNDAMENTAÇÃO TEÓRICA.....	25
2.1	INTRODUÇÃO.....	25
2.2	ESTUDOS EXPERIMENTAIS NA ENGENHARIA DE SOFTWARE	26
2.2.1	Survey	26
2.2.2	Experimentos em Engenharia de Software.....	27
2.2.3	Experimentos em outras áreas.....	30
2.3	PLANEJAMENTO EXPERIMENTAL	31
2.3.1	Dificuldades no Planejamento Experimental.....	32
2.4	AMEAÇAS À VALIDADE	33
2.4.1	Priorização das Ameaças à Validade.....	36
2.5	RESUMO	37
3	<i>SURVEY</i> – AS AMEAÇAS À VALIDADE EM EXPERIMENTOS CONTROLADOS: O QUE OS <i>EXPERTS</i> DIZEM E POR QUE ISSO É RELEVANTE?.....	38
3.1	INTRODUÇÃO.....	38
3.2	OBJETIVOS E DEFINIÇÕES.....	38
3.3	AMOSTRA DO ESTUDO	39
3.4	INSTRUMENTO DE COLETA DE DADOS	40
3.5	ESTUDO PILOTO.....	41
3.6	EXECUÇÃO.....	42
3.7	ANÁLISE DOS DADOS.....	42
3.8	RESULTADOS E DISCUSSÕES.....	44
3.8.1	Análise Quantitativa	45
3.8.2	Análise Qualitativa.....	50

3.9	AMEAÇAS À VALIDADE	53
3.10	CONCLUSÃO	54
3.11	RESUMO	54
4	PRIORITTVS: UM PROCESSO QUE SE DESTINA A APOIAR PESQUISADORES A PRIORIZAR AMEAÇAS À VALIDADE E AS SUAS AÇÕES DE MITIGAÇÃO AO PLANEJAR OS EXPERIMENTOS CONTROLADOS EM ES	56
4.1	INTRODUÇÃO.....	56
4.2	TRABALHOS RELACIONADOS.....	57
4.3	MODELAGEM CONCEITUAL.....	58
4.3.1	Identificação das Informações de um Plano Experimental (A1)	59
4.3.2	Classificação das Ameaças Preliminares (A2)	59
4.3.3	Classificação da assertividade das Ações de Controle (A3).....	62
4.4	INSTRUMENTO DE CARACTERIZAÇÃO DE UM PLANO EXPERIMENTAL	63
4.5	RESUMO	65
5	VALIDEPLAN: FERRAMENTA PARA PLANEJAMENTO DE EXPERIMENTOS DE ENGENHARIA DE SOFTWARE ORIENTADA À VALIDADE.....	67
5.1	INTRODUÇÃO.....	67
5.2	VALIDEPLAN – VALIDITY-DRIVEN SOFTWARE ENGINEERING EXPERIMENTS PLANNING TOOL	67
5.2.1	Definições e Requisitos da ValiDEPlan	68
5.2.2	Arquitetura da ValidEPlan	73
5.2.3	Funcionalidades da ValidEPlan	74
5.3	RESUMO	79
6	AVALIAÇÃO EXPERIMENTAL.....	80
6.1	INTRODUÇÃO.....	80
6.2	AVALIAÇÃO EXPERIMENTAL POR ESPECIALISTAS EM ESE	82
6.2.1	ETAPA 1: Análise dos Dados Demográficos e de Expectativas	85
6.2.2	ETAPA 2: Avaliação do processo PrioriTTVs.....	92
6.2.2.1	Identificação às ameaças à validade	92
6.2.2.2	Classificação das ameaças à validade	99
6.2.2.3	Classificação das ações de controle.....	102

6.2.3	ETAPA 3: Avaliação da ferramenta ValidEPlan	103
6.2.4	ETAPA 4: Entrevista Narrativa Episódica	108
6.3	AVALIAÇÃO EXPERIMENTAL POR ESTUDANTES EM ESE	115
6.3.1	ETAPA 1: Análise dos Dados Demográficos	116
6.3.2	ETAPA 2: Experimento Controlado.....	120
6.3.2.1	Definição	120
6.3.2.2	Questões de Pesquisa	121
6.3.2.3	Preocupações Éticas.....	121
6.3.2.4	Hipóteses, Variáveis e Medidas.....	121
6.3.2.5	Materiais Experimentais e Tarefas.....	123
6.3.2.6	Seleção dos participantes	125
6.3.2.7	Design Experimental	125
6.3.2.8	Procedimentos.....	126
6.3.2.9	Preparação da Execução do Experimento	128
6.3.2.10	Resultados Alcançados.....	130
6.3.2.11	Ameaças à Validade.....	140
6.3.3	Interpretação dos Dados do Questionário	145
6.3.3.1	Percepção de Aprendizagem.....	147
6.3.3.2	Percepção de Facilidade	149
6.3.3.3	Percepção de Satisfação	150
6.3.4	ETAPA 3: Entrevista Narrativa Episódica	152
6.4	RESUMO	164
7	CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS	165
7.1	ANÁLISE DAS QUESTÕES DE PESQUISA DA TESE	165
7.2	REVISÃO DAS CONTRIBUIÇÕES.....	169
7.3	LIMITAÇÕES DO TRABALHO.....	172
7.4	TRABALHOS FUTUROS	172
7.5	RESUMO	173
	REFERÊNCIAS.....	175
	APÊNDICE A – AN EXPERIMENTAL PLANS CHARACTERIZATION CHECKLIST FOR CONTROLLED EXPERIMENTS IN SOFTWARE ENGINEERING.....	182
	APÊNDICE B – ROTEIRO DE TAREFA: CENÁRIO 3 – IDENTIFICAÇÃO E CLASSIFICAÇÃO DE AMEAÇAS À VALIDADE.....	186

APÊNDICE C – ROTEIRO DE ENTREVISTA NARRATIVA EPISÓDICA – ESPECIALISTAS EM ENGENHARIA DE SOFTWARE EXPERIMENTAL	188
APÊNDICE D – FORMULÁRIO DE CONSENTIMENTO DO PARTICIPANTE	190
APÊNDICE E – ROTEIRO DE ENTREVISTA NARRATIVA EPISÓDICA COM PESQUISADORES INICIANTE	192
APÊNDICE F – QUESTIONÁRIO DE AVALIAÇÃO DA FERRAMENTA VALIDEPLAN	193
ANEXO A – DESCRIÇÃO DA UNIDADE EXPERIMENTAL – UE1	196
ANEXO B – DESCRIÇÃO DA UNIDADE EXPERIMENTAL – UE2	205
ANEXO C – BANCO DE DADOS DE AMEAÇAS E AÇÕES DE CONTROLE	214

1 INTRODUÇÃO

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.”

Arthur Schopenhauer

Este capítulo relata as principais motivações para a realização deste trabalho. Ele aborda a justificativa para realização desta tese, questões de pesquisa, objetivos a serem alcançados, abordagem de pesquisa e a estrutura deste documento.

1.1 PROBLEMA, MOTIVAÇÃO E JUSTIFICATIVA

Pesquisadores têm desenvolvido mecanismos de apoio que permitem planejar, executar, analisar e reportar Estudos Experimentais (EE) em Engenharia de Software (ES). Entretanto, reportar resultados experimentais válidos, ainda é um dos desafios da comunidade de ES, pois questão central a respeito de experimentos é o “quão corretos são seus resultados?” (KITCHENHAM *et al.*, 2002; JEDLITSCHKA; CIOLKOWSKI, 2005; WOHLIN *et al.*, 2012; FONSECA, 2016; FREIRE, 2015).

Segundo Neto e Conte (2013), para que conclusões corretas sejam extraídas de um experimento, os pesquisadores devem identificar e tratar os fatores que podem fazer com que os resultados sejam mal interpretados, desde o planejamento até a divulgação dos resultados do estudo.

As ameaças à validade são um dos fatores que podem alterar os resultados de um experimento comprometendo a validade (WAZLAWICK, 2017). Estes problemas podem trazer sérios riscos à confiabilidade de pesquisas e tecnologias promissoras da área de ES. Por exemplo, pode haver uma imprecisão nos resultados obtidos, fruto de uma medição incorreta ou uso de materiais e métodos impróprios ao delineamento experimental do estudo. Conseqüentemente, estas ameaças precisam ser identificadas e controladas adequadamente, para que os resultados do estudo sejam úteis e aplicáveis (DIAS-NETO *et al.*, 2004).

De acordo com entendimento de Wohlin et al (2012), ameaças a validade são os possíveis riscos que podem acontecer durante o planejamento e execução de estudos experimentais e podem ser categorizadas em quatro tipos de validade: interna, conclusão, constructo e externa.

Devido a isto, é importante encontrar as soluções que possam melhorar a confiabilidade das conclusões e recomendações resultantes de experimentos da área de ES (WOHLIN *et al.*, 2012), pois a partir das relações entre as ameaças à validade e suas respectivas ações de controle, podem surgir novas ameaças (NETO; CONTE, 2013).

O escopo deste trabalho foca em ameaças à validade no planejamento de experimentos controlados em ES. Wohlin *et al.* (2012) definem um experimento controlado como uma investigação empírica que manipula um fator ou variável do ambiente estudado de forma sistemática, disciplinada, quantificável e controlada, permitindo assim a avaliação de novas tecnologias. No âmbito deste trabalho, iremos usar o termo “experimento” no mesmo sentido de “experimento controlado”.

A literatura descreve exemplos, ferramenta, guidelines, modelos e checklists que ajudam o pesquisador a identificar e tratar ameaças à validade de experimentos da área de ES (DIESTER *et al.*, 2011; DIAS-NETO *et al.*, 2011; WOHLIN *et al.*, 2012; NETO; CONTE, 2013).

No entanto, estes estudos apoiam o processo de controle das ameaças à validade sem fornecer suporte em relação aos principais aspectos/características dos experimentos. As estratégias mencionadas na literatura, também, não apresentam mecanismos de suporte que apoiem o experimentador a priorizar ameaças à validade, consideradas importantes ao seu propósito de investigação. De acordo com Borges *et al.* (2015), um mecanismo é qualquer recurso que dê suporte a estratégias experimentais, como ferramentas, metodologias, processos, diretrizes, etc. Também se considera mecanismos de suporte os recursos utilizados para analisar os resultados do estudo (qualitativos e quantitativos) ou apenas para orientar sua validação.

Priorizar determinadas ameaças à validade permite que os pesquisadores identifiquem maiores riscos à confiabilidade dos resultados do estudo, planejando ações de controle mais assertivas. Segundo Wohlin *et al.* (2012), existem conflitos entre alguns dos tipos de ameaças à validade que, ao tentar controlar uma determinada ameaça, fazem surgir ameaças não identificadas inicialmente.

Por exemplo, para minimizarmos os efeitos de fadiga e cansaço durante a realização de um experimento (ameaça interna), podemos inserir intervalos entre as execuções do experimento (ação de controle). No entanto, durante as lacunas de tempo inseridas ao longo do experimento, pode haver comunicação entre os

participantes, surgindo a nova ameaça à validade do estudo (TEIXEIRA; FONSECA; SOARES, 2018).

Para superar as dificuldades mencionados, este trabalho propõe uma maneira semiautomática para orientar novos pesquisadores a decidirem como priorizar determinadas ameaças à validade, considerando os relacionamentos de causa e de consequência que existem entre as ameaças à validade e possíveis ações de controle.

Com isso, buscamos a melhoria da qualidade dos futuros experimentos da área de ES através de solução que possa ajudar aos pesquisadores, principalmente aos menos experientes, obter conclusões com um maior nível de validade interna, externa, de construção e de constructo sobre os resultados de seus estudos.

1.2 OBJETIVOS DA PESQUISA

O objetivo geral desta pesquisa é *propor uma abordagem que promova uma melhor qualidade aos experimentos controlados em ES no aspecto da validade de seus resultados*. O trabalho tem os seguintes objetivos específicos:

1. Investigar o estado da arte para apoiar o processo de identificação e controle de ameaças à validade de experimentos controlados de ES;
2. Definir processo de suporte à priorização e controle de ameaças à validade durante o planejamento de experimentos;
3. Desenvolver uma ferramenta web para viabilizar o processo de forma semiautomática aos experimentadores;
4. Avaliar a abordagem e a ferramenta propostas a partir da execução de diferentes estudos experimentais.

1.3 QUESTÕES DE PESQUISA

Com base na descrição do problema e objetivos anteriormente apresentados, definimos as seguintes questões de pesquisa que guiarão esta tese:

- **Questão 1:** Quais são os procedimentos realizados por pesquisadores para identificar e controlar as ameaças à validade ao planejar experimentos controlados em ES? Qual a importância desse processo?

- **Questão 2:** Quais são as recomendações indicadas por pesquisadores para melhorar a validade de resultados de experimentos da área de ES?
- **Questão 3:** Como o planejamento de experimentos da área de ES pode ser apoiado quanto ao processo de identificação e priorização das ameaças à validade?
- **Questão 4:** Quais são os benefícios e limitações da abordagem proposta para gerir o processo de identificação e priorização das ameaças à validade?

1.4 ESTRATÉGIA DE PESQUISA

Com o fim de responder a questões de pesquisas apresentadas anteriormente, e alcançar objetivos específicos do trabalho, foram definidas as seguintes etapas da estratégia de pesquisa:

- **Etapa 1:** Planejamento e realização de um *survey* através do uso de questionário para coleta de dados aplicado aos praticantes de experimentos controlados da área de ES com o propósito exploratório de: (i) compreender as ações dos pesquisadores ao realizar os procedimentos para controle das ameaças à validade de experimentos controlados e; (ii) verificar a percepção entre as práticas reais de tratamento de ameaças a validade e as orientações e recomendações existentes na literatura. Esta etapa da estratégia de pesquisa, foi elaborada, objetivando atender ao primeiro objetivo específico da pesquisa, assim como, responder à primeira e segunda questão de pesquisa investigada.
- **Etapa 2:** Definição e modelagem de um processo para a identificação, classificação e priorização de ameaças à validade e suas respectivas ações de controle ao planejar experimentos controlados da área de ES. Esta etapa da estratégia foi elaborada objetivando atender ao segundo objetivo específico da pesquisa, assim como, responder parte da terceira questão de pesquisa.
- **Etapa 3:** Elaboração de uma lista de verificação (*checklist*), com a finalidade de capturar informações sobre os principais aspectos de um plano experimental da área de ES, estruturado de acordo com as categorias: Objetivos, hipóteses, variáveis e medidas, participantes, materiais e tarefas experimentais, design experimental e coleta e análise de dados. A etapa três deste planejamento foi criada objetivando atender ao segundo objetivo específico da pesquisa, assim como, responder parte da terceira questão de investigação.

- **Etapa 4:** Especificação e desenvolvimento de uma ferramenta de suporte ao planejamento de experimentos controlados da área de ES, orientada à validade dos seus resultados. Já esta etapa da estratégia, foi elaborada objetivando atender ao terceiro objetivo específico da pesquisa, assim como, responder parte da terceira questão de pesquisa.
- **Etapa 5:** Realização de avaliações experimentais com o objetivo de: (i) avaliar os procedimentos propostos para identificação e priorização de ameaças à validade e suas respectivas ações de mitigação e; (ii) avaliar o impacto do uso da ferramenta desenvolvida, relacionado à facilidade de uso, satisfação e aprendizado dos usuários. Por fim, a etapa cinco da estratégia, foi empregada objetivando atender ao quarto e último objetivo específico e questão de investigação da pesquisa.

1.5 CONTRIBUIÇÕES

As principais contribuições do trabalho são relacionadas à:

Abordagem: (i) Redução do esforço do experimentador por meio do uso de um processo que possa ajudar no suporte à identificação e priorização de ameaças a validade e suas respectivas ações de mitigação durante a fase de planejamento de um experimento da área de ES e; (ii) Fornecer uma lista de verificação (*checklist*) que possa ajudar a capturar informações sobre as principais características de um plano experimental.

Ferramenta: (i) Desenvolvimento de uma ferramenta cujo objetivo está em fornecer suporte ao planejamento de experimentos controlados da área de ES, dirigido/orientado à validade dos seus resultados e; (ii) Contribuir para o ensino e aprendizagem da Engenharia de Software experimental, através do uso de uma ferramenta que possa ajudar a relacionar os conceitos teóricos estudados em sala de aula com a prática da experimentação.

Avaliação: Uso de entrevistas, experimento controlado e questionários para: (i) Avaliação do processo proposto para identificação e classificação de ameaças e suas respectivas ações de mitigação e; (ii) Verificação de percepções de aprendizagem, satisfação e facilidade do uso da ferramenta para planejar, identificar e classificar ameaças à validade em experimentos controlados da área de ES.

1.6 CONQUISTAS | RESULTADOS

Até o momento, os resultados advindos desta pesquisa são:

- Artigo publicado no periódico *Information and Software Technology*, Volume 115, 2019, Qualis A1.
 - **Teixeira, E.**, Fonseca, L., Cartaxo, B., & Soares, S. (2019). PrioriTTVs: A process aimed at supporting researchers to prioritize threats to validity and their mitigation actions when planning controlled experiments in SE. *Information and Software Technology*, <https://doi.org/10.1016/j.infsof.2019.07.008>.

- Artigo apresentado na Sessão de Ferramentas do X Brazilian Conference on *Software – Theory and Practice* (CBSOFT), 2019.
 - L. Fonseca, **E. Teixeira**, S. Soares, ValiDEPlan – *Validity-Driven Software Engineering Experiments Planning Tool*, in: Tools Session in X Brazilian Conference on Software: Theory and Practice (CBSOFT), Brazil, 2019, pp. 00–00. (in Portuguese), https://doi.org/10.5753/cbsoft_estendido.2019.7665
 - Escolhido como o Terceiro Melhor Artigo da Sessão de Ferramentas do Congresso Brasileiro de Software – CBSOFT 2019.

- Artigo apresentado no Simpósio Brasileiro de Engenharia de Software (SBES'18), in: IX Brazilian Conference on Software: Theory and Practice (CBSOFT), com o objetivo de compreender as ações dos pesquisadores ao realizar os procedimentos para controle das ameaças à validade, Qualis B2:
 - **Teixeira, E.**, Fonseca, L., & Soares, S. (2018, September). Threats to validity in controlled experiments in software engineering: what the experts say and why this is relevant. In *Proceedings of the XXXII Brazilian Symposium on Software Engineering* (pp. 52-61). ACM, <https://doi.org/10.1145/3266237.3266264>

- Artigo apresentado no 15º Simpósio Doutoral Internacional de Engenharia de Software Empírica (IDoESE 2017) da Semana Internacional de Engenharia de Software Empírica (Eseiw 2017) e em seguida publicado no *Journal ACM SIGSOFT Software Engineering Notes*, 2018:

- **Teixeira, E.** (2018). Improving the Quality of Controlled Experiments in Software Engineering. *ACM SIGSOFT Software Engineering Notes*, 43(1), 1-6, <https://doi.org/10.1145/3178315.3178321>
- Coautor de artigo apresentado na 19ª Conferência Internacional sobre Avaliação Experimental em Engenharia de Software (EASE 2015). Um dos resultados do estudo foi identificar principais mecanismos utilizados como referência ao planejamento e execução de experimentos controlados em ES, Qualis B1:
 - Borges, A., Ferreira, W., Barreiros, E., Almeida, A., Fonseca, L., **Teixeira, E.**, ... & Soares, S. (2015, April). Support mechanisms to conduct empirical studies in software engineering: a systematic mapping study. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering* (p. 22). ACM, <https://doi.org/10.1145/2745802.2745823>.

1.7 ORGANIZAÇÃO DA TESE

Esta proposta está organizada da seguinte forma:

O Capítulo 2 relata a fundamentação teórica através da síntese dos estudos experimentais utilizados na tese, conceito e fases do planejamento experimental e ameaças à validade.

O Capítulo 3 reproduz a execução de um survey cujo objetivo foi de entender as ações dos autores de experimentos em relação ao processo de controle das ameaças à validade no planejamento dos seus estudos.

O Capítulo 4 apresenta o PrioriTTVs, um processo que se destina a apoiar pesquisadores a priorizar ameaças a validade ao planejar experimentos. O Capítulo 5 detalha a especificação e desenvolvimento da ferramenta web ValidEPlan (*Validity-Driven Software Engineering Experiments Plan Tool*), uma aplicação para o planejamento de experimentos da área de ES, orientada à validade dos resultados.

O Capítulo 6 traz a avaliação experimental do processo proposto, bem como o impacto do uso da ferramenta desenvolvida e, por fim, o Capítulo 7 apresenta as considerações finais e trabalhos futuros desta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

“A verdadeira viagem de descobrimento não consiste em procurar novas paisagens, mas em ter novos olhos”.

Marcel Proust.

2.1 INTRODUÇÃO

Evidentemente, no Brasil, cada vez mais as instituições de ensino e seus planejamentos de formação estratégica têm desempenhado um papel crucial e de resultados na construção de políticas de formulação de conhecimento e de pesquisa científica significativa (NAKAGAWA *et al.*, 2017).

Nas universidades, enquanto fundamento para o desenvolvimento das áreas experimentais, têm sido estimulados programas e metodologias inovadoras e tecnológicas que agregam imensurável valor científico ao desempenho de discentes e ao engajamento de pesquisadores imbuídos na definição de critérios de qualidade e/ou validade para as suas análises de dados, fenômenos de controle e resultados.

Em um ambiente de pesquisa científica complexo, instável e desafiador, muitas vezes, ao pesquisador mais atento e rigoroso quanto aos monitoramentos de resultados, foge ao campo de visão crítica preditores que, a partir da evolução metodológica da realização de experimentos, podem desencadear uma ruptura na sistemática de qualidade de seus resultados, refletindo, dessa forma, em sensíveis ou severos riscos ao agrupamento de impactos que a presença desses fatores podem evocar dentro dos ambientes controlados (WOHLIN *et al.*, 2012).

Determinar a qualidade de um experimento não é algo habitual, uma vez que não há um conceito absoluto do constructo “qualidade” para esse escopo. Porém, autores como Kitchenham *et al.* (2002) e Dieste *et al.* (2011) apresentaram as recomendações onde o conceito de qualidade experimental deve estar relacionado ao viés da pesquisa, aumento da validade (planejamento, execução, generalização e aplicação) e correta interpretação dos resultados (TEIXEIRA; FONSECA; SOARES, 2018).

A validade de um experimento relaciona-se ao nível de confiança, por parte do pesquisador, que se pode definir no desenvolvimento de uma investigação experimental, considerando todo o processo. Refere-se ao quão confiáveis são os

elementos empregados, sua sistematização e arranjos, até a apresentação precisa dos resultados (LIMA; NETO; EMER, 2014).

Este capítulo apresenta uma síntese dos métodos de pesquisas experimentais utilizados nesta tese, assim como, a descrição dos principais conceitos e trabalhos anteriores relevantes, relacionados a esta pesquisa. Por questões de simplificação do texto escrito, a partir deste ponto usaremos os termos “experimento”, “quase-experimento” e “experimento controlado” como sinônimos.

2.2 ESTUDOS EXPERIMENTAIS NA ENGENHARIA DE SOFTWARE

Um método de pesquisa experimental é um conjunto de princípios organizados em torno do qual os dados são coletados e analisados (EASTERBROOK *et al.*, 2008). Estudos experimentais permitem avaliar as atividades abordadas de forma sistemática, quantificável e controlada (WOHLIN *et al.*, 2012). O objetivo desta seção é fornecer uma breve orientação para o entendimento dos métodos de pesquisa experimentais utilizados nesta pesquisa.

2.2.1 Survey

Survey é o método de pesquisa experimental originado das áreas de economia e sociologia. O pesquisador elabora um instrumento de coleta de dados para obter respostas de um conjunto de pessoas ou de uma população (censo). Normalmente, as informações são relacionadas às ações, experiências, comportamentos, opiniões ou perfis das pessoas com relação ao objeto de estudo pesquisado (PIMENTEL e FUKS, 2012).

Para Travassos e Barros (2003), o objetivo do survey é descrever, explicar e explorar informações preliminares (quantitativas e qualitativas) através de um instrumento de coleta, além de identificar variáveis do estudo que serão avaliadas.

De acordo com Easterbrook *et al.* (2008), além do método survey ser mais associado ao uso de questionários, a coleta dos dados também pode ser realizada por meio de entrevistas estruturadas ou dados técnicos. Uma pré-condição para a realização de um survey é uma questão de pesquisa clara sobre a natureza de uma população alvo (representativa da população). Informação adicional sobre este

método pode ser encontrada em (TRAVASSOS; BARROS, 2003; EASTERBROOK *et al.*, 2008; PIMENTEL, 2018).

Esta tese utiliza survey com o objetivo exploratório de entender as atividades realizadas pelos praticantes de experimentos controlados da área de ES em relação ao processo de controle das ameaças à validade no planejamento de seus estudos.

2.2.2 Experimentos em Engenharia de Software

O método experimento tem origem nas ciências naturais, como biologia e física. Normalmente é realizado em laboratório e oferece o maior nível de controle entre os diferentes tipos de métodos experimentais (PIMENTEL, 2018). Escolher o tipo de experimento para uma pesquisa depende dos objetivos do estudo ou dos resultados esperados. As principais características usadas para diferenciar os tipos de experimentos estão relacionadas ao controle de medição e execução, custo e nível de replicabilidade do estudo.

De acordo com Easterbrook *et al.* (2008), um pré-requisito para realizar um experimento é uma hipótese bem definida. A hipótese (e a teoria a partir da qual é desenhada) orientam todas as etapas do projeto experimental, incluindo as variáveis do estudo e como medi-las.

Conceitualmente, de acordo com o entendimento expresso no trabalho de Mafra e Travassos (2006 p. 19) sobre experimento controlado em ES:

[...] é construído e avaliado um modelo que representa ao uso de métodos, de técnicas e das ferramentas (objetos de estudo) durante atividades de desenvolvimento de software. Assim, os participantes selecionados aplicam sob as condições controladas, produzindo os resultados cuja análise permite obter as conclusões a respeito da aplicação. Para avaliar o objeto de estudo em experimento, é preciso isolar os fatores que o influenciam e, de alguma maneira, captar os efeitos obtidos com aplicação. Assim, o ponto de partida é a ideia de uma relação de causa e efeito em teoria.

Experimento (ou experimento controlado) em ES, ainda, é a investigação experimental que manipula um fator ou variável do ambiente estudado. Com base na aleatorização, diferentes tratamentos são aplicados a ou por diferentes sujeitos, mantendo outras variáveis constantes e medindo efeitos nas variáveis de resultado.

Os experimentos apresentam um maior nível de controle da execução e medição, no entanto, o custo da investigação e repetição dos estudos são maiores comparados a outros métodos. Borges *et al.* (2015) identificaram mecanismos de

suporte utilizados como referência para conduzir e reportar experimentos em ES. A tabela abaixo mostra os principais *guidelines* para conduzir e reportar experimentos.

Tabela 1 – *Guidelines* para conduzir e reportar experimentos em ES

Autor/Ano	Objetivo	Fases do Estudo
Wohlin (2012)	Pesquisa Empírica	Todas
Juristo e Moreno (2013)	Experimento Controlado	Todas
Pfleeger (1995)	Experimento	Todas
Jedlitschka (2005)	Experimento Controlado	<i>Report</i>
Ko <i>et al.</i> (2015)	Experimento Controlado	Todas
Kitchenham (2002)	Pesquisa Empírica	Todas

Fonte: Elaboração própria (2019).

Esta tese irá considerar a definição de experimento controlado a partir de Wohlin *et al.* (2012, p. 11), descrito abaixo:

Experimento (ou experimento controlado) em engenharia de software é um método empírico que manipula um fator ou variável do ambiente estudado. Com base em aleatorização, diferentes tratamentos são aplicados a ou por diferentes sujeitos, enquanto outras variáveis se mantêm constantes e medindo os efeitos sobre as variáveis de resultado. Em experimentos orientados a humanos, estes aplicam diferentes tratamentos a objetos, enquanto em experimentos orientados a tecnologia, diferentes tratamentos técnicos são aplicados a diferentes objetos.

Experimentos aleatorizados são caracterizados pela distribuição aleatória dos sujeitos aos grupos experimentais de forma imparcial, com o objetivo de minimizar o viés da pesquisa. Quase-experimentos também possuem tratamentos, medidas de resultados e unidades experimentais, porém, não é feita a distribuição aleatória dos sujeitos pelos tratamentos, nem grupos-controle. Para o escopo deste trabalho, iremos considerar experimentos ou quase-experimentos como objeto de estudo da proposta.

Um experimento controlado pode ser manipulado em laboratório (*off-line*), sob condições controladas ou realizado em campo (*on-line*), no viés real do fenômeno estudado, onde o nível de controle é mais difícil. Wohlin *et al.* (2012) classificam os experimentos em ES em duas categorias: experimentos orientados a seres humanos e experimentos orientados a tecnologia.

Nos experimentos orientados para o ser humano, os seres humanos aplicam tratamentos diferentes aos objetos (com menor nível de controle), enquanto em experimentos orientados para tecnologia, vários tratamentos técnicos são aplicados em diferentes objetos (com maior nível de controle).

Experimentos com a participação de seres humanos possuem maior conotação subjetiva, tornando-se um maior desafio aos experimentadores obter e/ou reportar resultados mais confiáveis e, ao mesmo tempo, com maior rigor de validade. Por este estímulo, somente experimentos orientados a seres humanos serão objetos de estudo para esta pesquisa.

A seguir, algumas das principais particularidades de experimentos em ES:

- Oferecer maior nível de controle do processo, bem como, a possibilidade de replicá-lo;
- Confirmar teorias, explorar os relacionamentos, avaliar a predição dos modelos, bem como, validar medidas;
- Os princípios básicos são aleatorização, controle local e replicação;
- Avaliar a confiabilidade de seus resultados a partir de quatro tipos de validades (interna, externa, de constructo e de conclusão);
- Apresentar fortes aspectos de controle do estudo como rigor estatístico, relacionamento tratamento-resultado, generalização dos resultados a uma população maior e relação entre teoria e prática.

A força de um experimento é investigar em que situações as alegações são verdadeiras, bem como, fornecer um contexto em que certos padrões, métodos e ferramentas são recomendados para uso.

Conduzir um experimento envolve várias etapas diferentes. Juristo e Moreno (2013) organizam as fases de experimento de acordo com as seguintes atividades: definição dos objetivos, design, execução e análise dos resultados. Segundo Wohlin et. al (2012), as etapas de um experimento são divididas em escopo, planejamento, operação, análise e interpretação, apresentação e empacotamento. Pfleeger (2005), apresenta o processo do experimento em seis estágios: concepção, design, preparação, execução, análise e disseminação e tomada de decisões.

Fonseca (2016) elaborou a combinação das fases de experimento controlado em ES a partir de autores mencionados anteriormen (JURISTO; MORENO, 2013; WOHLIN *et al.*, 2012; PFLEEGER, 2005):

- **STUDY DEFINITION:** Descrever os objetivos gerais e motivação de um experimento;
- **PLANNING:** Elaborar o planejamento de cada etapa do experimento, ou seja, é a fase de concepção do plano experimental. Segundo Wohlin *et al.* (2012), nesta fase do experimento os pesquisadores devem definir as hipóteses, variáveis, seleção

dos participantes, design e instrumentação do experimento, escalas de medição, como os dados serão analisados e finalmente a verificação da validade do experimento;

- **PREPARATION:** Produzir os materiais necessários à realização do experimento, de acordo com o plano experimental;

- **EXECUTION:** Realizar o experimento de acordo com o plano experimental com o objetivo de coletar os dados que devem ser analisados;

- **ANALYSIS AND INTERPRETATION:** Analisar e interpretar dados coletados de acordo com o rigor estatístico definido no plano experimental;

- **PRESENTATION AND PACKAGE:** Organizar e reportar corretamente os resultados do experimento aos diferentes públicos.

Para o escopo deste trabalho iremos evidenciar o controle das ameaças à validade durante a fase de Planejamento (*planning*) de um experimento controlado da área de ES.

2.2.3 Experimentos em outras áreas

Evidentemente, em outras áreas, como medicina, educação, psicologia e ciências sociais, a prática experimental existe há muito mais tempo do que em ES. Por isso, essas áreas de conhecimento, principalmente a medicina e as ciências sociais, influenciaram o desenvolvimento experimental em ES.

A partir de pesquisas da área médica, Kitchenham *et al.* (2002) propuseram um roteiro com diretrizes preliminares usadas para realizar pesquisas experimentais em ES. De acordo com um *survey* realizado por Teixeira, Fonseca e Soares (2018), as estratégias mais usadas para a realização de experimentos em Engenharia de Software foram as propostas por Wohlin *et al.* (2012) e Juristo e Moreno (2013).

Estes estudos, utilizados amplamente como referência em experimentos para área de ES, foram baseados em pesquisas da área de ciências sociais como Judd, Hoyle e Harris (2002) e Campbell (1963) e da área de estatística como Montgomery (2017) e Box, Hunter e Stuart (1978).

Especificamente sobre estratégias relacionadas à identificação e priorização de ameaças à validade de experimentos, alguns dos campos de estudo mais maduros do que a ES têm concentrado suas pesquisas na frequência e na gravidade das ameaças, como forma de priorizar qual delas serão mitigadas.

Por exemplo, Henderson *et al.* (2013) sugerem classificar e priorizar maiores riscos de acordo com as ameaças recorrentes na literatura, justificando que guias de medicina afirmam que recomendações recorrentes com maior frequência evidenciam as ameaças à validade prioritárias a serem mitigadas.

Já Singh *et al.* (2003) sugerem estimar as maiores ameaças à segurança do paciente de acordo com métodos de frequência e gravidade de cada um dos riscos, enfatizando que as ações de mitigação devem ser classificadas de acordo com a sua assertividade potencial.

Esta seção não se propõe a esgotar o assunto sobre experimentos em outras áreas de conhecimento, mas apenas fornece uma breve visão geral sobre a temática em áreas distintas da ES, assim como, algumas maneiras de como o tema classificação e priorização de ameaças à validade e suas respectivas ações de mitigação têm sido investigados em outras áreas de pesquisa.

2.3 PLANEJAMENTO EXPERIMENTAL

Um plano experimental é um registro documental em que os experimentadores definem e descrevem passo a passo procedimentos e direcionamentos específicos a serem seguidos na condução e/ou replicação de um experimento, assim como a análise de seus resultados (FONSECA, 2016).

Guidelines como os de Wohlin *et. al* (2012), Pfleeger (1995) e Juristo e Moreno (2013), usados para conduzir e relatar experimentos em ES (Tabela 1) têm visões distintas das etapas do processo experimental. A partir do panorama apresentado por esses autores, Fonseca (2016), agrupou em nove passos as fases de um plano experimental, o qual iremos utilizar para o escopo deste estudo:

- Objetivos;
- Hipóteses, variáveis e medidas;
- Participantes;
- Materiais Experimentais e Tarefas;
- Design Experimental;
- Procedimentos;
- Coleta e Análise dos Dados;
- Ameaças à Validade;
- Estrutura do Documento.

Durante a documentação das informações de cada etapa de um plano experimental, o pesquisador deve identificar as ameaças à validade e definir ações para controlá-las. Com isso, pode-se diminuir a probabilidade de erros sistemáticos o que minimiza o risco de viés e pode possibilitar conclusões corretas e confiáveis. O detalhamento de cada uma das etapas de um plano experimental, será definido na Seção 4.4, onde abordaremos a construção de um Instrumento de Caracterização de um Planejamento Experimental.

2.3.1 Dificuldades no Planejamento Experimental

Ao longo das análises realizadas em estudos localizados na literatura recente, observou-se, de início, que as interferências das principais ameaças à validade dos resultados experimentais em ES são manifestadas, em especial, na fase de execução de experimentos, ficando os pesquisadores responsáveis pela observação das mudanças como um todo, das ações corretivas e mensuração dos resultados definidos a priori.

Imaginava-se, durante esta fase executória, que as ameaças se manifestavam de maneira mais consistente, interrompendo o fluxo de qualidade e validade dos resultados e, ao mesmo tempo, estimulando do pesquisador o olhar dinâmico, a fim de empregar os recursos para se garantir a confiabilidade necessária à consecução de objetivos centrais do experimento (GABRIEL, 2010; FREIRE, 2016; NAKAGAWA, 2017).

Ademais, foram, aos poucos, os experimentadores percebendo que nem sempre os fatores ameaçadores à validade dos experimentos agiram durante a fase executória, mas também eram oriundos de falhas e ausência do método descritivo no próprio planejamento experimental.

Por meio de um planejamento amarrado, bem delineado e consistente, torna-se a identificação dos agentes de ameaça à validade dos resultados experimentais mais facilitada, repercutindo, assim, numa melhor interpretação do especialista e dos demais condutores da pesquisa, durante a execução do estudo.

Sem uma atuação eficiente no planejamento do experimento, com inclusão dos fatores que podem colocar em risco os resultados experimentais, dificilmente o pesquisador direciona seu enfoque para outras questões, senão as triviais ameaças

já discutidas no viés da literatura científica, que também permeiam o contexto deste trabalho.

Um dos principais obstáculos quanto ao planejamento de um experimento da área de ES é prever os possíveis riscos na execução que possam causar vieses no resultado trazendo em muitas vezes resultados irreais ou até mesmo inconclusivos (SIEGMUND; SIEGMUND; APEL, 2015). Uma das soluções, pois, para aumentar a confiabilidade nos resultados dos experimentos é identificar preventivamente os principais riscos que possam surgir e se antecipar a eles através do planejamento de ações de mitigação. No entanto, as atividades de identificação, análise e priorização a ameaças à validade não são atividades simples.

Um dos motivos da complexidade se deve ao fato do grande número de possibilidades de problemas que podem acontecer desde planejamento, execução, análise e divulgação da avaliação experimental, além de que o tratamento de uma ameaça pode implicar, dessa maneira, no surgimento de novas ameaças à validade, não identificadas inicialmente (TEIXEIRA; FONSECA; SOARES, 2018).

Por exemplo, para minimizar efeitos de fadiga e cansaço durante a realização de um experimento (ameaça interna), pode-se inserir intervalos entre as execuções do experimento (ação de controle). Contudo, durante as lacunas inseridas ao longo do experimento, pode haver comunicação entre os participantes, surgindo uma nova ameaça à validade do estudo.

2.4 AMEAÇAS À VALIDADE

Uma questão fundamental sobre os resultados de um experimento é quão válidos são seus resultados. Ameaças à validade são potenciais riscos que podem surgir desde a fase de planejamento até a divulgação dos resultados dos estudos experimentais (WOHLIN et al, 2012).

As ameaças à validade são preditores dos quais não se devem abreviar, nem deixar de analisar os aspectos e contribuições para desenvolvimento do experimento, já que seus efeitos nocivos podem ser evitados desde o planejamento de ações, da sistematização dos procedimentos de análise e medidas para saneamento (leia-se medida corretiva adequada) em cada caso (MAFRA; TRAVASSOS, 2006; TEIXEIRA; FONSECA; SOARES, 2018).

Assim, para que se resulte em eficácia e efetividade posterior do estudo

experimental, torna-se necessária a identificação quanto às ameaças à validade e confiabilidade, a fim de compreender os seus meandros no estudo, suas influências e comportamentos que resultem em perdas de valor, significado e síntese dos resultados.

Dias-Neto *et al.* (2004), entretanto, alertam para que se tome cuidado no sentido de não confundir validade com confiabilidade, já que são termos distintos entre si. Em termos conceituais, a confiabilidade é, pois, a aplicação repetida de técnicas, a partir do uso de métodos que implicam em resultados coerentes, observada a partir da comparação de resultados. Por sua vez, a validade está atrelada à precisão dos dados, definidos a priori, demonstrando a certeza dos procedimentos, assim como, à existência de efeitos não controlados e a capacidade de generalização.

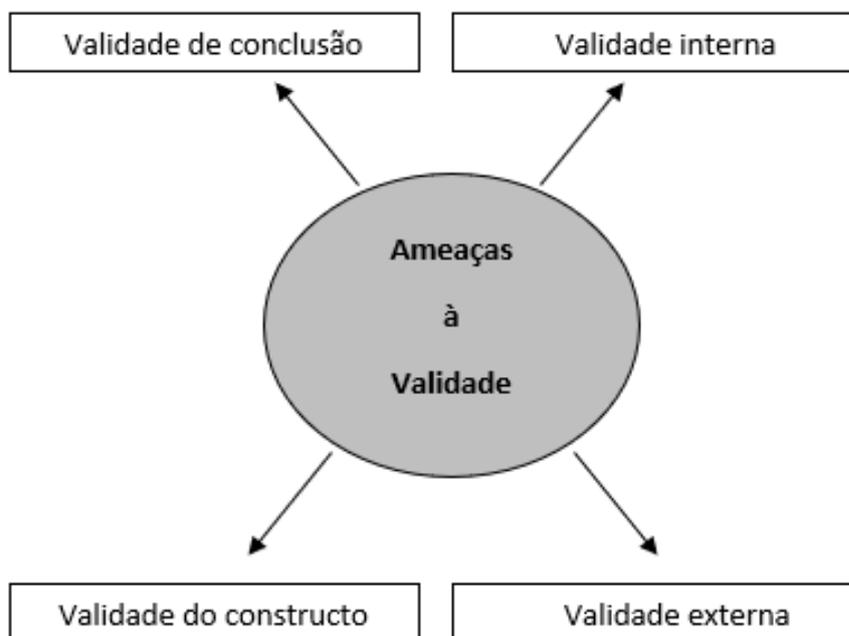
A atividade de identificação das variáveis de risco à pesquisa é, antes de tudo, prioritária para a consecução de objetivos e dos resultados pré-definidos, buscando atingir maior segurança e a coerência nas informações, sendo também uma vertente fundamental que deve ser desempenhada ainda no planejamento (DIAS-NETO, 2007).

A literatura científica tem cada vez mais discutido e aprofundado, embora ainda muito timidamente quanto aos avanços contínuos do desenvolvimento da ES, os critérios de definição e a classificação (taxonomia) das ameaças à validade dos experimentos, de modo a evidenciar um conjunto de variáveis que possam implicar em ruptura na qualidade dos resultados das investigações experimentais (TEIXEIRA; FONSECA; SOARES, 2018).

Como princípio da identificação, a literatura tem também auxiliado no que compete à acessibilidade das ameaças à validade aos pesquisadores diversos e aplicáveis a várias categorias experimentais. Para isso, a forma resumida de exposição dos fatores de risco aos experimentos, conhecida como *checklist*, é a abordagem mais aceita e padronizada, a fim de familiarizar os especialistas e outros agentes participantes dos procedimentos experimentais acerca da ação, atuação e interferências no projeto conduzido (TEIXEIRA; FONSECA; SOARES, 2018).

Segundo Wohlin *et al.* (2012), é importante considerar a questão das ameaças à validade desde a etapa de planejamento do experimento para que possamos ter resultados válidos para a população de interesse. Conforme entendimento de Cook e Campbell (2002), ameaças à validade e confiabilidade de experimentos são, assim, classificadas como mostra a Figura 1:

Figura 1 – Classificação das ameaças à validade dos experimentos



Fonte: Cook e Campbell (1979).

Cada um dos quatro tipos de validade, apresentados por Cook e Campbell (1979), está relacionado a uma questão metodológica na experimentação.

A validade de conclusão está relacionada à capacidade de tirar conclusões corretas sobre as relações entre os tratamentos usados no estudo e os resultados obtidos. Por exemplo, termos a certeza do efeito de um tratamento através da significância estatística, ou seja, quando o p-valor observado é menor que o nível de significância definido para o estudo.

Segundo Wohlin *et al.* (2012), alguns riscos estão relacionados com a validade de conclusão, são eles: baixo poder estatístico, violação de hipóteses de testes estatísticos, confiabilidade de medidas, confiabilidade de implementação de tratamento e heterogeneidade aleatória dos sujeitos.

A validade interna, verifica se há algum relacionamento causal entre tratamento e o resultado obtido, sem influência de outro fator que não é controlado ou medido. Em outras palavras, está preocupada com a capacidade de repetir o comportamento atual de estudo com os mesmos participantes e objetos para o qual foi executado.

São exemplos de ameaças à validade interna: efeitos de aprendizagem, risco de aplicação do histórico de tratamento que afeta os resultados, diferenças entre participantes relacionadas ao nível de habilidade, efeitos de fadiga e cansaço,

diferenças nos materiais utilizados e comunicação entre participantes.

A validade do constructo está relacionada ao grau em que as variáveis usadas no estudo conseguem medir com exatidão os conceitos teóricos que se quer medir (HENDERSON *et al.*, 2013). De acordo com Wohlin *et al.* (2012), algumas ameaças à validade de constructo referem-se ao design do experimento, outros a fatores sociais.

Por fim, a validade externa está relacionada à capacidade de repetir o mesmo comportamento da pesquisa em outros grupos de participantes, além daqueles em que o estudo foi aplicado, ou seja, relaciona-se à capacidade de generalização dos resultados. São exemplos práticos de ameaças à validade externa: representatividade dos participantes, representatividade dos artefatos, representatividade do tratamento comparativo e representatividade do ambiente operacional.

2.4.1 Priorização das Ameaças à Validade

Existem conflitos entre alguns dos tipos de ameaças à validade. Por exemplo, ao tentar priorizar um tipo de validade, outro tipo pode ser afetado (WOHLIN *et al.*, 2012). A ordem de prioridade entre os tipos de ameaças à validade pode ser definida em virtude dos propósitos de cada estudo experimental (TRAVASSOS *et al.*, 2002). Há, também, discordâncias em virtude dos relacionamentos que existem entre as ameaças à validade e suas respectivas ações de controle, o que faz com que ao tentar controlar uma determinada ameaça, outra não identificada inicialmente poderá surgir (NETO; CONTE, 2013).

Priorizar algumas ameaças à validade do experimento permite que os pesquisadores possam identificar os maiores riscos à confiabilidade dos resultados do estudo, planejando ações de controle mais assertivas.

A tarefa de priorizar determinadas ameaças à validade é definida em razão do propósito de cada estudo e não tem um processo determinístico. Dessa forma, o pesquisador terá um papel importante nessa decisão, pois é ele quem define os objetivos do estudo. Dependendo dos objetivos, algumas ameaças à validade são mais importantes do que outras (ANDERSSON *et al.*, 2003).

O trabalho propõe um processo, cujo fim é melhorar a qualidade dos resultados de experimentos controlados em ES por meio da identificação, da classificação e da priorização de ameaças à validade e suas respectivas ações de mitigação.

2.5 RESUMO

Este capítulo apresentou visão geral de qualidade e priorização de ameaças à validade no âmbito do planejamento de experimentos da área de ES. Evidenciamos, também, uma síntese dos métodos survey e experimentos utilizados nesta tese, assim como, a descrição dos principais conceitos e trabalhos anteriores relevantes.

3 SURVEY – AS AMEAÇAS À VALIDADE EM EXPERIMENTOS CONTROLADOS: O QUE OS *EXPERTS* DIZEM E POR QUE ISSO É RELEVANTE?

“A pesquisa básica é como atirar uma flecha para o ar e, onde ela cair, pintar um alvo”.

Homer Adkins Burton.

3.1 INTRODUÇÃO

Survey Research é um método de pesquisa considerado misto, pois inclui uma combinação equilibrada de atributos qualitativos e também quantitativos. Segundo Easterbrook *et al.* (2008), é considerado um dos cinco métodos de pesquisa experimental mais relevantes para Engenharia de Software (ES). Para esta pesquisa de doutorado, o *survey* foi utilizado como uma estratégia experimental preliminar de cunho exploratório e motivacional.

Realizou-se o *survey* com 115 experimentadores que planejaram e executaram experimentos controlados em suas pesquisas. O objetivo foi descrever e entender as ações dos pesquisadores ao realizar os procedimentos para controle das ameaças à validade; e verificar a percepção entre práticas reais de tratamento de ameaças à validade e as orientações existentes na literatura.

A taxonomia utilizada como parâmetro para descrição da estratégia do *survey* e definição do método e tipo de pesquisa em quantitativo e qualitativo, baseou-se respectivamente nos trabalhos de Juristo e Moreno (2013) e de Travassos e Barros (2003). Para condução deste *survey*, seguiram-se os guias propostos por Pfleeger (1995).

3.2 OBJETIVOS E DEFINIÇÕES

O objetivo geral do *survey* foi de entender as ações dos autores de experimentos da área de ES em relação ao processo de controle das ameaças à validade no planejamento dos estudos. Para isto, definiram-se as seguintes questões de pesquisa:

- Quais são os procedimentos realizados por pesquisadores para identificar e

priorizar as ameaças à validade ao planejar experimentos controlados em ES?

- Qual a importância desse processo?

Estruturamos os objetivos deste *survey* segundo a abordagem GQM (*Goal, Question, and Metric*) (VAN SOLINGEN *et al.*, 2002, p. 49), descrito a seguir:

Analisar a atividade de identificação de ameaças à validade do planejamento de experimentos controlados da área de ES. **Com propósito** de compreender os procedimentos para controle de ameaças à validade. **Com respeito** às estratégias utilizadas, importância e finalidade do processo e direcionamentos em prol da melhoria da validade dos experimentos. **Sob o ponto de vista** dos autores de experimentos controlados. **No contexto** dos pesquisadores que publicam na área de ES e na Engenharia de Software Experimental (ESE).

3.3 AMOSTRA DO ESTUDO

O critério para participar deste estudo foi o participante do *survey* ser um dos autores de pelo menos um experimento em conferências e periódicos científicos da área de ES e ESE. A seleção da amostra foi realizada através da inclusão de todos os artigos completos (*full papers*) dos seguintes locais:

- *Empirical Software Engineering Journal (ESEJ)*;
- *International Conference on Empirical Software Engineering and Measurement (EMSE)*;
- *International Conference on Empirical Assessment & Evaluation in Software Engineering (EASE)*;
- *International Conference on Software Engineering (ICSE)*;
- *Journal of Systems and Software (JSS)*;
- *Transactions in Software Engineering (TSE)*.

Estes locais foram escolhidos porque são considerados relevantes periódicos e conferências da comunidade de ES e ESE (FALCÃO *et al.*, 2015). Nesta fase, nenhuma busca automática foi necessária porque o processo de busca envolveu apenas os artigos anteriormente selecionados por Borges *et al.* (2015) e Falcão *et al.* (2015). Os estudos foram coletados através de bibliotecas digitais de artigos científicos, por exemplo, *IEEEExplore*, *ACM*, *Springer Link*, *Science Direct*. A pesquisa abrangeu os seguintes períodos de publicação das conferências e periódicos selecionados: TSE (2003 a 2013) e ESEM, EASE, ESEJ, ICSE, TSE (2003 a 2017).

Os critérios de exclusão adotados são descritos a seguir: (1) artigos curtos; (2)

artigos não técnicos; (3) artigos duplicados e, (4) artigos que não se enquadram na definição de experimento controlado definida neste estudo. Entre os procedimentos para seleção dos estudos, considerou-se o título, resumo, palavras chave, método da pesquisa e conclusão do artigo.

Tabela 2 – Estudos que reportaram experimentos

Periódico/ Conferência	Artigos que reportaram experimentos		
	Total de artigos investigados	Experimentos	%
ESEM	421	77	18,28
EASE	249	28	11,24
ESEJ	483	51	10,55
ICSE	1077	43	3,99
TSE	335	45	13,43
JSS	1778	41	2,30
Total	4.343	285	6,56

Fonte: Elaboração própria (2019).

A Tabela 2 resume em números os estudos investigados e quais de fato relataram experimentos da área de ES por local de publicação. O local de publicação com maior percentual de experimentos é o ESEM com 18,28% dos 421 artigos investigados. Isso, provavelmente, se deve ao fato de que o ESEM é uma das conferências mais relevantes na área de Engenharia de Software Empírica. A análise dos 4.343 artigos candidatos resultou na seleção de 285 estudos que reportaram experimentos, ou seja, 6,56% do total de artigos analisados.

3.4 INSTRUMENTO DE COLETA DE DADOS

Uma revisão *ad hoc* da literatura foi realizada no intuito de identificar potenciais instrumentos de coleta de dados relacionados à problemática abordada neste *survey*. No entanto, não foram encontrados estudos apropriados a serem aplicados.

Em razão disto, desenvolveu-se um questionário contendo 12 questões objetivas com alternativas de respostas pré-fixadas e mais uma questão discursiva. A construção dos enunciados do questionário se deu a partir de uma revisão *ad hoc* da literatura da área de ESE, levando principalmente em consideração os objetivos e definições deste *survey*.

O instrumento de coleta de dados foi enviado aos participantes por meio de endereços de *e-mail* obtidos através dos artigos extraídos das bibliotecas digitais. A

escolha desse método deu-se por diversas razões (OPPENHEIM, 2001), incluindo:

- A população alvo da pesquisa participa de comunidades conectadas através da tecnologia;
- Permite a aplicação não supervisionada do questionário;
- Baixo custo financeiro com a coleta de dados;
- Facilidade e rapidez quanto ao envio e retorno de respostas;
- Gerenciamento automatizado do processo de coleta.

Os *e-mails* enviados aos participantes do *survey* continham um formulário elaborado através da ferramenta *Google Forms*, utilizada por ser de fácil usabilidade para esse tipo de pesquisa. A primeira parte do formulário foi responsável por coletar dados demográficos dos participantes como país de origem, instituição de ensino e/ou trabalho e posição atual. A segunda parte teve por objetivo identificar as características do processo de controle de ameaças à validade de experimentos da área de ES.

Foram definidos os conceitos dos termos técnicos e evitou-se (quando possível) o uso de siglas. O instrumento original de coleta de dados (em inglês), bem como os dados brutos do *survey* encontra-se disponível em <https://bit.ly/2HnJzfr>.

3.5 ESTUDO PILOTO

Uma avaliação prévia foi conduzida com o objetivo de identificar erros, realizar refinamentos necessários e examinar a clareza e compreensão das perguntas do questionário. Para isso, quatro pesquisadores foram convidados por conveniência, sendo dois professores doutores e dois alunos de doutorado, ambos do programa de pós-graduação do Centro de Informática (CIn) da UFPE. Eles foram selecionados considerando experiência anterior com o método *survey* e disponibilidade para participar do estudo piloto.

As observações extraídas dessa avaliação trouxeram as seguintes melhorias ao instrumento de coleta de dados:

- Identificação de erros tipográficos e ambiguidades textuais;
- Indicação de uma melhor sequência dos enunciados;
- Padronização de alguns formatos de resposta;
- Identificação de perguntas desnecessárias ao objetivo do *survey*;

- Estimativa de tempo médio necessário para preenchimento do questionário.

3.6 EXECUÇÃO

A aplicação do instrumento de coleta de dados foi realizada de forma não supervisionada, ou seja, os participantes responderam, sem suporte por parte dos pesquisadores envolvidos na pesquisa.

No entanto, o questionário aplicado possuía todas as instruções necessárias para o seu correto preenchimento e, em caso de dúvidas sobre o preenchimento, os participantes tiveram livre acesso a questionamentos aos pesquisadores envolvidos na pesquisa através de *e-mail*.

Com o objetivo de motivar os entrevistados a participarem do *survey* a fim de garantir uma taxa de resposta aceitável ao estudo foram definidas as seguintes estratégias ao inserir as informações no corpo do *e-mail*:

- Nome personalizado de cada autor pesquisado, título, local e também o ano de publicação de seu artigo;
- Objetivos claros sobre a contribuição do estudo;
- Informações sobre a preservação da confidencialidade dos participantes;
- Especificação do número de questões do formulário;
- Informação do tempo médio necessário para responder os questionamentos.

Além disso, por ter integrantes de diversos continentes, os envios dos *e-mails* foram segmentados com base no fuso horário dos participantes, evitando (quando possível) envios fora do horário regular de trabalho.

Com o objetivo de acompanhar a aplicação do *survey*, criou-se uma planilha eletrônica com os dados agrupados da seleção da amostra. Os contatos foram classificados por nome, título, local e ano de publicação do artigo, continente de origem do e-mail, além das informações dos envios como dia e hora, lembretes, feedbacks e status. O estudo foi realizado em dois períodos, maio a junho de 2017 e fevereiro a março de 2018.

3.7 ANÁLISE DOS DADOS

Nesta seção, as principais questões que envolveram a análise dos dados deste

survey são discutidas. Antes de realizar a análise propriamente dita, verificou-se problemas relacionados a inconsistência e completude dos dados, assim como conversões ou ajustes dos dados em escalas do tipo nominal, ordinal ou escalar.

Para realizar a análise estatística dos dados e discussão dos resultados utilizou-se o pacote de software estatístico IBM SPSS (SPSS, 2011) e significância estatística de $\alpha = 0,05$. O SPSS permite realizar uma variedade de análises descritivas, inferenciais e também multivariadas.

Os dados do *survey*, com a devida preservação da confidencialidade dos participantes, estão disponíveis em <https://bit.ly/2HnJzfr>. A descrição quantitativa dos dados foi realizada a partir da distribuição de frequência das variáveis nominais, de ordinais e intervalares dos seguintes questionamentos:

- Qual o seu país de origem?
- Qual a sua instituição de Ensino e/ou Trabalho?
- Relacionado à educação. Qual a sua posição atual? (Estudante de Mestrado; Mestre; Estudante de Doutorado; Doutor; Outro?);
- Relacionado a atividade profissional. Qual a sua posição atual? (Professor; Pesquisador; Pesquisador Sênior; Profissional da indústria de ES; Outro);
- Quantos experimentos da área de Engenharia de Software você participou do planejamento? (Somente um experimento; dois a cinco experimentos; seis a nove experimentos; 10 ou mais experimentos);
- Quantos experimentos da área de Engenharia de Software você participou da execução? (Somente um experimento; dois a cinco experimentos; seis a nove experimentos; 10 ou mais experimentos);
- Você considera importante identificar as ameaças à validade de experimentos controlados da área de Engenharia de Software? (Não é importante; Pouco Importante; Importante; Muito importante; Extremamente importante);
- Você considera importante analisar como as ameaças à validade afetam o plano experimental do experimento? (Não é importante; Pouco Importante; Importante; Muito importante; Extremamente importante);
- Por que identificar as ameaças à validade de um experimento? (Divulgar para a comunidade as possíveis limitações do experimento; Para ajustar o plano experimental do experimento; Para ajustar a execução do experimento; Eu não identifico ameaças à validade; Outro);
- Qual é o melhor momento para identificar as ameaças à validade? (Durante o

planejamento experimental; Após o planejamento e antes da execução do experimento; Após a execução do experimento e antes da análise dos dados; Somente quando irei divulgar o experimento; Não identifico ameaças à validade; Outro);

- Você já utilizou guias, listas de verificação, ferramentas de software ou outra estratégia para identificar e/ou controlar as ameaças à validade de experimentos? Se sim, reporte a(s) referência(s) bibliográfica(s) utilizada(s);

- Segundo Wohlin *et al.* (2012), há quatro principais tipos de ameaças à validade: a validade de conclusão, a validade interna, a validade de construção e a validade externa. Em sua opinião, há uma ordem de importância entre estes tipos de validade apresentados por esse autor? Se sim, qual a ordem de prioridade? (Não existe diferença de prioridade entre os tipos de validade; Sim – existe uma ordem de prioridade entre os tipos de validade; Desconheço essa temática).

O procedimento utilizado para analisar questões discursivas foi a análise qualitativa, a partir de procedimentos sistemáticos e objetivos de descrição e codificação aberta da teoria fundamentada nos dados (*Grounded Theory*) (CORBIN; STRAUSS, 2014). Como o objetivo deste trabalho não foi criar uma teoria, a etapa de codificação seletiva da teoria fundamentada dos dados não foi realizada.

Além disso, as etapas da análise de conteúdo definidas por Bardin (2009) também foram seguidas para um melhor tratamento das informações, são elas, (i) análise prévia, (ii) exploração de conteúdo, (iii) tratamento dos resultados, inferência e interpretação. Seguem as questões discursivas analisadas:

- Ao aplicar uma ação para controlar um determinado tipo de validade, outras ameaças podem ser inseridas na pesquisa. Você se preocupa com esses relacionamentos entre as ameaças à validade e as ações de controle? Sim? Por que? | Não? Por que?

- Você tem algum comentário adicional a esta pesquisa, aos questionamentos ou relacionado ao estudo em geral?

3.8 RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados e as discussões sobre a análise realizada nos dados coletados neste estudo.

3.8.1 Análise Quantitativa

Foram enviados 748 *e-mails*, dos quais 102 envios foram rejeitados pelos servidores de domínio, classificadas como endereços inexistentes. Com isto, 646 *e-mails* foram enviados e entregues com sucesso.

Dos 646 *e-mails* enviados com sucesso, obteve-se a taxa de resposta de 115 participantes, o que equivale a 17,8% do total de mensagens recebidas pelos autores. É importante esclarecer que foram enviados *e-mails* para todos os autores de cada artigo e também foram utilizadas outras fontes de busca (bibliotecas digitais, sites pessoais e institucionais) para os endereços de *e-mails* retornados, muitos destes, foram de autores de artigos que mudaram de instituição.

A avaliação inicial da população-alvo deve surgir dos propósitos da pesquisa, e não da sensação de quem está disponível para responder às nossas perguntas (KITCHENHAM; PFLEEGER, 2003). A amostra obtida não é suficiente para afirmar que respostas reais são representativas de toda a população e conseqüentemente generalizável, mas é um bom ponto de partida, visto ser um *survey* com o objetivo exploratório e os seus dados não foram usados para inferência estatística onde há comparação de médias estatisticamente significativas.

Ao total, autores de 30 países responderam o *e-mail* enviado, destes, sete países do continente americano, cinco países asiáticos, 17 países europeus e um da Oceania. Os dados coletados mostram que há um total de 115 participantes de diferentes instituições de ensino e/ou trabalho.

Tabela 3 – Características dos participantes

Educação	N	Atividade Profissional						
		Professor		Pesquisador Sênior		Outros		
		%		%		%		
Doutorandos	7	0		0		100		
Doutores	108	37%		30%		33%		
Total	115	35%		28%		37%		
Experiência com								
Educação	Planejamento de Experimentos (%)				Execução de Experimentos (%)			
	1	2 a 5	6 a 9	>=10	1	2 a 5	6 a 9	>=10
Doutorandos	28,5	57,1	14,2	0	28,57	57,14	14,28	0
Doutores	2	24,78	18,45	54,6	2,7	34,45	21,21	40,79
Total	3,4	27,8	18,2	50,4	4,3	37,3	20,8	36,5

Fonte: Elaboração própria (2019).

A Tabela 3 mostra que 94% do total de participantes são doutores e 6% são doutorandos. É possível, ainda, identificar que 63% do total dos participantes declararam ser professores ou pesquisadores sêniores.

Aos participantes, foi sugerido selecionar sua posição mais relevante. Em relação a experiência com planejamento e execução de experimentos da área de ES 50,4% afirmaram que já participaram do planejamento de 10 ou mais experimentos e 36,5% que já participaram da execução de 10 ou mais experimentos.

A amostra tem participantes de 30 países em quatro diferentes continentes, sendo que a maioria dos entrevistados declararam ser professores ou pesquisadores sêniores com doutorado e com experiência no planejamento e execução de 10 ou mais experimentos da área de engenharia de software. Dessa forma, aumentando a confiabilidade dos resultados e conclusões deste *survey*.

Tabela 4 – Importância das ameaças à validade

Importância da	EI	MI	I	PI	NI
Identificação das ameaças	78,26%	15,65%	3,48%	1,74%	0,87%
Análise durante o plano experimental	50,43%	33,04%	10,43%	4,35%	1,74%

Fonte: Elaboração própria (2019).

Legenda:

EI: Extremamente importante

MI: Muito importante

I: Importante

PI: Pouco importante

NI: Não importante

A Tabela 4 manifesta a importância do tema em questão na visão dos participantes. Observa-se que, 78,26% consideraram extremamente importante identificar ameaças à validade de experimentos da área de ES. A Tabela 4 reporta a importância de analisar previamente como as ameaças à validade podem afetar o experimento. 50,43% dos participantes consideraram extremamente importante analisar de forma preliminar as ameaças à validade dos experimentos.

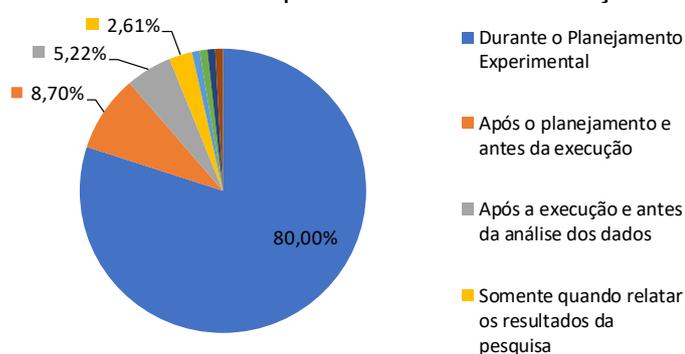
Realizar uma análise prévia das ameaças à validade implica em afirmar que o processo de identificação e controle das ameaças poderá melhorar o planejamento do experimento visto que o pesquisador poderá refazer ou mesmo corrigir as etapas do planejamento como redefinir objetivos, hipóteses, materiais e também as tarefas experimentais, procedimentos e até mesmo o design do experimento.

Figura 2 – Por que identificar as ameaças à validade de um experimento?



Fonte: Dados da pesquisa (2019).

Figura 3 – Melhor momento para identificar as ameaças à validade

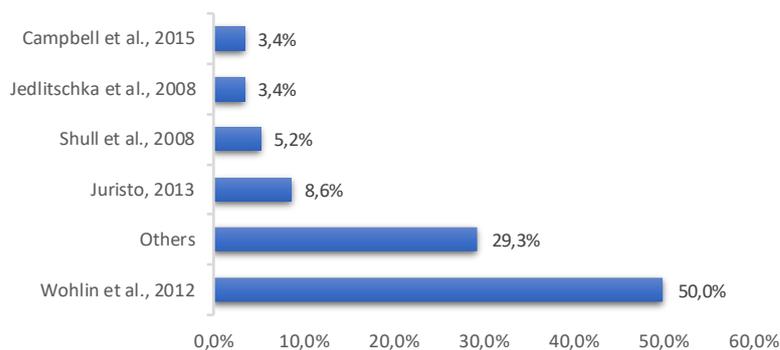


Fonte: Dados da pesquisa (2019).

A Figura 2 e Figura 3 mostram a visão dos participantes em relação ao porquê e quando deve-se identificar as ameaças à validade de experimentos da área de Engenharia de Software.

A maioria dos entrevistados declararam que identificam as ameaças à validade para: (i) possíveis limitações do experimento; (ii) ajustar o plano experimental e; (iii) ajustar a execução do experimento. 80% afirmaram que o melhor momento para identificar ameaças é durante o planejamento do experimento.

Figura 4 – Estratégias para identificar e controlar as ameaças à validade



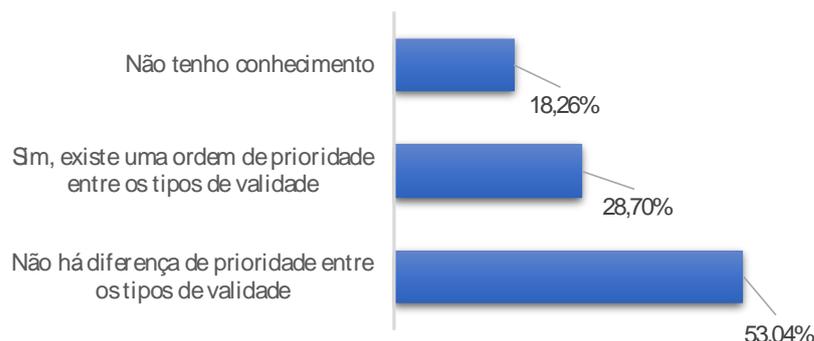
Fonte: Dados da pesquisa (2019).

Ao perguntar aos participantes se os mesmos usavam guias, ou mesmo listas de verificação, ferramentas de software, ou outras estratégias para identificar e controlar ameaças à validade de experimentos, 62,11% responderam que sim. Destes, 50% afirmaram que utilizaram como suporte o livro “*Experimentation in software engineering*” do autor Wohlin *et al.* (2012).

O item da tabela “Others” (29%), agrupa casos em que a estratégia/referência foi citada apenas uma vez. Os demais participantes informaram que utilizaram outras referências conforme mostra a Figura 4. Os dados mostrados pela Figura 4 reforçam a ideia de que as estratégias utilizadas para identificação e controle de ameaças à validade mencionadas na literatura pelos autores, podem não explorar mecanismos que fornecem apoio ao experimentador no sentido de ajudá-lo a priorizar as ameaças mais graves e com isso, deixar de implementar ações de controle mais concretas. Estes dados nos levam a acreditar que os pesquisadores/experimentadores podem não executar uma atividade de priorização de ameaças e ações de mitigação em seus experimentos.

A principal estratégia de suporte apontada pelos participantes é um guia para experimentos que contém um *checklist* com sete ameaças à validade de conclusão, 13 ameaças à validade interna, dez ameaças à validade de constructo e três ameaças à validade externa (WOHLIN *et al.*, 2012).

Figura 5 – Há uma ordem de importância entre tipos de validade?

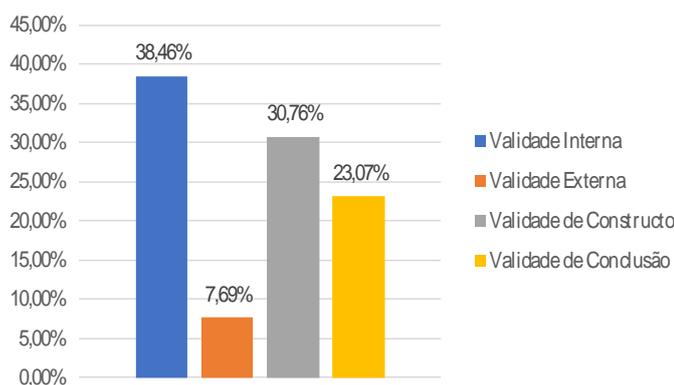


Fonte: Dados da pesquisa (2019).

Ao questionar os participantes se existe uma ordem de importância entre os tipos de ameaças à validade, 53,04% declararam que não há diferença de prioridade entre os tipos de ameaças apresentados, 28,70% disseram que sim, existe ordem de importância entre elas e 18,26% afirmaram desconhecer o assunto, conforme foi apresentado na Figura 5.

As informações relatadas na Figura 5 revelam certa preocupação, pois as orientações existentes na literatura mostram que existe sim uma ordem de prioridade entre tipos de validade em razão do propósito de cada experimento (ANDERSSON *et al.*, 2003; TRAVASSOS; BARROS, 2003; WOHLIN *et al.*, 2012). Tal desconhecimento pode levar a decisões equivocadas ao priorizar uma ameaça em razão de outra.

Figura 6 – Qual a ordem de importância entre os tipos de validade?

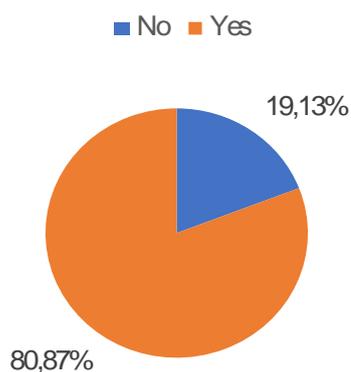


Fonte: Dados da pesquisa (2019).

Entre os 28,70% dos participantes que afirmaram haver ordem de prioridade, eles admitiram existir 13 diferentes combinações entre os tipos de validade (interna, externa, constructo e conclusão). A Figura 6 mostra que 38,46% definiram como maiores prioridades as seguintes opções: validade interna, 7,69% validade externa, 30,76% validade de constructo e 23,07% a validade de conclusão.

É importante ressaltar que apenas 15,15% desses participantes fizeram comentários adicionais, afirmando que dependendo da finalidade do experimento, diferentes tipos de validade recebem prioridades diferentes e que não podem impor uma classificação definitiva entre combinações apresentadas. Essa evidência mostra uma lacuna existente entre orientações existentes na literatura e as práticas reais observadas através dos dados coletados por este *survey*.

Figura 7 – Preocupação com relacionamentos entre ameaças e ações de controle



Fonte: Dados da pesquisa (2019).

A Figura 7 mostra que 80,87% dos participantes disseram preocupar-se com novas ameaças que podem ser inseridas na pesquisa causadas em virtude dos relacionamentos que existem entre as ameaças à validade e possíveis ações de controle. Esses dados revelam a relevância do tema, considerando a importância da aplicação de determinadas ações de mitigação e das consequências ao aplicá-las.

3.8.2 Análise Qualitativa

Os dados foram analisados pelo autor desta tese, supervisionado por dois revisores, orientadores desta pesquisa. Como os questionários foram respondidos em inglês, foi feita tradução livre das citações dos especialistas e, posteriormente, verificadas por outro pesquisador. Uma planilha eletrônica foi usada para facilitar a codificação do conteúdo e análise das questões discursivas.

Na etapa de análise prévia, foi realizada a primeira leitura das respostas ao questionário no intuito de organizar os dados que posteriormente foram analisados. Na fase de exploração de conteúdo, os textos foram estudados com uma maior profundidade no intuito de definir as unidades de registro e unidade de contexto. Nessa etapa, encontraram-se dez temas (unidades de registro) e em seguida

agruparam-se em três categorias de análise (unidades de contexto).

Na etapa de tratamento e interpretação dos resultados, os dados foram sintetizados e analisados de acordo com os eixos temáticos, ligados a uma das três categorias de análise. A Tabela 5 apresenta as categorias, temas identificados e as frequências das citações analisadas.

As evidências ligadas à preocupação de participantes a respeito do surgimento de novas ameaças a partir das relações entre as ameaças à validade e as suas respectivas ações de controle, mostram que houve consideráveis manifestações positivas e negativas.

Tabela 5 – Categorias de Análise

Categorias de Análise	Eixos Temáticos	% Freq.
Preocupação com o surgimento de novas ameaças: Aspectos positivos	Considerações sobre o objetivo do experimento	9,36
	Aspectos relacionados à priorização de ameaças	40,43
	Reflexões sobre a importância da temática	39,16
	Referências à generalização do estudo	5,76
Preocupação com o surgimento de novas ameaças: Aspectos negativos	Ponderações a respeito da irrelevância do processo	20,20
	Dificuldades de implementação do processo	34,36
	Alegações de indiferença ao problema	16,03
	Falta de informação para realizar o processo	14,21
Comentários gerais e adicionais à pesquisa	Sugestões ao processo e/ou pesquisa	39,73
	Considerações sobre a relevância do processo e/ou pesquisa	51,67

Fonte: Elaboração própria (2019).

Relacionado aos aspectos positivos, os principais argumentos dizem respeito à priorização de ameaças (Freq. 40,43%). Os especialistas justificam que não é possível mitigar todas as ameaças e por isso é necessário priorizá-las de acordo com o objetivo a ser alcançado pelo experimento, causando assim reflexões sobre a importância da temática (Freq. 39,16%). Alguns participantes enfatizaram o fato de que novas ameaças podem oferecer os maiores riscos que as originais e que a comunidade de ESE precisa estar ciente e preparada para lidar com a problemática.

Relacionado a aspectos negativos os principais argumentos estão relacionados às dificuldades de implementação do processo de identificação e priorização de ameaças à validade (Freq. 34,36%) e algumas ponderações a respeito da irrelevância desta etapa considerada por alguns pesquisadores (Freq. 20,20%).

De acordo com as citações de alguns dos participantes, a quantidade de fatores incertos como por exemplo fatores humanos, domínio de aplicação e natureza do domínio está relacionada diretamente às dificuldades de implementação do processo de ameaças. Já o enfraquecimento do uso do método de pesquisa experimento controlado pode ser resultado, também, da irrelevância dada ao processo de controle de ameaças e ações de mitigação consideradas por alguns experimentadores.

As principais evidências relacionadas aos comentários gerais e adicionais a esta pesquisa mostram que houve manifestações como sugestões sobre o processo de ameaças à validade (Freq. 39,73%) e considerações sobre a importância do processo e/ou pesquisa (Freq. 51,67%).

As principais sugestões sobre o processo de ameaças à validade citadas pelos participantes foram: utilizar os testes pilotos durante o planejamento experimental, verificar a abordagem em outros campos de pesquisa e usar os exemplos práticos, comuns e significativos sobre a priorização de ameaças.

Relacionado à relevância da temática abordada no estudo, os especialistas mostram-se preocupados com a forma com que os pesquisadores vêm tratando essa temática. Um dos participantes salientou: *"As experiências que não tratam as ameaças à validade são inúteis, porque podem sustentar conclusões falsas, fazendo-nos passar por ignorantes e errados"*.

Em geral, a maioria dos participantes enfatizaram o fato de que pouco se sabe e falha-se muito sobre o controle de ameaças à validade de nossos experimentos. Os especialistas também levantaram a questão de que muitos pesquisadores continuam a tomar decisões baseadas em crenças e não em fatos negligenciando o processo de identificação e também priorização de ameaças, atuando de forma não científica.

Um dos participantes relatou: *"Muitas vezes essa questão das ameaças à validade é negligenciada ou até mesmo feita de forma não científica. O resultado são experimentos caros sem valor científico! Desperdício!"*.

Infelizmente, este comportamento leva ao apoio, ainda que inconsciente, de conclusões falsas comprometendo a validade dos resultados dos estudos.

3.9 AMEAÇAS À VALIDADE

A validade das pesquisas que envolvem o método *survey* está fortemente relacionada ao fato da pesquisa medir o que se deve medir (CHAVA; NACHMIAS, 1996). Os resultados apresentados neste artigo devem ser interpretados dentro dos limites criados por algumas ameaças a seguir:

Validade Externa. A amostra obtida através da população estudada neste *Survey Research* não representa necessariamente a comunidade global de autores de experimentos da área de ES, por isso, a generalização é limitada. Amostras pequenas são dificuldades conhecidas em ES e difíceis de superar (CONTE *et al.*, 2007). Com o melhor de nosso conhecimento, não se tem ciência de pesquisas em ES que afirme que sua amostra é representativa de uma determinada área.

A análise qualitativa trata-se de uma apreciação exploratória e não se pretende generalizar seus resultados, apenas compreender aspectos da pesquisa cuja análise dificilmente poderia ser aprofundada utilizando métodos quantitativos. Ainda assim, conforme os resultados apresentados neste *Survey*, a diversificação de locais, filiações e características dos participantes representadas na amostra, contribuem para uma menor probabilidade de que um segmento específico de especialistas pesquisados influencie nossos resultados.

Validade Interna. No intuito de evitar influência nos resultados desta pesquisa foi implementado apenas um questionário (mesmo a aplicação tendo sido feita em duas etapas). Foi definida estratégia para motivar os entrevistados a responderem os questionamentos e com isso, obter uma taxa de resposta aceitável (Seção 3.6).

Validade de Constructo. No intuito de obter coerência e consistência dos questionamentos apresentados pelo instrumento de coleta de dados, foi realizada uma avaliação prévia do referido questionário através de estudo piloto. Ainda assim, há o risco do questionário influenciar os participantes devido às perguntas e opções de respostas escolhidas assim como, as respostas dos participantes podem não refletir inteiramente a realidade de suas práticas e isso poderá contribuir para imprecisões nos dados e limitações à interpretação dos resultados. Dessa maneira, apesar dos ajustes realizados ao questionário, a construção dos questionamentos pode não ter sido clara e inequívoca o suficiente para garantir que as respostas se baseiam no entendimento consistente das perguntas.

Validade de Conclusão. Relacionado a validade dos resultados foi usado o pacote estatístico IBM SPSS (SPSS, 2011), que permite realizar de forma eficiente análises estatísticas, de forma a evitar a imprecisão e erros estatísticos. Relacionado a validade e confiabilidade das medidas, estas foram descritas de acordo com regras claramente definidas na seção de análise dos dados (Seção 3.7).

3.10 CONCLUSÃO

Apesar da maioria dos participantes estar preocupada com a identificação de ameaças à validade e com as novas ameaças que podem ser inseridas a partir da aplicação de determinadas ações de controle e das consequências ao aplicá-las, algumas evidências apontadas na seção de resultados mostram a lacuna existente entre as orientações existentes na literatura e as práticas reais de tratamento de ameaças à validade.

Os resultados reforçam importância de melhorar confiabilidade das conclusões resultantes de experimentos da área de ES a partir da adoção das seguintes recomendações:

- Identificar e priorizar ameaças à validade durante o planejamento do experimento;
- Refazer ou ajustar as etapas do plano experimental a partir das ameaças à validade identificadas;
- Utilizar guias, *checklists* e ferramentas de software que auxiliem o processo de controle de ameaças à validade;
- Identificar ações de controle que podem causar novas ameaças ao estudo e;
- Identificar ordem de prioridade entre os tipos de ameaças baseado no propósito de cada experimento, evitando decisões equivocadas ao priorizar uma ameaça em detrimento de outra.

3.11 RESUMO

O objetivo do *survey* apresentado neste capítulo foi entender a importância e os procedimentos adotados para controle de ameaças à validade no planejamento de experimentos. A pesquisa foi realizada com autores que publicaram experimentos nas principais conferências e periódicos da área de ES e ESE.

Os resultados advindos da visão de autores de experimentos, avançam o saber da comunidade de Engenharia de Software Experimental, contribuindo, assim, para a melhoria da validade dos resultados de futuros experimentos da área de ES.

4 PRIORITTVS: UM PROCESSO QUE SE DESTINA A APOIAR PESQUISADORES A PRIORIZAR AMEAÇAS À VALIDADE E AS SUAS AÇÕES DE MITIGAÇÃO AO PLANEJAR OS EXPERIMENTOS CONTROLADOS EM ES

“Um passo à frente e você não está mais no mesmo lugar”.

Chico Science.

4.1 INTRODUÇÃO

Conforme apresentado no capítulo anterior, no intuito de melhorar a validade das conclusões resultantes de experimentos da área de ES, devemos identificar e priorizar a maioria das possibilidades de problemas que possam acontecer desde o planejamento até a divulgação dos resultados do estudo. O procedimento permite ajustar ou mesmo refazer as etapas do plano experimental a partir das ameaças à validade identificadas, bem como considerar a prioridade de uso de determinadas ações de mitigação, ciente das consequências ao aplicá-las.

Analisar e priorizar determinadas ameaças à validade e suas respectivas ações de mitigação permitem que os experimentadores possam identificar os maiores riscos à confiabilidade dos resultados do estudo, planejando ações de controle mais assertivas.

A tarefa de análise e priorização de ameaças à validade é definida em razão do propósito e particularidades de cada experimento e não tem processo determinístico (TEIXEIRA; FONSECA; SOARES, 2018). Assim, o pesquisador/experimentador tem um papel fundamental nesta decisão, pois é ele quem define os objetivos do estudo e em dependência destes propósitos, algumas ameaças à validade são mais críticas e algumas ações de mitigação são mais eficazes que outras (ANDERSSON *et al.*, 2003).

O Survey Research descrito no Capítulo 3 revelou que, embora existam na ESE alguns estudos que tratam da temática das ameaças à validade, as estratégias utilizadas para controle de ameaças à validade mencionadas atualmente na literatura não fornecem apoio ao experimentador no sentido de ajudá-lo a definir os maiores riscos de acordo com o propósito de experimentação de seu estudo. A partir destas circunstâncias, apresentamos o processo PrioriTTVs.

A sigla PrioriTTVs é uma abreviação cujo significado reduz as palavras da frase “Priorização de Ameaças à Validade”, do inglês *Prioritization of Threats to Validity*, com o intuito de ser mais ágil e significativa ao falar e ao escrever.

O objetivo geral do processo PrioriTTVs é dar suporte à identificação, análise e priorização de ameaças à validade e suas respectivas ações de mitigação para os riscos identificados, durante o planejamento de experimentos controladas da área de ES. Dessa forma, o experimentador, especialmente o iniciante, poderá de forma cíclica e iterativa, rever o seu plano experimental, ajustando-o a diversos riscos identificados a partir das atividades do processo proposto, como forma de aprimorá-lo e, em seguida, executar o experimento.

Os principais objetivos específicos do processo proposto são:

- Reduzir o esforço exigido pelo pesquisador por meio de um processo estruturado que possa ajudar no suporte à identificação, priorização e mitigação de ameaças à validade de experimentos da área de ES;
- Fornecer uma lista de verificação que possa ajudar a capturar informações sobre as principais informações de um plano experimental;
- Apoiar o processo decisório de priorizar determinadas ameaças à validade de acordo com o propósito do experimento realizado;
- Contribuir para o ensino e aprendizagem da Engenharia de Software Experimental, relacionando os conceitos teóricos estudados em sala de aula com a prática da experimentação;
- Melhorar a qualidade dos futuros experimentos controlados da área de ES através de um processo que possa ajudar o pesquisador/experimentador a lidar de forma sistemática com potenciais ameaças à validade e com isso, agregando uma maior validade a seus resultados.

4.2 TRABALHOS RELACIONADOS

Ainda é restrito o número de abordagens para apoiar o controle das ameaças à validade dos experimentos da área de ES, pesquisadores da comunidade de ESE tem proposto modelo, *checklists* e resumos (FELDT; MAGAZINIUS, 2010).

Tem-se como trabalho relacionado a esta proposta um modelo conceitual apresentado por Neto e Conte (2013). Tal modelo relata uma estratégia de suporte ao pesquisador durante o processo de identificação e controle de ameaças à validade,

identificando relações entre ameaças à validade e suas respectivas ações de controle. O objetivo foi melhorar o processo de identificação e controle de ameaças à validade de estudos controlados em ES e, com isso, aumentar o número de ameaças identificadas e controladas.

Feldt e Magazinius (2010) apresentam um *survey* elaborado com o intuito de averiguar como as ameaças à validade são analisadas e mitigadas em estudos na área de ESE. Como resultado, guidelines para realizar as análises e mitigações foram apresentados para que a validade dos resultados seja maximizada. A partir da análise dos artigos, 136 ameaças à validade foram identificadas dentre os quatro tipos de validade e 69 ações de controle.

Wohlin *et al.* (2012) apresentam através de um livro, um *checklist* que contém sete ameaças à validade de conclusão, 13 ameaças à validade interna, dez ameaças à validade de constructo e três ameaças à validade externa. Este *checklist* tem sido amplamente utilizado pela comunidade de ESE como uma estratégia para a verificação se as ameaças presentes na lista podem ocorrer em seu experimento (TEIXEIRA; FONSECA, SOARES, 2018).

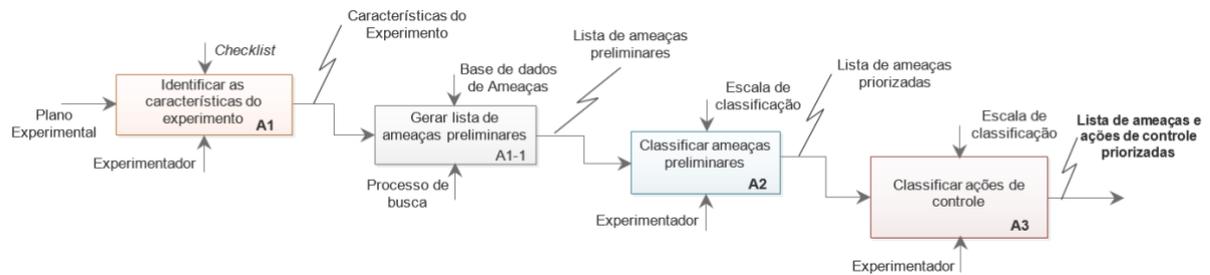
Embora existam na ESE estudos que tratam da temática das ameaças à validade e apresentem propostas de apoio ao pesquisador quanto a identificação e controle das ameaças à validade, as estratégias mencionadas atualmente na literatura não fornecem apoio ao experimentador no sentido de ajudá-lo a priorizar os maiores riscos de acordo com o propósito de experimentação de seu estudo e com isso, dificulta a implementação de ações de controle mais assertivas.

4.3 MODELAGEM CONCEITUAL

Para elaborar o modelo conceitual com a visão geral do fluxo funcional do processo PrioriTTVs foi utilizada *Structured Analysis* (SA), uma linguagem de análise estruturada projetada especificamente para ajudar as pessoas a descrever modelagens funcionais para análise e entendimento de sistemas de ES e processos de negócio.

SA foi desenvolvida no final dos anos 1960 por Douglas T. Ross (ROSS, 1977), formalizada e publicada como uma metodologia de modelagem de função IDEF0 (PRESS, 2001).

Figura 8 – PrioriTTVs: processo para identificar e priorizar ameaças à validade



Fonte: Elaboração própria (2019).

Como mostrado na Figura 8, a posição de cada seta nos retângulos que acomodam as atividades A1, A2 e A3 representa um papel específico no processo. As informações de **entrada**, dados ou objetos de cada etapa são representadas pelas setas à esquerda, as **saídas** de cada operação são produzidas por setas à direita, as informações de **controle** por setas na parte superior e os **mecanismos** que fornecem meios de suporte pelas setas na parte inferior, expressando a relação de cada etapa com todo o processo.

4.3.1 Identificação das Informações de um Plano Experimental (A1)

Nesta atividade, as particularidades de um determinado experimento da área de ES serão coletadas através um instrumento (*checklist*), cujo objetivo é capturar as principais informações do plano experimental. As informações do experimento são a saída da Atividade A1 que, por sua vez, são a entrada para a próxima atividade.

A Atividade A1-1 usa uma operação automática de busca como mecanismo de suporte para relacionar as respostas da lista de verificação preenchidas pelo pesquisador com um banco de dados de ameaças à validade e das ações de controle. Ao final, a saída é uma lista de ameaças em potencial e suas respectivas ações de mitigação como sugestão para o experimentador. Mais informações a respeito do instrumento de caracterização podem ser encontradas na Seção 4.4.

4.3.2 Classificação das Ameaças Preliminares (A2)

Nesse ponto, o pesquisador classifica as ameaças identificadas anteriormente na Atividade A1 de acordo com os critérios de impacto, urgência e tendência dos riscos apresentados. IMPACTO (consequências para o experimento provocada pela ameaça), URGÊNCIA (rapidez com que uma ameaça específica deve ser resolvida)

e TENDÊNCIA (prognóstico temporal de como a ameaça tende a evoluir ao longo da execução do experimento). Cada um dos três critérios recebem uma pontuação de classificação de intensidade definida como baixa (valor 1), média (valor 2) e alta (valor 3), sendo que no cálculo de magnitude da ameaça o critério impacto terá peso 3, urgência peso 2 e tendência peso 1, conforme a Tabela 6. Torna-se válido destacar que esse modelo de matriz é uma adaptação de uma ferramenta que auxilia na priorização de resolução de problemas identificados (DAYCHOUW, 2007).

Tabela 6 – Critérios, pesos e intensidades de classificação de ameaças

Pesos	Intensidade	Impacto	Urgência	Tendência
3	Alta	Muito Grave	Muito Urgente	Ir agravar imediatamente
2	Mdia	Grave	Urgente	Ir agravar ao longo da execuo
1	Baixa	Pouco Grave	Pouco Urgente	No ir agravar

Fonte: Elaborao prpria (2019).

A matriz  uma possibilidade de classificao de cada problema/risco que  julgado pertinente para o desenvolvimento do experimento pela tica do impacto (ou gravidade da interferncia da ameaa e como afeta a validade dos resultados), da urgncia (de resoluo dele) e tendncia (de evoluo da situao de risco se agravar ao longo da execuo do experimento). , pois, um recurso utilizado para a definio de prioridades de tarefas no contexto do experimento executado (MEIRELES, 2001; SELEME; STADLER, 2008).

A partir da classificao dos riscos sugeridos  preciso organizar em formato de matriz as informaes sobre o impacto, urgncia e tendncia das ameaas para que possamos perceber quais ameaas so mais ou menos crticas. Para isso, o processo utiliza uma escala de magnitude (Tabela 7), desenvolvida a partir de uma matriz de riscos.

A Tabela 7 mostra a escala e categorias de magnitude elaborada a partir da adaptao da matriz de Garvey e Lansdowne (1998) que apresentaram abordagem estruturada para conhecer quais riscos so mais ou menos crticos. A partir da elaborao desta escala, o experimentador poder definir o grau de relevncia e criticidade que cada ameaa identificada ter em seu experimento, definindo uma ordem de prioridade atravs das categorias de magnitude: (5) Muito Alta, (4) Alta, (3) Moderada, (2) Baixa e, (1) Muito Baixa. Desse modo, com uma tica clara e objetiva

do problema a ser priorizado, o pesquisador poderá definir ações de mitigação necessárias e assertivas à solução das ameaças à validade de seus resultados.

O cálculo de magnitude de uma ameaça será realizado em virtude da pontuação (pesos) de cada critério definido pelo pesquisador, em razão dos fatores mais críticos à realidade da experimentação.

A equação se dará a partir da multiplicação dos valores de impacto, urgência e tendência de cada uma das ameaças classificadas, conforme a fórmula: $P = I \times U \times T$, sendo que “I” refere-se ao critério de impacto, “U” critério de urgência e “T” ao critério de tendência. Por exemplo, o cálculo de magnitude para uma ameaça classificada pelo pesquisador com impacto 3, urgência 2 e tendência 2 será: $3 \times 2 \times 2$, obtendo-se como produto o valor 12, ou seja, magnitude alta, de acordo com Tabela 7.

Tabela 7 – Escala e Categorias de Magnitude de Ameaças à Validade

CRITÉRIOS			PONTOS (P)	LIMITES	MAGNITUDE	DEFINIÇÃO
IMPACTO	URGÊNCIA	TENDÊNCIA				
3	3	3	27	P ≥ 15	Muito Alta (5)	Ameaças onde a maioria dos critérios foram classificados como de alta (3) intensidade e ausência de critérios de baixa (1) intensidade. Se ocorrida, possui probabilidade muito alta de impactar negativamente a validade dos resultados da pesquisa.
3	3	2	18			
3	2	3	18			
2	3	3	18			
3	2	2	12	9 ≤ P < 15	Alta (4)	Ameaças onde houver a classificação de um critério de baixa (1) e demais de alta (3) intensidade ou classificação de um critério de alta (3) e demais de média (2) intensidade. Se ocorrida, possui probabilidade alta de impactar negativamente a validade dos resultados da pesquisa.
2	3	2	12			
2	2	3	12			
3	3	1	9			
3	1	3	9			
1	3	3	9			

2	2	2	8	5 ≤ P < 9	Moderada (3)	Ameaças onde todos os critérios foram classificados como de média intensidade ou quando há uma distribuição de diferentes intensidades entre todos os critérios, distinguindo-se entre si. Se ocorrida, possui probabilidade moderada de impactar negativamente a validade dos resultados da pesquisa.
3	2	1	6			
3	1	2	6			
2	3	1	6			
2	1	3	6			
1	3	2	6			
1	2	3	6			
2	2	1	4	3 ≤ P < 5	Baixa (2)	Ameaças onde houver a classificação de um critério de alta (3) e demais de baixa (1) intensidade ou classificação de um critério de baixa (1) e demais de média (2) intensidade. Se ocorrida, possui probabilidade baixa de impactar negativamente a validade dos resultados da pesquisa.
2	1	2	4			
1	2	2	4			
3	1	1	3			
1	3	1	3			
1	1	3	3			
2	1	1	2	P < 3	Muito Baixa (1)	Ameaças onde nenhum critério foi classificado como de alta intensidade (3) e a maioria de baixa intensidade (1). Se ocorrida, possui probabilidade muito baixa de impactar negativamente a validade dos resultados da pesquisa.
1	2	1	2			
1	1	2	2			
1	1	1	1			

Fonte: Elaboração própria (2019).

O resultado desta atividade (A2) será uma lista de ameaças à validade classificadas pelo pesquisador de acordo com os objetivos e informações de cada plano experimental.

4.3.3 Classificação da assertividade das Ações de Controle (A3)

Nesta atividade do processo, o pesquisador classifica as ações de mitigação para cada ameaça à validade identificada. A referida classificação utiliza uma escala de importância/efetividade, considerando os seguintes valores possíveis: (i) não é importante; (ii) pouco importante; (iii) importante; (iv) muito importante; e (v) extremamente importante. Com isso, o pesquisador define quais as ações são mais

assertivas ao mitigar determinadas ameaças.

4.4 INSTRUMENTO DE CARACTERIZAÇÃO DE UM PLANO EXPERIMENTAL

Conforme a Seção 4.3.1, para a identificação das principais informações de um plano experimental, criou-se um instrumento cujo objetivo é capturar os atributos de um experimento com base em um conjunto de perguntas objetivas, respondidas por cada pesquisador.

A criação deste instrumento de coleta e a organização das perguntas que podem ajudar ao pesquisador a identificar as principais informações do experimento foram baseadas na estrutura de plano experimental definido por Fonseca (2016), dividido em oito categorias:

1) Definição de Objetivos: Essa categoria de perguntas permite que os pesquisadores revisem seu plano experimental em relação aos objetivos do estudo, às questões de pesquisa, à escolha do experimento controlado como a técnica de pesquisa mais apropriada a ser usada e preocupações com aspectos éticos (DYBÅ; DINGSØYR, 2008; WOHLIN *et al.*, 2012). Exemplo: As questões de pesquisa foram definidas de forma clara e relacionadas aos objetivos do estudo?

2) Hipóteses, Variáveis e Medidas: Esse grupo de itens tem o objetivo de revisar a relação entre hipóteses, variáveis e medidas e os objetivos da pesquisa (JEDLITSCHKA; CIOLKOWSKI; PFAHL, 2008; KITCHENHAM *et al.*, 2009). Exemplo: Existe alguma possibilidade de os participantes terem conhecimento prévio das hipóteses do experimento?

3) Participantes: Este conjunto de elementos diz respeito às preocupações com os seres humanos participantes do experimento, estratégia e descrição do processo de recrutamento, coleta de informações dos participantes, preocupações com a população para a qual os participantes foram escolhidos, tamanho da amostra e até mesmo a maneira de como lidamos com eles (JEDLITSCHKA; CIOLKOWSKI; PFAHL, 2008; KO; LATOZA; BURNETT, 2015). Exemplo: Haverá participantes com diferentes níveis de experiência ou habilidades na execução de tarefas durante o experimento?

4) Materiais e Tarefas experimentais: Este conjunto de questionamentos é focado em materiais e tarefas que devem ser usados no experimento controlado (JEDLITSCHKA; CIOLKOWSKI; PFAHL, 2008). Exemplo: Artefatos e procedimentos

experimentais como instrumentos, materiais, tecnologias, ferramentas, tarefas etc., são representativos daqueles utilizados num contexto real?

5) Design Experimental: Essa classe de perguntas faz com que os pesquisadores verifiquem se o design experimental escolhido é o mais apropriado ao experimento, se os tratamentos estão bem definidos, se a aleatorização está bem descrita e se algum procedimento de mascaramento ou ocultação (*blinding*) deve ser aplicado para reduzir o viés do estudo (MONTGOMERY, 2017). Exemplo: O uso de um determinado tratamento pode influenciar outros tratamentos usados no experimento?

6) Procedimentos: Esta seção de questionamentos permite revisar a seção de procedimentos, incluindo uma descrição adequada do contexto do experimento controlado, treinamento, estudo piloto e cronograma de execução do experimento (OPPENHEIM, 2001; KO; LATOZA; BURNETT, 2015). Exemplo: O experimento será realizado em diferentes ambientes experimentais?

7) Coleta e Análise de Dados: Esta categoria inclui itens relacionados aos procedimentos utilizados para coleta e análise de dados. Além disso, esta seção apresenta preocupações sobre métodos e o rigor estatístico do experimento (JEDLITSCHKA; CIOLKOWSKI; PFAHL, 2008). Exemplo: Dados que serão coletados podem apresentar-se de forma incompleta, imprecisa e/ou não confiável?

8) Estrutura do Documento: Esta categoria concentra-se na redação geral do plano experimental em relação à adequação do público e sua facilidade de leitura. Exemplo: Você planeja apresentar ou indicar onde os dados brutos de sua avaliação experimental estarão disponíveis?

Analisando a estrutura das 43 questões categóricas do instrumento de coleta de dados, observou-se que a categoria Participantes foi a que mais concentrou um maior número de questões de verificação, com representatividade de 30,2% em relação ao total de itens identificados. Justifica-se investigações realizadas em experimentos, participantes do processo desempenham papel indispensável na execução, na condução e na avaliação dos procedimentos experimentais, sendo os mesmos elementos fundamentais para o desenvolvimento de resultados mais confiáveis e, ao mesmo tempo, com maior rigor de validade.

Como são os agentes que influenciam, por meio de atitudes e também de comportamentos, impactos pré-definidos dos projetos experimentais, percebe-se, pois, que também podem configurar em diversas ameaças à validade dos dados. Os

participantes, em si, também determinam outra forma de ameaça, atrelada aos Objetivos, em especial no que compete à comunicação e interação entre os mesmos, quando não permitido o engajamento e a sociabilidade, a fim de não apresentar quaisquer riscos aos experimentos.

A categoria de Procedimentos obteve significativa representatividade no número de questionamentos do instrumento de coleta, adquirindo 18,6%, enquanto Materiais, Métodos e Tarefas teve percentual de 13,9%. Os procedimentos dentro da área de projetos experimentais são, também, variáveis essenciais para atingir a melhor consecução dos resultados, uma vez que, quando sistematizados e organizados em planejamento, tendem a resultar em uma maior confiabilidade, além de certeza e de segurança dos dados e informações pertinentes ao experimento controlado.

Por sua vez, a categoria de materiais, métodos e tarefas são contextos combinatórios, juntamente com procedimentos, para melhoria nos resultados e na monitoria dos projetos controlados em ES, a fim de suscitar, a experimentadores, os critérios de identificação de possíveis ameaças à validação e sua consequente definição das medidas corretivas compatíveis a cada fator de risco.

O instrumento possui 43 perguntas associadas às oito categorias definidas anteriormente. O instrumento completo (em inglês), pode ser encontrado no **Apêndice A**. As questões do *checklist* podem ser respondidas com uma das quatro opções a seguir:

- **Sim**: significa que plano experimental possui certa informação/característica;
- **Parcialmente**: o plano experimental pode ter essa informação/característica, mas o pesquisador não tem certeza;
- **Não**: o plano experimental não possui certa informação/característica;
- **Não aplicável**: o item avaliado não é aplicável ao plano experimental avaliado.

4.5 RESUMO

Este capítulo apresentou a modelagem de um processo que se destina a apoiar os pesquisadores a priorizar ameaças à validade e suas ações de mitigação ao planejar experimentos em ES, assim como, o instrumento de caracterização, utilizado

para ajudar a capturar os atributos de um experimento com base em um conjunto de 43 perguntas objetivas, respondidas por cada pesquisador.

Os principais benefícios almejados pelo processo são: a) reduzir o esforço exigido por pesquisadores iniciantes ou experientes; b) fornecer um *checklist* que possa ajudar a capturar informações sobre os principais aspectos de um plano experimental; c) apoiar o processo decisório de priorizar ameaças de acordo com o objetivo do experimento realizado e; d) contribuir para o ensino e aprendizagem da Engenharia de Software Experimental.

5 VALIDEPLAN: FERRAMENTA PARA PLANEJAMENTO DE EXPERIMENTOS DE ENGENHARIA DE SOFTWARE ORIENTADA À VALIDADE

“O valor de uma ideia está na sua utilidade”.

— Thomas Edison, cofundador da General Electric.

5.1 INTRODUÇÃO

A área de ES, inegavelmente, tem desempenhado um papel estratégico e preponderante no desenvolvimento de mercados e de atendimentos a diversas demandas, seja na área industrial, social, científica e também de pesquisa, graças à necessidade de formulação e concepção de novos recursos e instrumentos que simulem fenômenos, identifique eventuais problemáticas e apresente, pois, as soluções mais convenientes ao exercício de sua finalidade.

Embora existam na ESE alguns estudos que tratam da temática das ameaças à validade (KITCHENHAM *et al.*, 2002; WOHLIN *et al.*, 2012; NETO; CONTE, 2013; JURISTO; MORENO, 2013), nenhum deles apresenta ferramenta ou mecanismo automático ou semiautomático que dê suporte de como essas ameaças e essas ações de mitigação podem ser classificadas, priorizadas e tratadas, de acordo com características de cada estudo.

Como forma de semiautomatizar processo **PrioriTTVs** exposto anteriormente (Capítulo 4), propomos a **ValidEPlan**, uma ferramenta web que auxilia pesquisadores iniciantes e experientes a planejar experimentos controlados da área de ES, com o enfoque na identificação e priorização das ameaças à validade e suas ações de mitigação.

5.2 VALIDEPLAN – VALIDITY-DRIVEN SOFTWARE ENGINEERING EXPERIMENTS PLANNING TOOL

A ValiDEPlan¹ é uma ferramenta de planejamento de experimentos da área de ES, orientada à validade dos resultados. A ValiDEPlan é uma evolução de uma outra

¹ ValiDEPlan: Validity-Driven Software Engineering Experiments Planning Tool.
<https://valideplan.cin.ufpe.br>

ferramenta chamada *Reviewer EP* (FONSECA, 2016), plataforma web colaborativa para revisão da completude de planos experimentais.

A transformação permitiu ao novo aplicativo gerenciar as ameaças à validade durante a fase de planejamento de experimento, antecipando de forma direta riscos em potencial que possam acontecer na execução do referido estudo, permitindo a classificação de possíveis ameaças à validade, indicando quais devem ter prioridade ao serem tratadas e quais são as ações de controle mais assertivas para mitigá-las.

A ferramenta foi escrita na língua inglesa para facilitar acesso a pesquisadores de diversos países. Esta sessão contempla a definição e análise dos requisitos, projeto (arquitetura do sistema, linguagens de programação, gerenciador de banco de dados e padrões de interface), e as principais funcionalidades da ValiDEPlan.

5.2.1 Definições e Requisitos da ValiDEPlan

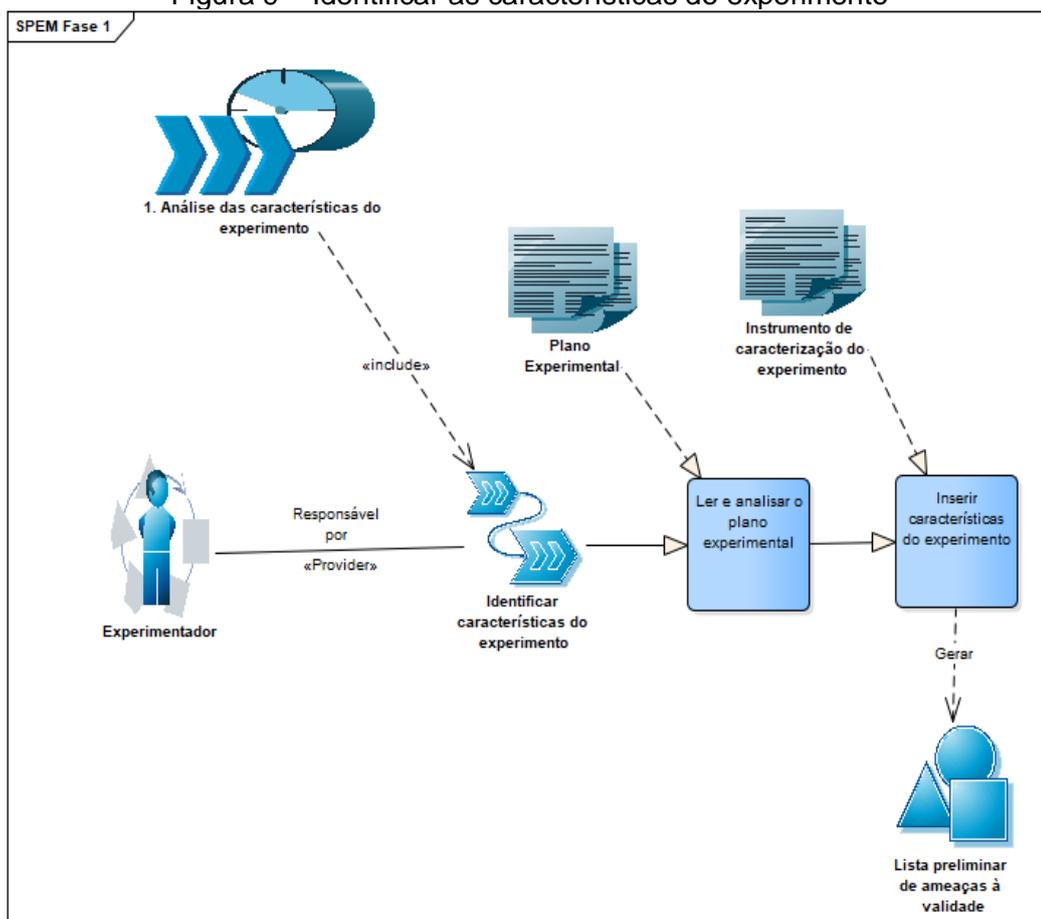
Esta atividade teve como objetivo criar uma estratégia de solução do problema e fornecer uma visão do que deve ser construído e o que o sistema deve realizar, antes de definir como isso será feito. Nessa fase, buscamos identificar e priorizar as necessidades dos usuários da ferramenta, usando como referência a descrição do processo PrioriTTVs proposto (Capítulo 4).

De acordo com o processo proposto PrioriTTVs, o desenvolvimento de ValidEPlan tem o intuito de reduzir o esforço do pesquisador em relação ao processo de identificação, análise, priorização e controle de ameaças a validade e respectivas ações de mitigação para o risco identificado, durante planejamento de experimentos controlados da área de ES, tornando o processo semiautomatizado, sistemático e organizado.

Para facilitar o entendimento e melhor especificar principais fases do processo, foram desenvolvidos novos diagramas através da notação SPEM² (*Software Process Engineering Meta Model*), usando a ferramenta *Sparx Systems Enterprise Architect Version 12.0*.

² O SPEM é um padrão adotado pela OMG (Object Management Group) para a definição do processo de desenvolvimento de software, incluindo os processos que envolvem ou exigem o uso de UML.

Figura 9 – Identificar as características do experimento

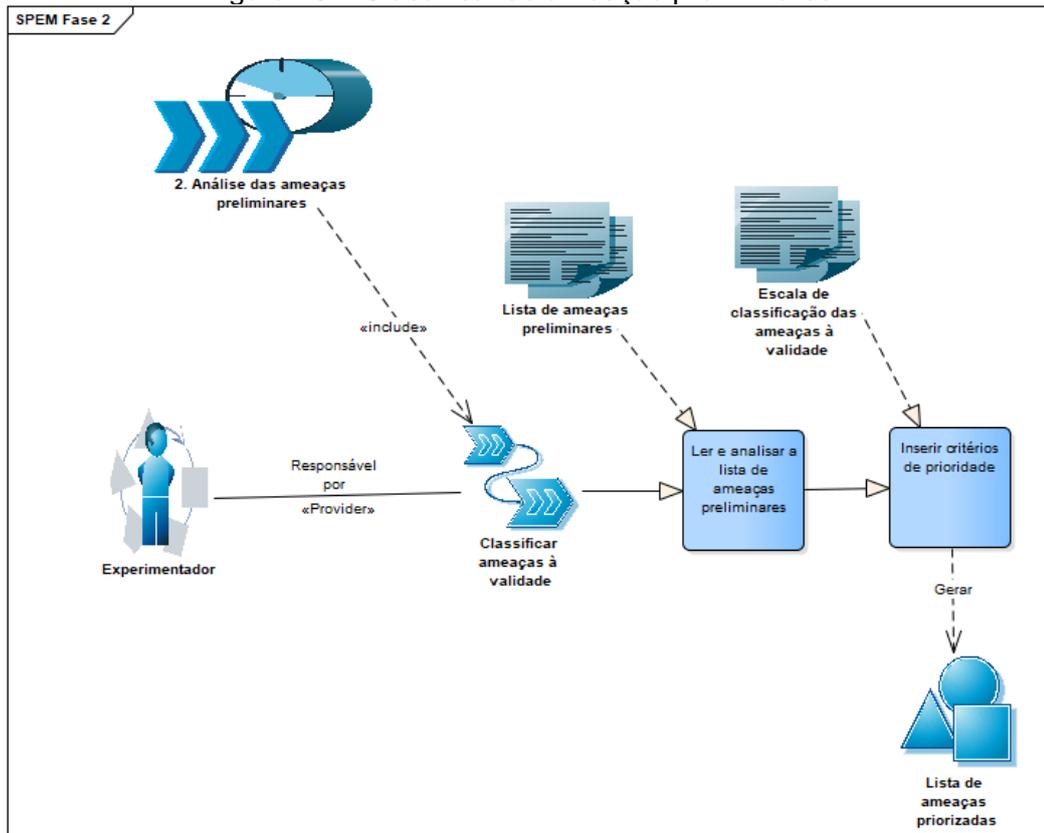


Fonte: Elaboração própria (2019).

A Figura 9 mostra o passo a passo da fase de identificação das principais informações de um experimento. O objetivo é reduzir o número de ameaças à validade e suas respectivas ações de controle para um universo relacionado apenas aos propósitos do experimento.

Para isso, iremos capturar principais características de um plano experimental com base em um instrumento de caracterização que deverá ser preenchido pelo experimentador. Ao final desta etapa, será gerado lista preliminar com as possíveis ameaças e ações de controle que podem influenciar no referido experimento.

Figura 10 – Classificar as ameaças preliminares

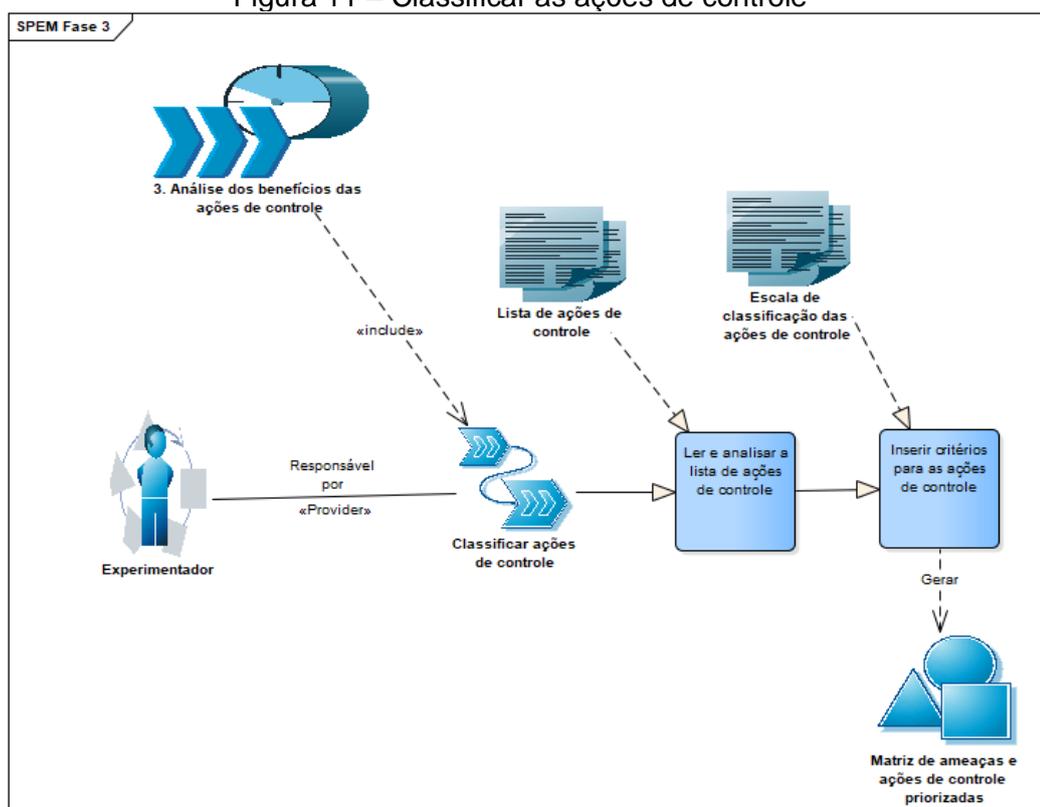


Fonte: Elaboração própria (2019).

A Figura 10 mostra o passo a passo da fase de classificação da lista de ameaças preliminares gerada pela fase anterior. O objetivo é obter uma lista de ameaças priorizadas de acordo com critérios definidos pelo pesquisador.

Com isso, o pesquisador poderá implementar as ações de controle mais assertivas a partir da distinção de determinadas ameaças, consideradas mais graves ao propósito de investigação do seu experimento.

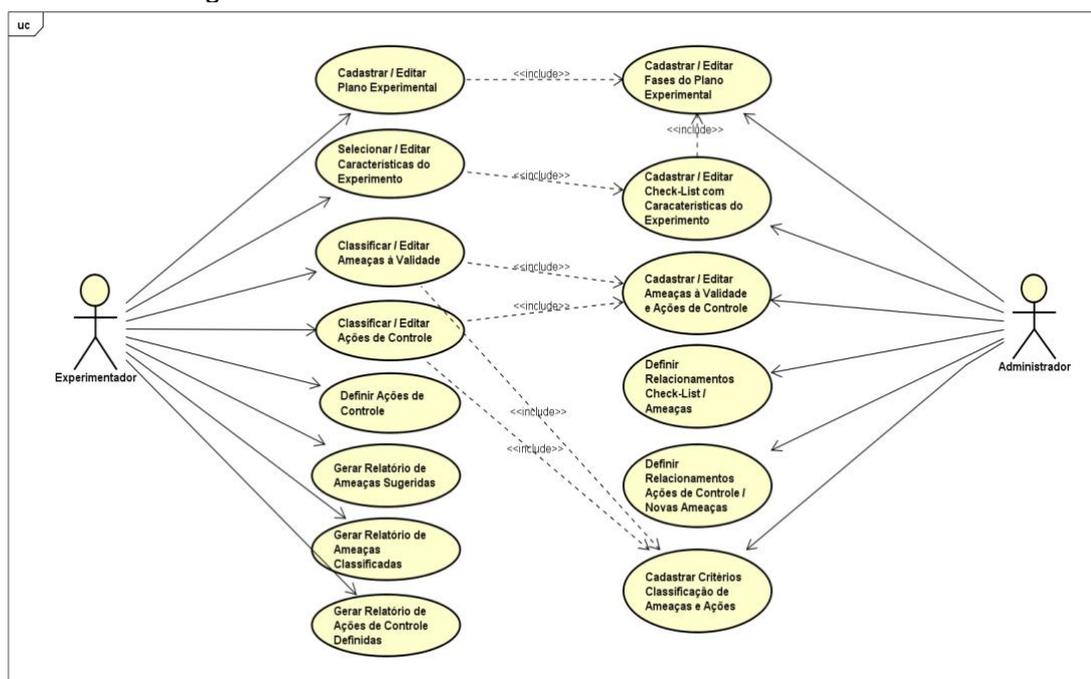
Figura 11 – Classificar as ações de controle



Fonte: Elaboração própria (2019).

A Figura 11 mostra o passo a passo da fase de classificação das ações de controle. O objetivo é classificar as ações de controle das ameaças anteriormente selecionadas de acordo com uma escala de importância, garantindo que as ações de controle que causem novas ameaças sejam classificadas em virtude de seu grau de relevância ao experimento.

A Figura 12 representa como os casos de uso interagem entre si no sistema e com o usuário experimentador (pesquisador usuário do sistema) e administrador (responsável pela administração do sistema), dispondo das seguintes unidades funcionais: Cadastrar plano experimental, identificar características do experimento, classificar ameaças e ações de mitigação, definir as ações de controle e gerar os relatórios. Todas as informações específicas de cada caso de uso como descrição, fluxo de eventos básicos e alternativos, pré e pós condições, pontos de extensão, estão disponíveis em <https://bit.ly/35DVxtm>.

Figura 12 – Visão Geral dos Casos de Uso da *ValidEPlan*

Fonte: Elaboração própria (2019).

Além dos requisitos acima descritos, foram definidos os seguintes requisitos não funcionais:

- **Usabilidade:** Interface amigável (o sistema deve possuir uma interface simples e de fácil utilização, que possa atender ao usuário de forma satisfatória e intuitiva) e Documentação do Sistema (o sistema deve possuir uma documentação mínima, que responda às perguntas relativas a modelagem e implementação do sistema);

- **Confiabilidade:** Correção de falhas (o sistema deve possuir um log com todas as operações realizadas num determinado espaço de tempo, que deve ser verificado após cada transação possibilitando a verificação de erros com envio de mensagem ao administrador);

- **Padrões:** Padrão de interface (o sistema deve possuir uma interface obedecendo ao padrão WEB com telas, ícones e botões, facilitando o atendimento do requisito Usabilidade), Padrão de Codificação (o sistema deve ser escrito em JAVA (*Jersey*, *Hibernate* e *Jetty*) seguindo os padrões de codificação adotados pela linguagem, facilitando a obtenção do requisito usabilidade) e Padrão para impressão de relatórios (o sistema deve fazer a geração de relatórios de acordo com os protótipos gerados);

- **Hardware e Software:** Implementação (o sistema deve se adaptar a qualquer

tipo de sistema operacional, ou seja, multiplataforma em uma máquina com no mínimo 1 GB RAM, com processador Core i3 (ou compatível) de 2,3 GHZ ou superior);

- **Desempenho:** Rapidez (o sistema deve possuir um tempo máximo para a execução de uma determinada transação, oferecendo um controle de *timeout*, caso a operação não tenha sido realizada por motivos independentes do sistema, por exemplo operações via rede. Os avisos de erros ocorridos devem ser dados num curto espaço de tempo oferecendo ao usuário a possibilidade de correção o mais rapidamente);

- **Segurança:** Segurança de acesso (o sistema deve oferecer um módulo de segurança, que possa ajudar a gerenciar os usuários e seus acessos, restringindo visões e associando a cada usuário uma senha que o identifique unicamente no sistema.) e Integridade (para garantir a integridade dos dados e a consistência das operações, o sistema utilizará um sistema gerenciador de banco de dados – SGBD).

Os documentos completos com Caso de Negócio, Diagrama e Especificação de Casos de Uso, Modelo Entidade Relacionamento (ER), Cenários e Protótipos estão disponíveis em: <https://bit.ly/35DVxtm>.

5.2.2 Arquitetura da ValidEPlan

A ValidEPlan é uma aplicação web desenvolvida com tecnologia Angular, responsável por fazer chamadas para API desenvolvida em Java. A comunicação é efetuada através de REST *web services*.

O *setup* de configuração do ambiente de desenvolvimento é simples e dependente da (s) IDE (s) e SGBD que o desenvolvedor preferir utilizar. Nós utilizamos o *IntelliJ* para desenvolver as funcionalidades do *back end* e *front end* a fim de fazer acesso/*queries* ao banco de dados. Para ambientes de produção ou quaisquer outros que porventura venham a ser necessários, a aplicação é executada em contêineres do *Docker*, sendo um contêiner isolando o *front end* e outro isolando a API.

Para as tecnologias utilizadas no *back end* utilizamos *Java*, *Jersey*, *Hibernate* e *Jetty*. Para o *Front end* utilizamos tecnologias *Angular*, *Materialize* e *Nginx*. Utilizamos *PostgreSql* para o banco de dados e o *deploy*, *shell script* e utilização de *containers* e *dockers*.

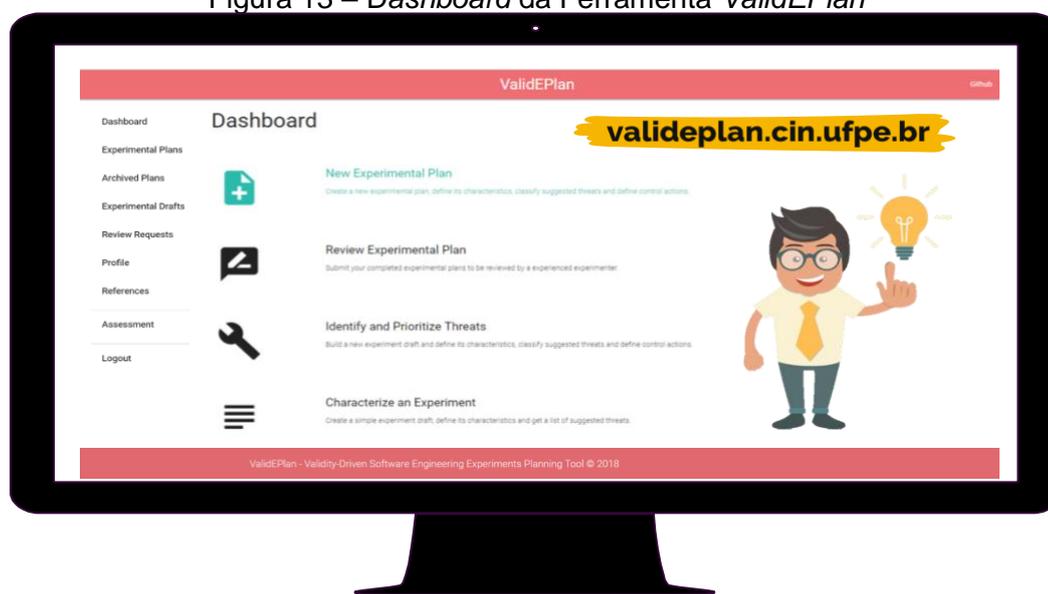
5.2.3 Funcionalidades da ValidEPlan

A ValiDEPlan pode ser utilizada por qualquer pesquisador, seja ele experiente ou iniciante, na construção de planos experimentais orientados à validade de seus resultados. Dentre os perfis com acesso à ferramenta, temos: administrador e experimentador (usuário que pode, também ser revisor colaborador).

Algumas funcionalidades herdadas da versão anterior para a ValiDEPlan: a) Submeter planos experimentais a serem revisados pelos colaboradores; b) Visualizar status da revisão dos planos experimentais e; c) Fazer *upload* e *download* dos planos experimentais. A funcionalidade de criar e revisar PE foi adaptada às modificações do planejamento orientados a ameaças à validade.

As funcionalidades seguintes foram adicionadas à ValiDEPlan: a) Identificar características de um experimento; b) Classificar ameaças à validade; c) Identificar e priorizar ameaças à validade; d) Analisar e selecionar as ações de mitigação e; e) Adicionar as novas ameaças e as ações de controle relacionadas ao experimento.

Figura 13 – Dashboard da Ferramenta ValidEPlan



Fonte: Elaboração própria (2019).

A Figura 13 mostra um screenshot (captura de tela) da ferramenta ValidEPlan onde podemos visualizar o menu e principais cenários de uso da aplicação. A seguir trataremos a descrição dos principais cenários funcionais da ferramenta ValidEPlan:

Produzir um Novo Plano Experimental (C1): Permite ao usuário criar um plano experimental desde o seu início até sua completude. No entanto, além disto, ela ao mesmo tempo em que ajuda o pesquisador a elaborar o plano experimental,

permite identificar suas principais características (C3).

Ao selecionar a funcionalidade de criar um Novo Plano experimental, o usuário informa nome e descrição do plano e inicia preenchendo o plano experimental de acordo com o instrumento desenvolvido por Fonseca (2016). Ao longo do processo de criação do plano, o usuário também será exposto a uma lista de verificação (Seção 4.4), que é um instrumento de caracterização criado para ajudar a especificar as principais características do plano experimental.

Caso deseje, a ValiDEPlan permite ao usuário enviar o plano experimental criado para ser revisado por um pesquisador mais experiente (Figura 14).

Figura 14 – Produzir um novo plano experimental

Edit Plan

Plan Name
Plano Experimental

Experiment Description
Descrição do Plano Experimental

- + Stating the goals
- + Hypotheses, Variables and Measurements
- + Participants
- + Experimental Materials and Tasks
- + Experimental Design
- + Procedure
- + Data Collection and Data Analysis
- + Threats to Validity
- + Document

SAVE

Fonte: Elaboração própria (2019).

Revisar um Plano Experimental (C2): Caso o usuário deseje iniciar por esta funcionalidade, ele pode submeter o plano experimental completo no formato PDF para ser revisado por um pesquisador mais experiente. O usuário deverá incluir revisores disponíveis na base e inserir uma data limite para o retorno da revisão. Os revisores cadastrados como disponíveis na plataforma receberão um *e-mail* e poderão aceitar ou recusar o pedido de revisão. Esta funcionalidade pode ser acessada

também quando o usuário finaliza um plano experimental (Figura 14).

Figura 15 – Revisar um Plano Experimental



Name	Version	State	Action
Teste de plano 1110	1	Ready to review	⊕

Fonte: Elaboração própria (2019).

Identificar características do experimento (C3): Esta funcionalidade pode ser acessada diretamente, caso o usuário já tenha elaborado um plano experimental do experimento ou pode ser associada a criação de um novo plano experimental (C1). O objetivo desta função é ajudar o pesquisador a identificar as principais características do plano experimental através de uma lista de verificação (Seção 4.4). Esta lista é composta por 43 questões dividida em oito categorias.

Cada questão é associada a uma escala de respostas: Sim (significa que o experimento possui a característica), Parcialmente (significa que o experimento tem parcialmente a característica), Não (significa que o experimento não tem a característica) e Não Aplicável (significa que o item perguntado não é aplicado ao experimento).

A partir da inserção das informações sobre as características do plano experimental, a ValiDEPlan acessa um banco de dados de ameaças à validade existente e gera uma lista de possíveis ameaças e ações de mitigação, conforme a Figura 16.

Inicialmente usamos um conjunto de ameaças e ações de mitigação identificadas através de uma revisão sistemática realizada por Neto & Conte (2013) como entradas para o banco de dados da ferramenta (**Anexo C**). Como trabalho futuro propomos realizar um estudo de mapeamento sistemático para identificar novas ameaças e ações de controle de experimentos para diferentes contextos/propósitos em ES.

Figura 16 – Identificar informações do experimento

ValidEPlan

Experiment Characteristics

Draft: Novo Plano Experimental

Answer the following questionnaire to generate a list of suggested threats for your experiment.

Answer the following questions considering the options:

- Yes: Means that your experiment has this characteristic.
- Partially: Means that your experiment may have this characteristic, but you are not sure.
- No: Means that your experiment has not this characteristic.
- Not applicable: Means that the item is not applicable for your experiment.

+ Stating the goals

+ Hypotheses, Variables and Measurements

+ Participants

+ Experimental Materials and Tasks

+ Experimental Design

It makes the researchers check if the experiment design chosen is the most appropriate, if the treatments are well defined, if the randomization is well described, and if an appropriate blinding procedure should be applied to reduce bias.

1. Could the treatments not represent the treatments used in the real context?

Yes Partially No

Not applicable

2. Could the use of a treatment influence other treatments used in the experiment?

Yes Partially No

Not applicable

3. Will each group use only one experimental artifact, or task, or treatment?

Yes Partially No

Not applicable

+ Procedure

+ Data Collection and Data Analysis

+ Document

Fonte: Elaboração própria (2019).

Classificar ameaças à validade sugeridas para o experimento (C4): O usuário deve já ter realizado a identificação das principais informações do experimento através da C1 ou C3. Baseado nos dados identificados, a ValiDEPlan gera uma lista de possíveis ameaças.

O usuário deverá classificar cada ameaça sugerida em uma escala de 1 a 3 com base na intensidade nos seguintes três critérios: IMPACTO (consequências para o experimento provocada pela ameaça), URGÊNCIA (rapidez com que determinada ameaça deve ser resolvida) e TENDÊNCIA (prognóstico sobre como a ameaça tende a evoluir). A saída dessa atividade é uma lista de ameaças à validade priorizadas (Figura 17).

Figura 17 – Classificar ameaças à validade sugeridas para o experimento

ValidEPlan

Threats

Plan: Plano Experimental

DEFAULT THREATS
CUSTOM THREATS

1. Impossibility to replicate the study in similar context ?

- Impact
- 1 2 3
- Urgency
- 1 2 3
- Trend
- 1 2 3

2. Researcher influence ?

- Impact
- 1 2 3
- Urgency
- 1 2 3
- Trend
- 1 2 3

3. Experimental hypotheses guessing ?

- Impact
- 1 2 3
- Urgency
- 1 2 3
- Trend
- 1 2 3

ValidEPlan - Validity-Driven Software Engineering Experiments Planning Tool © 2018

Fonte: Elaboração própria (2019).

Identificar e priorizar as ameaças à validade (C5): Esta funcionalidade unifica as funcionalidades C3 e C4, ou seja, o usuário tem a opção de, a partir de um plano experimental construído, identificar as principais características do plano e classificar por ordem de prioridade as ameaças à validade sugeridas juntamente com suas respectivas ações de controle.

Analisar e Selecionar ações de controle (C6): Nesta funcionalidade, o experimentador classifica um conjunto de ações de mitigação (anteriormente sugeridas pela ferramenta) para cada ameaça à validade. Depois de classificar as ameaças sugeridas pela ValidEPlan para o experimento, ela define uma escala de magnitude para cada ameaça de acordo com sua prioridade (Muito alta, Alta, Moderada, Baixa e Muito Baixa). O usuário pode acessar, através de um filtro, qual prioridade interessa.

Em seguida, o usuário classifica cada ação de controle utilizando uma escala de importância/efetividade, considerando os seguintes valores possíveis: (i) não

importante; (ii) ligeiramente importante; (iii) moderadamente importante; (iv) importante; e (v) muito importante. O resultado desta atividade é uma lista de ações de controle para cada ameaça identificada.

5.3 RESUMO

Neste capítulo, apresentamos a ferramenta ValidEPlan, uma aplicação web útil ao planejamento de experimentos com o diferencial de auxiliar os pesquisadores a planejarem seus experimentos orientados a ameaças à validade. A ValiDEPlan dá suporte a identificação, priorização e controle das ameaças à validade já na fase de planejamento, a fim de reduzir os potenciais riscos e problemas que podem ocorrer desde o planejamento, execução e reporte do experimento.

A ferramenta representa um importante instrumento de verificação e de análise que possibilita ao pesquisador/experimentador novo horizonte a ser contemplado, compreendendo as diversas abordagens de ameaças inseridas nos espaços experimentais controlados, muitos dos quais o pesquisador não consegue, de uma certa forma, vislumbrar durante a fase de execução e também de desenvolvimento, seja pelo seu envolvimento contínuo, efeitos naturais, contato entre participantes ou mesmo pelo esgotamento físico e mental provenientes das tarefas e atividades experimentais, que exigem dos especialistas maior cuidado, atenção e exercício decisório.

6 AVALIAÇÃO EXPERIMENTAL

“O que não pode ser medido não pode ser controlado, e o que não pode ser controlado não pode ser melhorado”.

William Thomson, físico irlandês que desenvolveu a escala Kelvin de temperatura.

6.1 INTRODUÇÃO

Os autores Basili, Selby e Hutchens (1986) iniciaram ampla discussão sobre a necessidade de inclusão e aplicação da avaliação experimental em ES. Ao longo dos anos, autores como Kitchenham *et al.* (2002), Travassos (2003), Jedlitschka, Ciolkowski e Pfahl (2008), Kitchenham *et al.* (2009), Wholin *et al.* (2012), Juristo e Moreno (2013) e outros pesquisadores têm estudado a importância e desenvolvido mecanismos (guias, metodologias, métodos, processos, etc.) de apoio para pesquisa e avaliação em ES que permitem planejar, executar, analisar e reportar resultados de estudos experimentais.

Para definição dos resultados de identificação e mensuração dos dados, bem como, avaliar na prática o impacto do uso do processo PrioriTTVs e da ferramenta ValidEPlan, foram analisados no viés desta avaliação experimental, experimento controlado, questionário e entrevistas, compreendidos a partir do ponto de vista de participantes iniciantes e especialistas em Engenharia de Software Experimental (ESE), conforme o quadro da Tabela 8.

Tabela 8 – Visão geral dos estudos avaliados

Participantes	Estudo Experimental	Objetivo do Estudo	Objeto Avaliado
6 (seis) pesquisadores experientes em ESE	Avaliação do Processo	Avaliar procedimentos para identificação e classificação das ameaças à validade e ações de mitigação	<ul style="list-style-type: none"> ▪ Instrumento de caracterização de um experimento; ▪ Processo PrioriTTVs para a priorização de ameaças e ações de controle.
	Avaliação da Ferramenta	Avaliar percepções positivas e negativas relativas à facilidade	Ferramenta ValidEPlan

Participantes	Estudo Experimental	Objetivo do Estudo	Objeto Avaliado
		e utilidade da ferramenta	
	Entrevista Narrativa	Analisar indícios de funcionalidade, aceitação e satisfação da ferramenta e do Processo	<ul style="list-style-type: none"> ▪ Processo PrioriTTVs ▪ Ferramenta ValidEPlan
20 pesquisadores iniciantes em ESE	Experimento Controlado	Avaliar a corretude da tarefa e o número de ameaças e ações de mitigação da ferramenta proposta em comparação ao uso das práticas <i>ad hoc</i>	Foi avaliada uma unidade experimental para cada tratamento (ValidEPlan vs. Práticas <i>ad hoc</i>).
	Questionário de Avaliação	Verificar a facilidade de uso e percepção de satisfação e aprendizagem da ferramenta	Ferramenta ValidEPlan
	Entrevista Narrativa	Analisar indícios de satisfação dos participantes e percepções positivas e negativas sobre o uso da ferramenta	Ferramenta ValidEPlan

Fonte: Elaboração própria (2019).

A seguir relatamos o planejamento, execução e análise dos dados das avaliações desta tese. Os participantes determinados para fins da pesquisa foram, pois, especialistas em ESE, membros do grupo Empirical Software Engineering Group (ESEG3) e alunos de pós-graduação matriculados na disciplina de ESE do Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE).

³ ESEG – Empirical Software Engineering Group. <https://sites.google.com/site/eportal>.

6.2 AVALIAÇÃO EXPERIMENTAL POR ESPECIALISTAS EM ESE

Com o objetivo de avaliar na prática o processo PrioriTTVs, bem como, a ferramenta ValidEPLan, apresentamos o planejamento, execução e mensuração dos dados de um estudo experimental a partir do ponto de vista de pesquisadores experientes da área de ESE, considerando que este grupo de agentes possui habilidades necessárias para a compreensão dos fenômenos ocorridos em contexto desse estudo e da interferência de ameaças à confiabilidade do mesmo. Nesse viés, os participantes para fins desta pesquisa foram, pois, o total de 6 (seis) especialistas em ESE componentes do grupo ESEG.

A execução do planejamento da avaliação com foco na participação dos especialistas foi desenvolvida entre julho e agosto de 2018, sendo compartimentada em 4 (quatro) etapas principais, conforme pode ser observado através da Tabela 9, abaixo.

Tabela 9 – Etapas do Planejamento da Avaliação com Especialistas

AVALIAÇÃO: ESPECIALISTAS GRUPO ESEG			
1ª Etapa (30/07)	2ª Etapa (31 a 02/08)	3ª Etapa (03 a 05/08)	4ª Etapa (12 a 26/08)
Questionário Demográfico e de Expectativas	Avaliação do Processo proposto	Avaliação da Ferramenta	Entrevista Narrativa Episódica

Fonte: Elaboração própria (2019).

Em relação à **1ª etapa**, estipulou-se, na qualidade de principais objetivos deste estágio de avaliação: obter informações sobre o perfil dos participantes, além de também analisar as expectativas dos especialistas sobre a importância e as características que deverão ser encontradas no objeto dessa atividade.

Por sua vez, o objeto de estudo tratou-se de possível ferramenta utilizada para planejar experimentos da área de ES que auxilia pesquisadores na identificação, priorização e classificação de possíveis ameaças à validade, indicando quais ameaças devem ter prioridade ao serem tratadas e quais são as ações de controle mais assertivas para mitigá-las.

Para validação das informações e resgate de dados para esta etapa, foi, a partir

de critérios pré-definidos sob a ótica da literatura, utilizado um questionário com 10 (dez) indagações de cunho aberto e fechado a fim de identificar informações gerais dos especialistas e os seus olhares e interpretação acerca dos fenômenos em análise neste estudo.

No que compete à **2ª etapa** do estudo em voga, objetivou-se avaliar os procedimentos para a identificação e a classificação das ameaças à validade e as suas respectivas ações para controle, além do mapeamento e classificação de ameaças à validade, por meio da verificação dos seguintes objetos: a) instrumento de caracterização de um experimento; b) Processo proposto PrioriTTVs para a priorização e o controle de ameaças à validade e suas respectivas ações de controle e; c) Ferramenta ValidEPlan.

Para tanto, foi apresentado, com base no levantamento bibliográfico e nas discussões da literatura sobre o tema abordado, um *checklist*, cuja finalidade é identificar principais informações de um determinado experimento da área de ES de acordo com as fases (categorias) de um plano experimental (Seção 4.4).

O instrumento para a identificação das ameaças à validade é composto por 43 questões específicas, divididas em oito categorias, onde o participante foi convidado também a manifestar outras indicações de fator de risco a instrumento apresentado.

Com base nessa abordagem de análise, foi apresentada uma lista preliminar de ameaças, sugeridas a partir das respostas aos questionamentos dos itens do questionário. O especialista deverá, dessa forma, classificar essas ameaças de acordo com uma escala de 1 a 3, considerando a intensidade dos seguintes critérios da matriz de avaliação: Impacto (impacto que o risco poderá causar a resultados do experimento), Urgência (urgência de resolução da ameaça) e Tendência (tendência de evolução da situação de risco). Torna-se válido destacar e justificar que esse modelo de matriz é uma ferramenta que auxilia na priorização de resolução de problemas identificados (CRESWELL, 2010).

Após essa nova etapa, foi apresentado, a especialistas da pesquisa, uma lista de ameaças e, respectivamente, ações de controle. Com as mesmas, o participante classificou estas práticas de controle, de acordo com uma escala de importância, informando quais delas eram mais assertivas ao mitigar ameaças identificadas do processo de investigação e mapeamento de fatores de risco ao experimento.

Além disso, o especialista pôde definir quais delas eram empregadas em seu experimento controlado. Por fim, foram gerados os relatórios contendo todas as

situações identificadas, com proposição sugerida e ações de controle das ameaças.

A **3ª etapa** do planejamento da avaliação objetivou coletar informações de cunho quantitativo e qualitativo ligados às percepções positivas e negativas da ferramenta ValidEPlan (*Validity-Driven Software Engineering Experiments Plan Tool*, objeto de avaliação da etapa) e, também, coletar informações qualitativas (cognitivas e comportamentais) sobre aceitação do referido instrumento no que diz respeito à percepção da facilidade do seu emprego para identificação das ameaças à validade e sua utilidade nas ações de controle dos fatores de risco para os experimentos controlados.

O instrumento de confiabilidade empregado nesta etapa visou à análise e investigação, do ponto de vista dos participantes (especialistas), acerca do uso da ferramenta ValidEPlan, sobre aspectos relevantes do programa, a saber: aspectos gerais do software, os textos e a facilidade de uso do mesmo para os experimentos controlados na área de ES.

Durante a **4ª etapa**, foi executado o roteiro de entrevista narrativa episódica, este com o fim de coletar os depoimentos dos especialistas (participantes), de maneira a analisar indícios de funcionalidade, aceitação e satisfação da ferramenta ValidEPlan e do Processo PrioriTTVs. Para tanto, foram utilizados para a coleta dos dados as seguintes ferramentas: Guia de Entrevista, Gravador de Áudio e nos casos em que a entrevista não pode ser feita presencialmente, utilizamos o aplicativo *Skype*, um software que permite comunicação pela Internet através de conexões de voz e vídeo.

As entrevistas narrativas episódicas definem-se como instrumentos para a coleta de dados acerca da observação de um determinado fenômeno e efeito, visando maior profundidade, de aspectos específicos, a partir das quais emergem pontos de vista do objeto de estudo (LOZARES; VERD, 2008).

Esta abordagem de entrevista narrativa episódica, muito pouco comum dentro de pesquisas experimentais na área de engenharia, enseja encorajar e estimular o participante a revelar resultados sobre o fenômeno observado e experienciado, para suscitar identificação de motivações e também das circunstâncias que são indispensáveis ao desenvolvimento dos fatores em análise (GÜNTHER, 2006).

O roteiro da entrevista, composta por 13 questionamentos, demandam o ponto de vista dos especialistas acerca do processo proposto, o papel das ameaças à validade e da ferramenta ValidEPlan, atividades de identificação e classificação de ameaças e também ações de controle e, por fim, relatórios disponíveis, elementos

constituintes das etapas de planejamento da avaliação.

Objetivando o recorte de novas ameaças identificadas pelos participantes e, a partir de sugestões significativas de novos vieses explorados na pesquisa e sugestões para manutenção da ferramenta ValidEPlan, foi empregado, no viés das transcrições das falas reproduzidas no viés da entrevista narrativa episódica, a técnica metodológica de Análise de Conteúdo (BARDIN, 2009).

A Análise de Conteúdo é uma metodologia de pesquisa usada para descrever e interpretar o conteúdo de toda classe de documentos e textos. A análise, conduzindo descrições sistemáticas, qualitativas ou quantitativas, contribui para interpretação de mensagens e atingimento da compreensão de seus significados (BARDIN, 2009). Essa metodologia constitui-se em mais do que uma simples técnica de análise dos dados coletados em entrevistas ou outros recursos exploratórios, representando uma técnica que possibilita que o pesquisador explore abordagens diferenciadas e estabeleça categorias de análise.

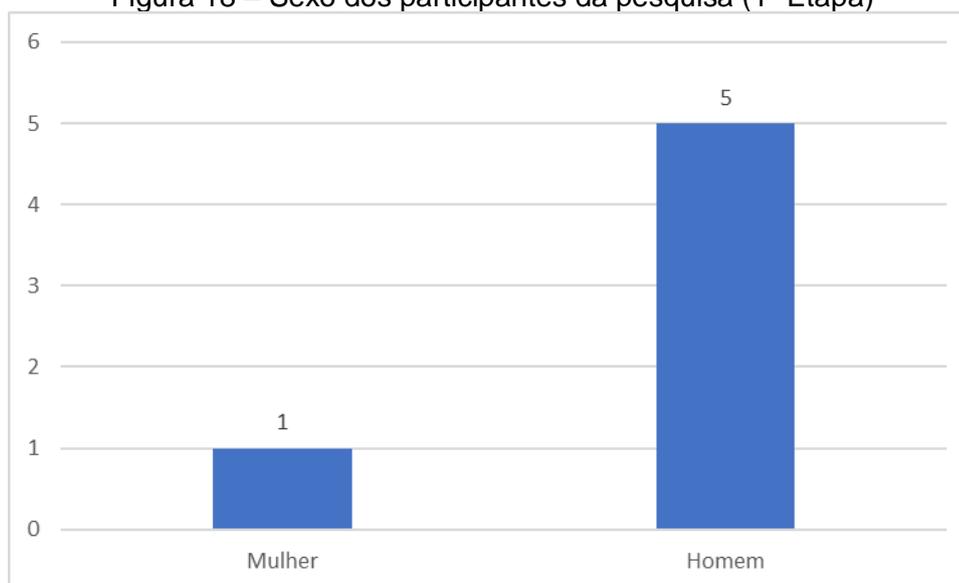
6.2.1 ETAPA 1: Análise dos Dados Demográficos e de Expectativas

O questionário para a coleta de dados de informações demográficas e de expectativas foi conduzido com 6 (seis) participantes (especialistas) no campo de ESE componentes do grupo ESEG – *Empirical Software Engineering Group*, um grupo de pesquisadores envolvidos com estudos experimentais em ES, organizando e publicando um corpo de conhecimentos para auxiliar e apoiar instituições e os profissionais na tomada de decisões sobre a avaliação de novas tecnologias em ES.

O grupo ESEG é multi-institucional formado por pesquisadores de mestrado e doutorado do Centro de Informática da UFPE, Universidade Federal da Paraíba (UFPB), Universidade de Pernambuco (UPE), Instituto Federal de Pernambuco (IFPE), Instituto Federal do Sertão Pernambucano (IF Sertão-PE) e Instituto Federal Baiano (IF Baiano).

Com base na identificação de questões, observaram-se resultados percentuais abaixo relacionados, seguidos de aprofundamento teórico, quando necessário, a fim de elucidar, sob a perspectiva de outros estudiosos da área de ES, as várias ações e percepções manifestadas pelos respondentes do estudo.

Figura 18 – Sexo dos participantes da pesquisa (1ª Etapa)



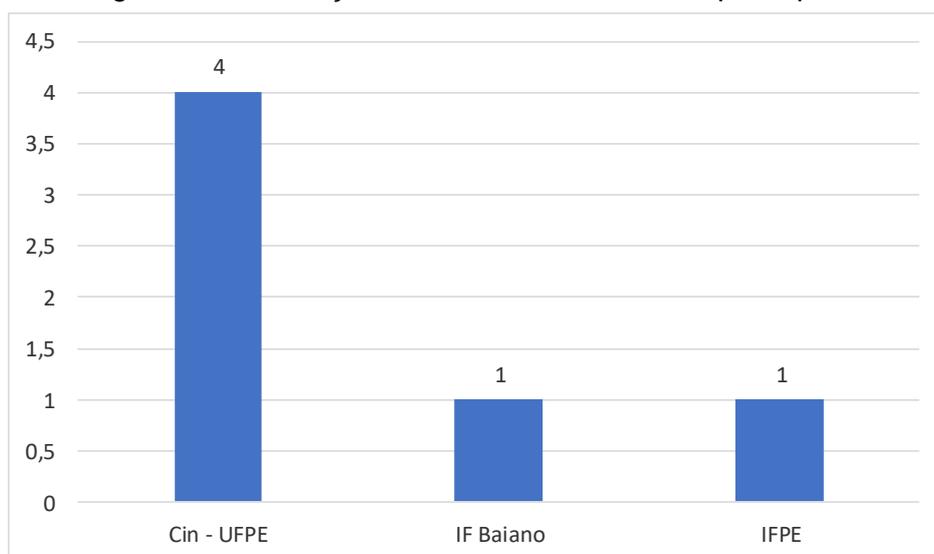
Fonte: Dados da pesquisa (2019).

Primeiramente, em relação ao sexo dos participantes, notou-se que, em sua grande maioria, eram de homens (5) em detrimento de mulheres (1), conforme o gráfico da Figura 18, acima. Cada vez mais, tem-se observado maior participação e presença de pesquisadores do sexo masculino em comparação ao feminino nas áreas de Computação, o que reforça o resultado percentual deste estudo, em caráter inicial.

Em relação às idades dos participantes, observou-se intervalo entre 28 e 37 anos de idade, evidenciando um público-alvo de especialistas com um perfil ainda jovem, com potencial de desenvolvimento de competências, ações e habilidades, em fase de contribuições para o mercado ativo de trabalho.

Quanto à instituição de ensino ou trabalho em que se encontram os participantes da pesquisa, o gráfico da Figura 19, abaixo, revelou que quatro dos participantes pertencem ao quadro do CIn-UFPE, evidentemente esperado, considerando que, em parte, o estudo de investigação com especialistas admitiu, nesse sentido, a escolha dos participantes por intencionalidade ou por conveniência, por facilidades quanto ao acesso aos especialistas (CRESWELL, 2010).

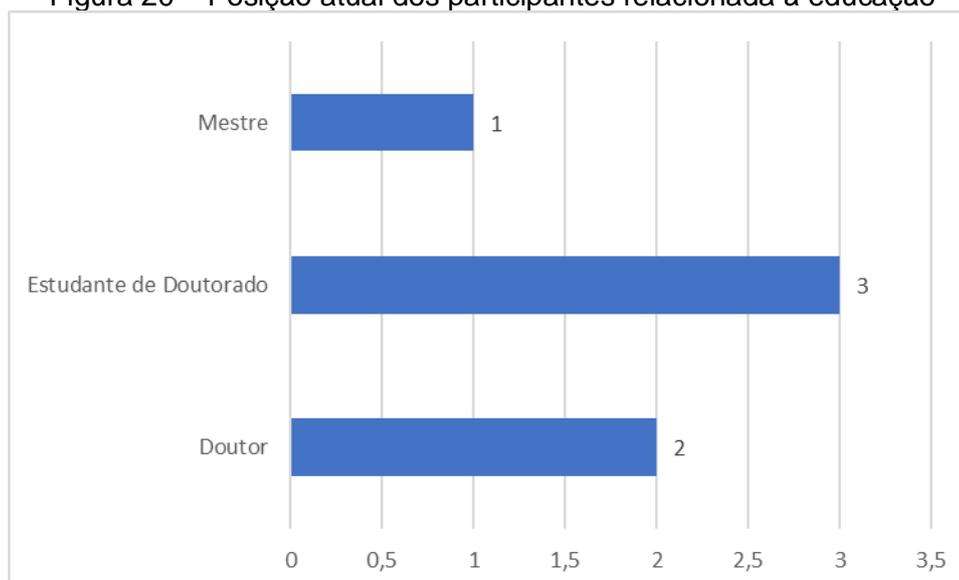
Figura 19 – Instituição de ensino ou trabalho do participante



Fonte: Dados da pesquisa (2019).

Por sua vez, um dos participantes pertencem ao quadro profissional do Instituto Federal de Pernambuco (IFPE) e outro ao Instituto Federal Baiano (IF Baiano), apontando uma participação e engajamento de outras Instituições Federais de Ensino Superior (IFES), com potencial de desenvolvimento de possíveis parcerias para a análise de ações, monitoramento e resultados na execução de experimentos controlados, com fim acadêmico ou não.

Figura 20 – Posição atual dos participantes relacionada à educação

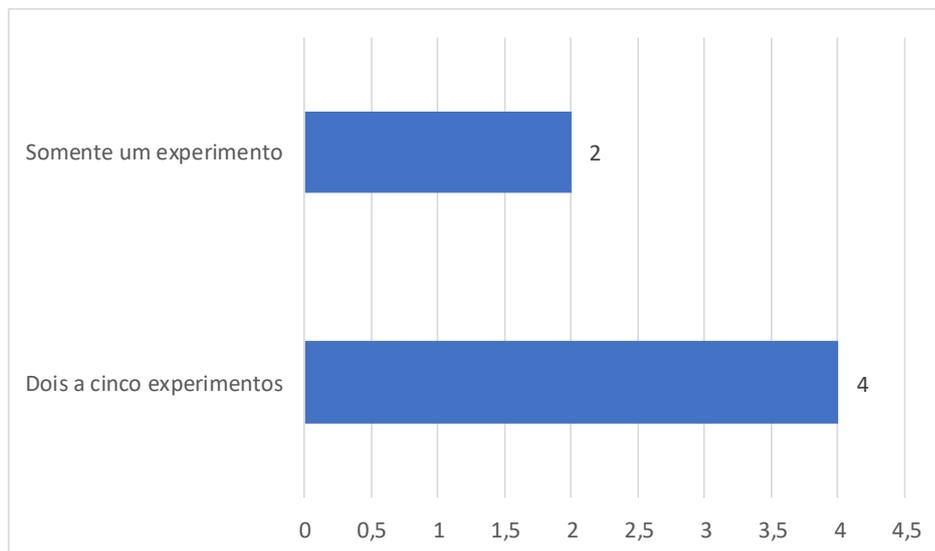


Fonte: Dados da pesquisa (2019).

Sobre a posição atual dos participantes em relação à educação e também à formação acadêmico-profissional, conforme exposição a partir do gráfico da Figura

20, evidenciou-se que três dos participantes encontram-se no doutorado, dois são doutores e um possui o título de mestre.

Figura 21 – Total de experimentos que o especialista participou do planejamento

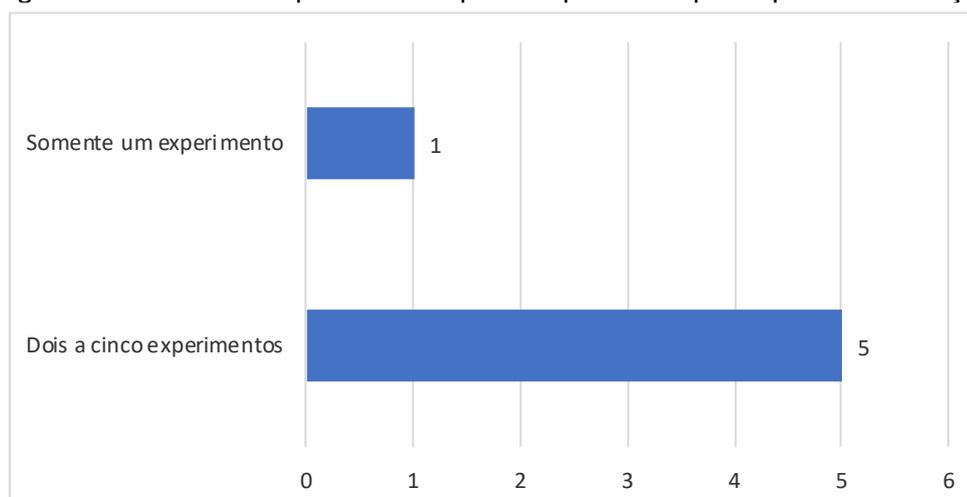


Fonte: Dados da pesquisa (2019).

No que compete ao número de experimentos, na perspectiva da ES, em que o especialista participou da atividade de planejamento, de acordo com o gráfico da Figura 21, dois participantes se engajaram em apenas um experimento e quatro, entre dois e cinco experimentos científicos.

O percentual indica que existe, ainda, uma deficiência no número de pesquisadores com um interesse na participação na ação de planejamento dos experimentos em ES, tendo em vista, como discutido em uma seção anterior, a significância do planejamento das ações e dos procedimentos da pesquisa (antes, durante e após a execução) para a identificação de um consequente controle das ameaças à validade e confiabilidade de resultados experimentais (MUYLAERT *et al.*, 2014; FREIRE, 2015).

Figura 22 – Total de experimentos que o especialista participou da execução

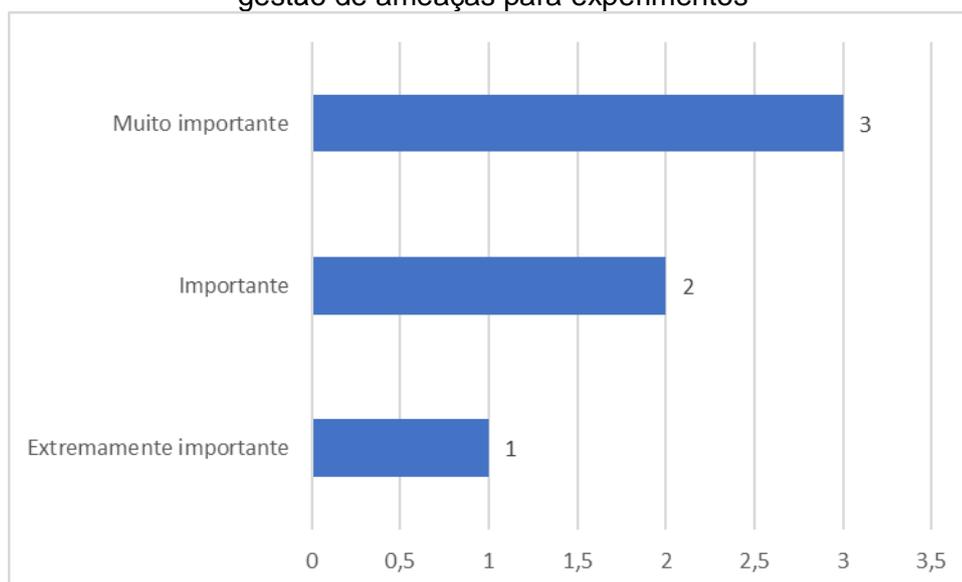


Fonte: Dados da pesquisa (2019).

Quanto ao gráfico da Figura 22, acima, cinco dos especialistas apontaram interação e engajamento na fase executória de dois a cinco experimentos, ao passo que um participou da observação de somente um experimento. Igualmente, é a fase de execução um importante passo para o desenvolvimento de um experimento científico, uma vez que é a partir dessa que são analisados a coesão das variáveis, a sua sistematização em relação aos parâmetros estabelecidos previamente e os comportamentos que se manifestam na consecução de resultados, podendo, com estes, coincidir ou até mesmo divergir, na presença de algumas ameaças à validade do estudo (LIMA; NETO; EMER, 2014).

É, em outra ótica de entendimento, no processo de planejamento experimental que o pesquisador deve analisar e definir ameaças à validade do seu experimento a partir da percepção de quatro tipos de validade: interna, externa, de construção e de conclusão. A eventual existência de determinada ameaça que não foi previamente definida, com apresentação da ação corretiva, pode comprometer os resultados do experimento, e com isso invalidar o estudo como um todo (MAFRA; TRAVASSOS, 2006; WOHLIN *et al.*, 2012).

Figura 23 – Nível de importância que o pesquisador oferece à ferramenta de planejamento e gestão de ameaças para experimentos



Fonte: Dados da pesquisa (2019).

No que compete à percepção de importância dos participantes em relação às ferramentas de planejamento e gestão de ameaças para os experimentos, o gráfico da Figura 23 demonstrou os seguintes resultados: Três dos participantes consideraram os instrumentos muito importantes, dois consideraram importantes e um enxerga esses recursos como de extrema relevância.

Os resultados supramencionados, mais uma vez, apontam a tendência no emprego, por parte de pesquisadores, de práticas de planejamento de experimentos existentes, considerando as suas potencialidades e as contribuições na gestão de ameaças e correções necessárias à validação dos resultados. Apesar de novas, as principais técnicas e abordagens de planejamento de experimentos e gestão de ameaças à validade são remotas e, em termos de eficácia, poderiam estar sendo sistematicamente aplicadas na concepção e desenvolvimento de projetos há muitos anos.

Contudo, devido à quantidade de dados e informações, além de um alto número de possibilidades de problemas que podem acontecer no planejamento, execução e análise de uma avaliação experimental e até mesmo do tratamento de uma ameaça poder implicar no surgimento de novas ameaças não identificadas inicialmente, há um desestímulo dos pesquisadores em muitos campos do conhecimento (MUYLAERT *et al.*, 2014; FREIRE, 2015; TEIXEIRA; FONSECA; SOARES, 2018).

Ao mesmo tempo em que se isola o planejamento como ponto valoroso para o atingimento de resultados eficientes num experimento, pouca ênfase também é

conferida, nesta etapa ou em posteriores, à investigação das ameaças à validade de resultados de estudos experimentais. Percebe-se, dessa forma, no campo estrutural da investigação experimental, uma carência conceitual na definição de um espaço para a identificação de ameaças à validade, ainda mais que este fator reflete no desempenho e impacto dos estudos (FREIRE, 2015; PIMENTEL, 2018).

As etapas da investigação experimental, independente da forma adotada, são: Definição das variáveis de controle e resultados, Planejamento, Avaliação, Execução, Análise e Interpretação de resultados auferidos e, por fim, realizada em paralelo, Empacotamento. A Definição e Planejamento são referências para as etapas seguintes, e, aqui, deveriam ser introduzidos meandros de um mapeamento de ameaças e correções admissíveis a cada fator que possa ser identificado na fase executória (MAFRA; TRAVASSOS, 2006; CORTÉS *et al.*, 2018).

É no procedimento de planejamento do experimento onde o pesquisador atenta-se a como o estudo experimental será conduzido, pois, aos seus estágios de desenvolvimento e sistematização de resultados (WOHLIN *et al.*, 2012).

Conforme notado, da estrutura de investigação experimental apresentada, preocupa-se mais com desenho do processo e a chegada do resultado do que com a identificação de eventuais barreiras à validade e possíveis formas de saneamento e correções das falhas observadas na execução.

No decorrer desta pesquisa com especialistas (participantes), foi questionado sobre quais requisitos funcionais e não funcionais deveriam ser incluídos às práticas executáveis da ferramenta ValidEPlan, com objetivo de torná-la mais produtiva às ações com experimentos da área de ES.

Em condições sugestivas, foram elencados os seguintes requisitos funcionais: Cadastro de banco de dados com as ameaças à validade mais conhecidas, procedimentos para identificar e caracterizar ameaças; cadastro de novas ameaças observadas na fase de execução experimental; cadastro dos critérios que tornem a identificação e priorização mais robustas; especificação amigável do experimento; gerar recomendações de ameaças a validade e suas mitigações baseado no método de pesquisa e ao tópico sob investigação; sugerir ameaças de acordo com as informações do experimento; listar as ameaças por grupos/tipologias; e listar as soluções corretivas por ameaças e; listar as soluções por grupos de ameaças.

No que compete aos requisitos não funcionais: Estar disponível na web; capacidade de integração facilitada com outras ferramentas; facilidade de uso

(usabilidade); bom desempenho da ferramenta e portabilidade. Quando indagados sobre as contribuições da ferramenta ValidEPlan para a maximização da qualidade e validade sobre o resultado decorrente da investigação experimental, os especialistas apontaram os seguintes impactos:

Porque aumenta a probabilidade de que o experimentador fará a identificação e irá se planejar para ameaças à validade do experimento (E1).

Poderia aumentar se auxiliar a identificar determinadas ameaças e ações de controle que pudessem, em primeiro momento, passar despercebido pelo pesquisador (E2).

Em geral, avaliar ameaças é um passo complicado durante um experimento, em que existe uma subjetividade da escolha por parte do pesquisador, uma ferramenta que dê suporte, elimina uma boa parte do viés, o que deixaria o resultado bem mais confiável (E3).

Existem outras formas de avaliar e identificar ameaças a validade do experimento, mas uma ferramenta que auxilia nesse processo por facilitar a identificação de ameaças e soluções, bem como aumentar a completude dessa avaliação. Então de alguma forma essa ferramenta aumenta as possibilidades de tratar a validade dos resultados do experimento, principalmente para pesquisadores iniciantes e menos experientes (E4).

Quando questionados sobre se a existência de uma ferramenta que dê suporte à identificação e priorização de determinadas ameaças à validade de acordo com características e/ou objetivos do experimento da área de Engenharia de Software (ES) aumentaria, dessa forma, a qualidade e/ou validade dos resultados de um estudo experimental conduzido, observou-se o percentual de 83% que evidenciaram concordância total.

Por sua vez, indagados sobre se a existência de uma solução que dê suporte à identificação e priorização de determinadas ameaças à validade de acordo com as características e/ou objetivos de um experimento da área de ES promoveria melhor e correta interpretação de resultados do estudo, observou-se que 67% concordaram totalmente com esta contribuição significativa do instrumento para investigação experimental.

6.2.2 ETAPA 2: Avaliação do processo PrioriTTVs

6.2.2.1 Identificação às ameaças à validade

Conforme aprofundamento teórico-conceitual do tópico anterior, observou-se uma relação casuística entre fatores de risco à validade de um experimento. Isto é, a

presença de ameaças perceptíveis ou não na investigação experimental tende, pois, a prejudicar confiabilidade dos dados e resultados, implicando em comprometimento de todo o estudo e os impactos sociais (acadêmicos e industriais).

Para melhoria na identificação das ameaças à validade dos experimentos, foi disponibilizado, aos participantes desta etapa da pesquisa (especialistas), o quadro abaixo exposto (quadro da Tabela 10), para que os mesmos se posicionassem diante da percepção acerca da presença desses fatores de risco aos projetos experimentais e, por sua vez, se os mecanismos e ações corretivas sugeridas eram compatíveis para a superação dessas ameaças. O quadro da Tabela 10 é um conjunto de ameaças à validade organizadas por tipo de ameaça (interna, externa, constructo e conclusão), identificadas através de uma revisão sistemática da literatura realizada por Neto & Conte (2013). As ameaças e suas

Esses dados, juntamente com o modelo estrutural de planejamento de um experimento, definido por Fonseca (2016), propiciaram a elaboração dos itens de verificação quanto a ameaças à validade (*checklist*) (quadro da Tabela 11, abaixo).

Tabela 10 – Classificação das ameaças à validade experimental identificadas por tipologia

TIPOLOGIA	AMEAÇA À VALIDADE EXPERIMENTAL
INTERNA	<p>Diferenças entre os participantes relacionadas ao nível de experiência</p> <p>Familiaridade com o material experimental</p> <p>Participante pode adquirir conhecimento durante o experimento</p> <p>Diferenças no material experimental</p> <p>Precisão dos dados coletados</p> <p>Diferenças entre os participantes relacionadas ao nível de habilidade</p> <p>Participantes podem abandonar o experimento</p> <p>Comunicação entre os participantes durante o experimento</p> <p>Falta de motivação dos participantes</p> <p>Problemas de especificação do material experimental</p> <p>Fatos do cotidiano podem afetar o desempenho dos participantes durante os intervalos</p> <p>Comunicação entre os participantes durante os intervalos</p> <p>Treinamento desbalanceado</p> <p>Idioma dos participantes diferente do idioma do material experimental</p> <p>Seleção dos participantes com base em dados históricos</p> <p>Aplicação de atividade pré-experimento pode influenciar o desempenho do participante</p> <p>No caso de replicação, alteração do pacote experimental do estudo original</p> <p>Estabelecer tempo para execução do experimento</p> <p>Diferenças entre as tarefas experimentais</p> <p>Rivalidade entre os grupos</p> <p>Os participantes podem ter realizado estudos semelhantes anteriormente</p> <p>Ordem de apresentação do material experimental para os participantes</p> <p>Falta de conformidade com o processo a ser seguido</p> <p>Treinamento pode ser insuficiente</p> <p>Atribuição de recompensa pela participação e não pelo desempenho</p> <p>Fatores técnicos podem causar dificuldade aos participantes</p> <p>Treinamento conduzido por pessoas diferentes</p> <p>Diferenças entre os participantes relacionadas aos gêneros</p> <p>Treinamento conduzido por apenas uma pessoa</p> <p>Coleta de dados incompleta ou incorreta</p> <p>Os participantes podem procurar respostas em outras fontes</p> <p>Parte dos participantes recebem recompensa e outra parte não</p> <p>Diferenças entre os participantes relacionadas ao nível de personalidade</p> <p>Design experimental inadequado aos objetivos do experimento</p>
	Representatividade dos participantes

EXTERNA	<p>Representatividade dos artefatos. Representatividade do domínio dos artefatos Representatividade do processo Representatividade das tarefas Representatividade dos tratamentos. Representatividade dos defeitos. Representatividade dos requisitos Representatividade do ambiente experimental. Representatividade do contexto da pesquisa e/ou resultados Replicação do estudo em contextos similares</p>
CONSTRUCTO	<p>Tratamento pode não representar a causa. Métrica pode não representar o efeito. Interação entre os tratamentos. Participantes se comportam de maneira diferente quando são observados. Influência do pesquisador. Adivinhação das hipóteses experimentais. Utilização de apenas uma métrica. Utilização de apenas um artefato experimental, ou tarefa, ou tratamento. Os participantes podem preencher o questionário de forma incorreta ou inverídica. Métricas de caracterização dos participantes podem não estar corretas</p>
CONCLUSÃO	<p>Confiabilidade das medidas. Aplicação de teste estatístico inadequado. Amostra heterogênea. Amostra de tamanho pequeno. Confiabilidade na implementação do tratamento. Seleção de dados para satisfazer as hipóteses Problemas aleatórios que podem ocorrer no ambiente experimental. Intervalo de confiança inadequado. Confiabilidade das variáveis</p>

Fonte: Neto e Conte (2013).

Tabela 11 – Lista de verificação por categoria em instrumento de coleta quanto à avaliação do processo

CATEGORIAS	LISTA DE VERIFICAÇÃO
OBJETIVOS (C1)	Interação com os participantes durante a execução do experimento (Q1)
HIPÓTESES, VARIÁVEIS E MEDIDAS (C2)	Conhecimento prévio das hipóteses do experimento (Q2) Existência de guia, método ou protocolo para medição de dados (Q3) A(s) métrica(s) definida(s) não representa(m) o que se pretende(m) medir (Q4) Métrica única a ser usada para responder às perguntas de pesquisa (Q5) O intervalo de confiança não é usado para tirar conclusões (Q6)
PARTICIPANTES (C3)	Diferentes níveis de experiência ou habilidades dos participantes em qualquer tarefa durante o experimento (Q7) Participantes com diferentes gêneros, personalidades ou quaisquer outras características (Q8) Participantes com línguas nativas diferentes da linguagem do material experimental (Q9) Participante que lidará com mais de um tratamento (Q10) Recompensa aos participantes para a realização do experimento (Q11) Participação dos participantes em experimentos semelhantes no passado (Q12) Falta de motivação do participante para realização do estudo (Q13) Opção para o participante deixar a pesquisa em qualquer estágio (Q14) Comunicação entre os participantes durante o experimento (Q15) Coleta de dados dos participantes via caracterização e/ou questionário demográfico (Q16) Falta de representatividade dos participantes como população-alvo (Q17) O tamanho pequeno da amostra selecionada (Q18) Eventos que influenciam o desempenho dos participantes podem ocorrer no ambiente experimental (Q19)
MATERIAIS EXPERIMENTAIS E TAREFAS (C4)	Os artefatos e procedimentos experimentais não representar aqueles usados no contexto real (Q20) Os procedimentos experimentais são diferentes para qualquer grupo de participantes (Q21) Existência de documento de requisitos entre os artefatos experimentais do experimento (Q22) Possibilidade de inserir defeitos nos artefatos, se o contexto da experiência for inspeção (Q23) As tarefas que serão executadas pelos grupos são distintas (Q24) As tarefas serão executadas por pesquisadores (Q25)
DESIGN EXPERIMENTAL (C5)	Os tratamentos poderiam não representar os tratamentos usados no contexto real (Q26) O uso de um tratamento poderia influenciar outros tratamentos usados no experimento (Q27) Cada grupo utilizará apenas um artefato, tarefa ou tratamento experimental (Q28)
PROCEDIMENTO (C6)	O ambiente em que o experimento será conduzido não representa o contexto real do estudo (Q29) O experimento será realizado em diferentes ambientes experimentais (Q30) Haverá intervalos entre as seções de experimentos (Q31) O treinamento dado aos participantes e/ou grupos será dado por pessoas diferentes em diferentes momentos (Q32) Algum dos participantes podem não entender ou não ter conhecimento suficiente do material experimental (Q33) Poderia o estudo piloto não ser suficiente para apontar possíveis problemas com a execução do experimento (Q34)

	<p>As atividades experimentais têm uma restrição de tempo para sua execução (Q35)</p> <p>Os eventos e a ordem de apresentação do material experimental para os participantes descreveram (Q36)</p>
PROCEDIMENTO DE COLETA DE DADOS E ANÁLISE (C7)	<p>Os dados coletados podem ser incompletos, precisos e/ou confiáveis (Q37)</p> <p>Você planeja incluir pesquisadores observadores durante o experimento (Q38)</p> <p>Algum conjunto de procedimentos, protocolos ou métodos será usado para análise de dados (Q39)</p> <p>Algum procedimento será usado para resolver possíveis desacordos na análise de dados (Q40)</p> <p>Os procedimentos e critérios que suportam o rigor estatístico do estudo descrito ou referenciado (Q41)</p>
DOCUMENTO (C8)	<p>Se o experimento é uma replicação, o pacote experimental é modificado a partir do estudo original (Q42)</p> <p>Você planeja apresentar ou indicar onde os dados brutos de sua experiência estarão disponíveis (Q43)</p>

Fonte: Elaboração própria, baseado em Fonseca (2016).

O instrumento de coleta de dados (Tabela 11) utilizado para análise de caracterização para planos experimentais da área de ES, pois, tratou-se da elaboração de uma lista de verificação (*checklist*), com a finalidade de capturar as principais informações do experimento, baseado na estrutura do plano experimental proposto por Fonseca (2016), em que dividem-se os problemas identificados em 8 (oito) categorias específicas e tangíveis às investigações experimentais.

Para cada resposta orientada ao desafio identificado (Tabela 11), algumas propostas de ameaças à validade e também as ações para os validar são sugeridas, a partir do relacionamento entre as respostas fornecidas ao *checklist* e o banco de dados de ameaças utilizado (Tabela 10). Nesse tocante, os 6 (seis) especialistas foram convidados a responder a este instrumento de coleta e validação dos dados a fim de se compreender suas percepções e possíveis sugestões e comentários acerca da avaliação do processo de identificação de ameaças à validade. As possibilidades de resposta foram: a) Concordo; b) Concordo Parcialmente, Por que? e c) Discordo, Por que?

Para identificação e facilitação para a análise dos resultados da avaliação, buscou-se considerar as categorias como C1, C2, C3... C8 e as questões da lista de verificação (*checklist*), de forma correspondente a cada categoria, sequencialmente, Q1, Q2, Q3... Q43.

De maneira sintetizada, em relação a grupos de resultados manifestados pelos participantes deste estudo acerca da avaliação do processo, por meio do instrumento de coleta de dados supramencionado, observou-se que, em relação a número totalizado da lista de verificação (*checklist*), com 43 questionamentos em geral, poucos foram os que, a partir da percepção dos especialistas apresentaram concordância de todos quanto às ameaças e possíveis medidas de saneamento (correção e controle) adequadas ao fator de risco à validade identificado, sendo o percentual de apenas 39,5%, mais uma vez, a categoria de Participantes (C3) a que mais contribuiu, pois, para a constituição do percentual total, com 29,4%.

O percentual de 39,5% evidencia que, apesar das manifestações da totalidade de itens da lista de verificação (*checklist*), a fim de se compreender, ao máximo, o número de casos e situações que constituem um objeto de ameaça à validade dos resultados dos experimentos, o instrumento ainda apresenta problemas desafiadores quanto à correta e compatível associação entre as ameaças e as ações corretivas (GABRIEL, 2010; WOHLIN *et al.*, 2012).

Em relação ao agrupamento de itens que receberam, dos especialistas participantes, concordâncias parciais, observou-se que 32,5% foi o percentual total de variáveis do *checklist*, de modo que, desse valor, 64,3% foram os itens em que pelo menos 2 (dois) participantes se mostraram concordando parcialmente com ameaças e as ações corretivas listadas e 35,7% foi o percentual em que 3 (três) ou mais especialistas concordaram parcialmente em relação aos itens propostos.

Quanto ao resultado total, de 32,5%, evidencia-se que, de fato, é preciso que o instrumento de coleta de dados, construído a partir do modelo de Fonseca (2016), seja, em si submetido a revisões de coesão e coerência textuais a fim de se evitarem ambiguidades ou dúvidas ligadas ao procedimento de análise, entre outros pontos de observação.

Quanto aos participantes com discordância às questões apresentadas dentro do esboço estrutural do instrumento de coleta de dados (*checklist*), em relação ao total de itens, observou-se que foi baixo o percentual de especialistas com esse posicionamento, sendo de 2,3%, que reforça o entendimento de que, de fato, os especialistas concordaram com a abordagem e critérios utilizados no processo para identificação de ameaças e ações de controle, no entanto, existe a necessidade de ajustes no instrumento de coleta apresentado, a fim de se explorar com uma maior fidedignidade a presença de riscos à validade da investigação experimental.

6.2.2.2 Classificação das ameaças à validade

Para este estágio, exigiu-se do participante a sua percepção quanto ao modelo proposto para classificação das ameaças à validade dos experimentos. Da mesma forma que a avaliação anterior, o participante foi inquirido a responder se concorda, discorda e caso necessário, apresentasse sugestões e comentários. Uma instância do modelo utilizado para classificação de ameaças à validade é mostrada conforme a Figura 24.

Figura 24 – Modelo de avaliação das ameaças sugeridas

Lista de Ameaças Sugeridas – Experimento X				C	D	Sugestões e Comentários
Influência do Pesquisador: Ocorre quando os participantes podem ter o desempenho afetado quando são observados por outras pessoas durante o experimento. Tipo: Constructo						
Critérios	Intensidade			Impacto → Impacto que a ameaça poderá causar aos resultados do experimento. Urgência → Urgência de resolução da ameaça. Tendência → Tendência de evolução da situação de risco identificada.		
Impacto						
Urgência						
Tendência						

Fonte: Elaboração própria (2019).

Foi, dessa maneira, apresentada lista preliminar de ameaças, sugeridas a partir das respostas aos questionamentos dos itens do questionário avaliado na Etapa 1. O experimentador precisa, pois, classificar as ameaças (*checklist*) conforme escala de 1 a 3, considerando a intensidade das seguintes variáveis de observação no experimento: Impacto (impacto que a ameaça poderá causar aos resultados do experimento), Urgência (urgência de resolução da ameaça) e Tendência (tendência de evolução da situação de risco identificada), recurso mais conhecido como matriz de prioridades, para alguns autores da literatura.

A matriz de prioridades utilizada é um recurso que auxilia na priorização de resolução de problemas identificados em um campo específico. A análise por meio dessa ferramenta é muito utilizada naquelas questões em que é preciso de uma orientação para tomar decisões complexas e que exigem uma análise de vários problemas, inclusive variáveis de ameaças à validade, por isso, a aplicabilidade no âmbito do estudo (SELEME; STADLER, 2008).

Logo, ameaças à validade definidas na etapa anterior serão agrupadas de acordo com a escala de prioridade, pela combinação dos valores de impacto (I), urgência (U) e tendência (T). Por meio da atribuição de valores (1 a 3) para as possíveis ameaças apresentadas e com possibilidades de gerar riscos a experimentos controlados, os especialistas evidenciaram seus pontos de vista sobre questões de priorização dos fatores de risco à validade dos resultados de estudos experimentais.

Para determinarmos quais ameaças são mais ou menos críticas, definimos uma escala de magnitude, criada a partir da combinação dos parâmetros de impacto, urgência e tendência, utilizados para gerar a lista de classificação das ameaças.

Figura 25 – Escala de magnitude

Agrupamento de Ameaças – Por Magnitude									C	D	Sugestões e Comentários			
As ameaças classificadas anteriormente serão agrupadas de acordo com escala de prioridade, pela multiplicação dos valores de impacto (I), urgência (U) e tendência (T). Abaixo segue as combinações:														
Muito Alta			Alta			Moderada			Baixa			Muito Baixa		
I	U	T	I	U	T	I	U	T	I	U	T	I	U	T
3	3	3	3	2	2	2	2	2	2	2	1	2	1	1
3	3	2	2	3	2	3	2	1	2	1	2	1	2	1
3	2	3	2	2	3	3	1	2	1	2	2	1	1	2
2	3	3	3	3	1	2	3	1	3	1	1	1	1	1
			3	1	3	2	1	3	1	3	1			
			1	3	3	1	3	2	1	1	3			
						1	2	3						

Fonte: Elaboração própria (2019).

A instância da escala de magnitude (Figura 25) acima, foi elaborada a partir de uma adaptação de uma matriz de Garvey e Lansdowne (1998), que apresentaram uma abordagem estruturada que identifica quais riscos são mais ou menos críticos.

Quanto à avaliação da metodologia para classificação das ameaças para a confiabilidade dos experimentos, questionou-se o seguinte ponto: Você concorda com os procedimentos (matriz de prioridades e escala de magnitude), usados para classificação de ameaças à validade?

Nesse sentido, 100% dos participantes concordaram com a abordagem e critérios empregados para a taxonomia das ameaças e de fatores de risco às investigações experimentais, resultado que reforça e implica que o recurso utilizado, de fato, denota, nesses modelos de estudos experimentais e das práticas de controle, uma maneira simplificada e facilitada de compreender e também identificar essas variáveis fundamentais ao desenvolvimento dos estudos experimentais.

Apesar do grau de concordâncias nos questionamentos, torna-se imperioso destacar que, ainda assim, o instrumento para validação dos critérios de classificação das ameaças no viés de experimentos não ficou imune a sugestões e críticas pertinentes, em especial acerca da formulação de maiores esclarecimentos a fim de promover maior coerência e a segurança nas hipóteses definidas, gerando, evidentemente, dados mais claros e mais confiáveis para fins de investigação de um cenário experimental (GABRIEL, 2010; WOHLIN *et al.*, 2012; FREIRE, 2015).

6.2.2.3 Classificação das ações de controle

Para esta etapa da avaliação, após identificação das principais ameaças à validade e confiabilidade dos resultados experimentais, foi apresentada, junto com o *checklist*, as respectivas ações de controle e saneamento, considerando cada caso e, também, os possíveis efeitos e reflexos da aplicabilidade das medidas corretivas, uma vez que, quando há relacionamentos entre ações de controle e ameaças à validade, a aplicação de uma ação de mitigação, tende a desencadear uma nova fonte de ameaça.

Durante esse estágio, o experimentador foi convidado a classificar estas ações de controle de acordo com uma escala de importância, informando quais eram mais assertivas ao mitigar de forma apropriada essas ameaças, sem, contudo, gerar novo fator de risco ao projeto experimental.

Além disso, o participante poderia, com base no mesmo instrumento, definir quais das práticas corretivas apresentadas eram aplicadas em experimento e que modelos de contribuições eram percebidas quanto ao atingimento dos resultados mais fidedignos (MAFRA; TRAVASSOS, 2006; MUYLAERT *et al.*, 2014; FREIRE, 2015), de acordo com o modelo da Figura 26, abaixo.

Figura 26 – Modelo de avaliação das ações de controle do experimento

Lista de Ameaças e Ações de Controle – Experimento X		C	D	Sugestões e Comentários
Exemplo: Influência do Pesquisador : Ocorre quando os participantes podem ter o desempenho afetado quando são observados por outras pessoas durante o experimento. Tipo: Constructo				
Ações de Controle:				
1. <input type="checkbox"/> Incluir pesquisadores independentes ao experimento.				
<input type="checkbox"/> Não é importante	<input type="checkbox"/> Pouco importante			
<input type="checkbox"/> Importante	<input type="checkbox"/> Muito importante			
<input type="checkbox"/> Extremamente importante				
2. <input type="checkbox"/> Não permitir que pesquisadores interessados nos resultados da pesquisa possam interagir com os participantes.				
<input type="checkbox"/> Não é importante	<input type="checkbox"/> Pouco importante			
<input type="checkbox"/> Importante	<input type="checkbox"/> Muito importante			
<input type="checkbox"/> Extremamente importante				

Fonte: Elaboração própria.

As ações de controle (corretivas) são abordagens que adquirem a capacidade de agir sobre as ameaças à validade identificadas dos resultados do experimento controlado, gerando mudanças no comportamento e sistematização (arranjo) dos elementos constituintes de forma a preservar retornos e contribuições fidedignas.

Logo, medidas corretivas constituem-se em práticas de saneamento, refletidas como eficientes no contexto dos projetos experimentais, controlados ou não, mas que podem resultar e também implicar em novas problemáticas aos resultados do plano experimental (FREIRE, 2015; TEIXEIRA; FONSECA; SOARES, 2018).

Dessa maneira, na investigação com os participantes (especialistas) no viés da presente pesquisa, questionou-se concordância sobre avaliar níveis de importância da ferramenta de ações de controle de ameaças à validade que foram identificadas durante as fases anteriores já discriminadas. 83,3% deles concordaram plenamente sobre a relevância do instrumento de avaliação das práticas de controle em investigação com experimentos controlados, ao passo que 16,7% concordam parcialmente com este recurso.

Em relação a sugestões para melhoria e ajustes no instrumentos, seguiram-se os mesmos argumentos das ferramentas anteriores, no sentido de formular o pesquisador deste estudo maior esclarecimento sobre os questionamentos, além de se abrir um campo para que os pesquisadores ao utilizar tal abordagem também possam sugerir novas ações de controle, observadas na prática cotidiana dos experimentos controlados em ES.

6.2.3 ETAPA 3: Avaliação da ferramenta ValidEPlan

Após avaliarmos o processo PrioriTTVs proposto para identificar e classificar ameaças à validade e suas respectivas ações de mitigação, a 3ª etapa da avaliação teve como objetivo coletar informações quantitativas e qualitativas relacionadas às percepções positivas e negativas dos especialistas ao usar a ferramenta ValidEPlan (objeto de avaliação da etapa).

A literatura oferece inúmeras possibilidades e teorias responsáveis por analisar a influência da tecnologia no comportamento humano, contudo no âmbito desta avaliação optou-se pelo modelo de aceitação de tecnologia: *Technology Acceptance Model* (TAM), proposto por Davis (1989). Esse modelo de aceitação de tecnologia é

uma teoria de sistema de informação, que modela como os usuários aceitam e usam uma tecnologia.

Dessa forma, propomos avaliar informações coletadas a partir de percepções cognitivas e comportamentais sobre aceitação da ferramenta ValidEPlan no que diz respeito à percepção da facilidade do seu uso e utilidade para identificação e classificação das ameaças à validade e suas respectivas ações de mitigação. O instrumento de coleta de dados empregado nesta etapa visou análise e investigação de tópicos relevantes da ValidEPlan, quanto aos aspectos gerais da ferramenta, termos, palavras e conceitos usados e facilidade de uso.

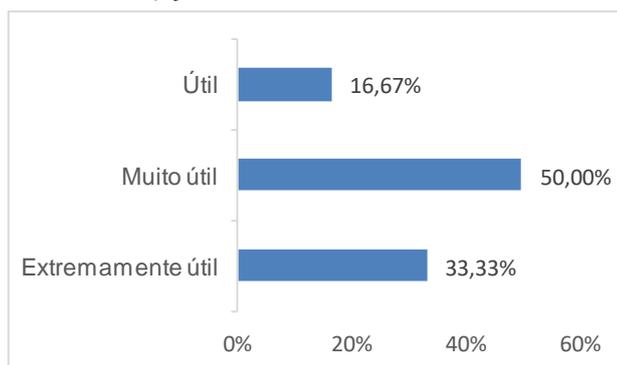
O questionário foi elaborado através da ferramenta Google Forms⁴, por permitir facilidade quanto à coleta e análise dos dados. Foram utilizados trechos dos comentários e sugestões que expressam a opinião dos especialistas em relação aos questionários para justificar resultados alcançados. Assim, estes serão referenciados no texto por E1, E2, E3, E4, E5 e E6.

Ao iniciar a avaliação, o participante (especialista) recebeu um roteiro de tarefa (**Apêndice B**) o qual continha as informações passo-a-passo para realizar um exemplo real de uso na ferramenta e em seguida proceder com o preenchimento do questionário com suas percepções positivas e negativas sobre a aceitação da ferramenta ValidEPlan. O cenário de aplicação do exemplo real de uso na ferramenta esteve centrado na identificação, classificação e priorização de ameaças à validade, cujo contexto foi criar um novo plano experimental, definir suas principais características, classificar as ameaças sugeridas e definir quais ações de mitigação são mais assertivas ao planejamento do experimento.

Ao questionarmos aos especialistas que considerando que a ferramenta "ValidEPlan" será empregado para auxiliar os pesquisadores/experimentadores na identificação, na classificação e na priorização de possíveis ameaças à validade e indicação de quais riscos devem ter prioridade ao serem tratados e quais ações de controle são mais assertivas para mitigá-los, 33,33% considerou a ferramenta extremamente útil, 50% admitiu a ferramenta como muito útil e 16,67% apenas como útil, conforme mostra a Figura 27, abaixo.

⁴ Formulários Google para criar fichas para pesquisas e coleta de dados *on-line*.

Figura 27 – Percepção de utilidade da ferramenta ValidEPlan



Fonte: Dados da pesquisa (2019).

Utilizando uma escala linear de 5 pontos (1 a 5) onde o participante pôde classificar sua reação entre o marcador inicial **Frustrante** (1) e marcador final **Satisfatório** (5), 66,7% dos especialistas optou pela posição 4 da escala e 33,3% pela posição 5, o que revela uma mediana de 4 em que todos os participantes consideraram a sua reação em relação ao uso da ferramenta como extremamente satisfatória.

Ainda no contexto dos aspectos gerais da ferramenta, ao questionarmos sobre qual foi a reação dos especialistas em relação à percepção da facilidade da ferramenta, onde o marcador inicial da escala linear correspondia a **Difícil** (1) e o marcador final a **Fácil** (5), 33,3% optaram pela posição 4 da escala e 66,7% pela posição 5, revelando uma mediana de 5, o que indica um excelente grau de percepção de facilidade geral em relação à ferramenta ValidEPlan.

Ao questionarmos os especialistas sobre qual foi sua reação em relação à adequação da ferramenta, levando em consideração que o marcador inicial da escala correspondia a **Inadequado** (1) e o marcador final como **Adequado** (5), 50% decidiram pela posição 4 e 50% pela posição 5, totalizando uma mediana de 4,5, o que aponta que os participantes consideraram sua reação geral em relação à ferramenta como extremamente adequada ao seu objetivo de gerenciar ameaças à validade de experimentos da área de ES.

Quando indagados sobre quais eram seus comentários sobre os tópicos gerais da ValidEPlan, os especialistas apontaram aspectos interessantes de funcionalidade e usabilidade da ferramenta:

Parou de funcionar algumas vezes (E1).

Eu comecei a classificar as ações de controle, e no meio da etapa eu utilizei o filtro, não sei o que aconteceu com as ações que eu defini antes (E2).

A classificação das ameaças e consequentemente ações podem ser um pouco cansativas dependendo da quantidade de ameaças sugeridas (E3).

Ferramenta cumpre com a proposta apresentada (E4).

De maneira geral, achei a ferramenta simples e fácil de usar. Acredito que ela possa ajudar o pesquisador a identificar e mitigar as ameaças a validade dos seus experimentos (E5).

A ferramenta está bem organizada e fácil de ser utilizada. Como existem algumas ameaças/ações relacionadas (mesmo em grupos diferentes), talvez a ferramenta pudesse fazer esse link de maneira mais intuitiva (E6).

Quanto aos textos da ValidEPlan, especificamente termos, palavras e conceitos usados, levando em consideração que o participante pôde classificar sua reação entre o marcador inicial **Inconsistente** (1) e marcador final **Consistente** (5), 16,7% dos especialistas optaram pela posição 3 da escala, 16,7% pela posição 4 e 66,7% pela posição 5, o que revela uma mediana de 5 pontos, ou seja, um excelente nível de consistência dos textos utilizados pela ferramenta.

Quando solicitados a descrever os problemas que foram encontrados relacionados aos termos, palavras e conceitos usados, os especialistas indicaram os seguintes aspectos:

Na parte de "Experiment Characteristics" tem muitas perguntas negativas, então para responder positivamente é necessário negar a negação. Isso torna tudo muito confuso (E3).

Na parte de "Threats" talvez seja legal usar a terminologia de análise de risco: Impacto, probabilidade, etc. Além disso, algumas ameaças poderiam ser reescritas para ficar com mais cara de ameaça. Por exemplo, "Participation rather than performance reward" não tem cara de ameaça, já "Inappropriate statistical test application" é bem sugestivo. Acho que a definição dos termos impacto, urgency e trend deveriam ficar fixos no topo da página (mesmo já tendo sido ditos em etapas anteriores), pois quando começo a preencher as vezes me perco na definição devido ao grande número de ameaças a classificar (E4).

Na tela de "Control Actions" acho que deveria ter o botão next e save não só embaixo da tela, mas também em cima da control action. Isso inclusive é um padrão de usabilidade, só não me recordo o nome. Fazendo assim você mantém os botões num ponto fixo, e o usuário não precisa ficar deslocando o mouse sempre para baixo (E5).

Questionados os participantes se a execução das tarefas na ferramenta ValidEPlan leva a resultados previsíveis, levando em consideração que o marcador inicial correspondia a **Nunca** (1) e o marcador final representava **Sempre** (5), 16,7% optaram pela posição 3 da escala, 66,7% pela colocação de número 4 e 16,7% pela

posição 5, resultando numa mediana de 4 pontos, considerada muito boa.

Tabela 12 – Resumo quantitativo das respostas sobre aprendizado do sistema

Critério	Mediana [1-5]
Reações sobre o aprendizado do sistema	4,25
Início do uso (Difícil – Fácil)	4
Tempo para aprender a usar o sistema (Demorado – Rápido)	4,5

Fonte: Elaboração própria (2019).

De acordo com a Tabela 12 (acima), observa-se que em relação aos aspectos (reações) gerais sobre a terminologia tais como: início do uso da ValidEPlan (Difícil – Fácil) e tempo para aprender a utilizar a referida ferramenta (Demorado – Rápido), obtiveram-se reações medianas de 4,25, consideradas ótimas. A seguir são apresentados os relatos em relação à facilidade de uso da ferramenta pelos especialistas:

A ferramenta é bem fácil de usar, segue o padrão de click forward típico desse tipo de sistema (E1).

Primeiro, na classificação das ameaças, já que se pode salvar sem ter terminado então ter algo tipo um next ou coisa do tipo além de salvar. Isso ajuda em situações que algum checkbox ficar pendente, usuário esqueceu de marcar e etc., porque do contrário a ferramenta não vai pra próxima etapa e o usuário pode ficar perdido sem saber o que está faltando. Segundo, meio lento a animação, marcação das checkboxes ao se clicar e; Terceiro, o filtro de magnitude não funciona muito bem, só aparecem de 2 em 2 as ações de controle, por filtro. Essa foi minha percepção inicial, foi um pouco confuso no início até entender como de fato funcionava nessa etapa (E3).

Ferramenta bastante intuitiva e simples de utilizar (E4).

De modo geral, a usabilidade da ferramenta é boa, com algumas pequenas melhorias que listei no comentário anterior (E5).

Boa usabilidade (E6).

Quando indagados quais eram os comentários sobre observações, sugestões, críticas ou problemas identificados na ferramenta os especialistas apontaram as seguintes considerações:

A ferramenta parou de funcionar algumas vezes (E1).

Acho que talvez seria útil que ferramenta gerasse os relatórios em formato pdf e/ou algum formato editável tipo docx, odt, ou google docs (E2).

No geral ok a ferramenta. e complementando as sugestões e comentários já descritos em seções anteriores: 1. sugestão: talvez permitir que o usuário sugira ameaças ou ações de controle e 2. questão: porque escolher ação de

controle que eu classificaria como não importante? melhor simplesmente ignorá-la? (E3).

No checklist havia colocado que os participantes não poderiam deixar o experimento, no entanto na tela de classificação das ameaças apareceu a ameaça "Participantes podem deixar o experimento". Acredito que esta ameaça não deveria ter aparecido na tela de classificação. Quando fiz o sign up não recebi o e-mail com a senha. Tive que clicar em "Esqueci a senha" para que pudesse redefinir a senha e fazer o login. Na tela dos relatórios notei que a numeração das ameaças sugeridas é diferente da numeração que aparece no relatório, isso pode confundir ou dificultar a análise do experimentador. Gostei bastante do trabalho e da ferramenta, acredito que podem auxiliar bastante (E4).

Ferramenta interessante como proposta de identificar problemas, principalmente para pesquisadores menos experientes (E6).

A partir da análise sobre a aceitação (usabilidade e utilidade) da ValidEPlan na visão de participantes especialistas, esta versão da ferramenta foi considerada aceitável com alguns pontos de melhorias.

6.2.4 ETAPA 4: Entrevista Narrativa Episódica

Para atender à finalidade metodológica e também validação do conteúdo de respostas propostas pelos participantes em nível especialista, foi aplicada o roteiro de entrevista semiestruturada (**Apêndice C**) contendo 13 (treze) questões divididas em 3 (três) segmentos centrais, quais sejam: a) Referente ao processo PrioriTTVs, o papel de ameaças à validade e ferramenta "ValidEPlan"; b) Referente a atividades de identificação e classificação de ameaças e ações de controle e; c) Referente aos relatórios disponíveis na ferramenta ValidEPlan.

Considerando as respostas e concepções dos especialistas, aqui sendo os mesmos, para fins de sigilo e anonimato, identificados por meio de códigos (E1, E2, E3, E4, E5, E6), procedeu-se ao uso de duas abordagens comuns para exploração das categorias pertinentes aos objetivos do trabalho: a entrevista de forma Narrativa Episódica e, de maneira complementar, a técnica de Análise de Conteúdo (BARDIN, 2009).

Esta última considera-se como uma das metodologias mais empregadas para identificação de cenários e categorias correspondentes a partir da análise dos diversos pontos de vistas que, em alguns vieses e perspectivas, tornam-se coincidentes ou também divergentes, quando for o caso, com os problemas de pesquisa formulados a priori pelo pesquisador.

Considerando a abordagem teórico-conceitual mais difundida na literatura e nos meios didático-metodológicos, tem-se percebido que:

As entrevistas narrativas se caracterizam como ferramentas não estruturadas, visando profundidade, de aspectos específicos, a partir das quais emergem as histórias de vida, tanto do entrevistado como entrecruzadas no contexto situacional. Esse tipo de entrevista visa encorajar e estimular o sujeito entrevistado (informante) a contar algo sobre algum acontecimento importante de sua vida e contexto social, **além de exteriorizar, em outros casos igualmente comuns, a percepção explicativa de resultados observados de determinado fenômeno, em que o participante pode manifestar a apreciação, ponto de vista ou a exposição dos efeitos deste fenômeno** (BAUER, 2002, grifo nosso).

Por meio da ótica episódica, tangível às narrativas, é possível estruturar e diagnosticar contextos intrínsecos aos experimentos, evidenciando os desafios, problemáticas e proposição de possíveis soluções de saneamento ou reforma dos riscos identificados sob a premissa do conhecimento realizado pelo especialista. Sem o apego à ênfase em informações diversas ao ambiente do experimento controlado, a entrevista narrativa episódica possibilita que tanto o pesquisador quanto o especialista se detenham, evidentemente, nos fatos e fenômenos oriundos dos objetivos da pesquisa, sem, contudo, se distanciar do real propósito (BAUER; GASKELL, 2002; MUYLAERT *et al.*, 2014).

Por meio da narrativa episódica, como forma de entrevista semiestruturada, valida-se um esforço empreendido no ato da pesquisa, com enfoque estratégico na busca e descoberta de variáveis fundamentais à consecução dos objetivos práticos e os pressupostos teóricos da investigação.

Com identificação de diversas concepções, então, a Análise de Conteúdo de Bardin (2009) emprega-se como meio subsidiário para identificação das principais categorias de estudo, respeitando os pontos de vista do especialista e o alinhamento à proposta metodológica do presente trabalho.

Nesse sentido, observou-se unanimidade, sem confrontos narrativos, no reconhecimento por parte dos especialistas da importância dos fatores de riscos para a validade experimental, em especial quanto à identificação, à compreensão dos seus impactos nos projetos experimentais e consequente ponto de partida para uma maior verificação dos enfrentamentos (saneamento) dos problemas expressos em experimentos em ES, implicando em uma maior qualidade e confiabilidade das informações experimentais.

Essa concepção se tornou perceptível em trechos expressos por alguns especialistas, ao longo da realização da entrevista, reproduzindo-se o seguinte:

Acredito que tratar as ameaças à validade serve para se aumentar a qualidade de resultados obtidos e confiança no experimento realizado (E1).

Eu acho que é fundamental a questão de ameaças a validade pros resultados, porque, a depender das ameaças que a gente conseguiu identificar em um determinado experimento, pode ter a validade muito restrita a ponto dos resultados serem completamente irrelevantes até, a depender do nível de criticidade das ameaças, se elas não forem tratadas. Então, eu acredito que a relação é direta e é extremamente importante, fundamental, porque para ficar com um resultado do experimento de alta qualidade você precisa ter um tratamento ideal de ameaças a validade (E2).

Eu acho que ameaças a validade tem total relação com a qualidade do experimento, porque o nome já diz tudo, vai estar ameaçando a validade do experimento, então está relacionado com questões que: as pessoas que estão realizando o experimento podem aprender com o experimento, enfim, tem vários aspectos que podem afetar tanto o resultado do experimento, como se por um caso vier a ser replicado; enfim, tudo que está relacionado com ameaças à validade, ele pode afetar diretamente a qualidade do experimento (E3).

Dos excertos acima, reproduzidos na prática narrativa e compreendidos sob a perspectiva da Análise de Conteúdo (BARDIN, 2009), percebe-se, pois, a percepção das ameaças à validade nos experimentos em ES como reflexos diretos na qualidade, confiabilidade e segurança dos dados explorados a partir dos fenômenos em estudo.

Outro ponto de vantagem apontado pelos especialistas foi a identificação de possíveis gargalos (*gaps*) no processo, configurando em problemas ao experimento, ainda mais na observância do cenário hipoteticamente configurado do estudo e as reais implicações na prática.

Quando você faz um experimento e não descreve as ameaças à validade, com certeza você vai ter leque, um “gap”, digamos assim, entre o que poderia ser realmente colocado e o que você colocou, então, descrever as ameaças à validade seria, assim, uma questão muito importante na hora de você colocar isso, quando você esquece alguma ou não descreve uma corretamente, fica parecendo que o seu trabalho está fugindo das ameaças à validade que ele realmente tem. Quem realmente compreende o assunto, vai ler aquilo e vai observar que, realmente, não se descrevem as ameaças corretamente (E5).

O desenho de concepção experimental de um recurso computacional, em sua estrutura de formulação e desenvolvimento, deve integrar, assim, os riscos e ameaças à confiabilidade e segurança dos experimentos, trazendo elementos e variáveis de ameaça compatíveis com finalidade de cada experimento, além de abrir um espaço para a inserção de novos fatores de risco, caso existam no ambiente controlado e não

anteriormente percebido pelo pesquisador.

A compreensão da validade de um experimento também foi objeto de uma formulação de questão da entrevista, associando-o à atividade de controle de variáveis e planejamento do experimento:

Quando você fala que um experimento é válido, eu acho que é quando o experimentador fez lá todo o planejamento do experimento dele e levantou as possíveis ameaças que poderiam, de alguma forma, invalidar aquele experimento e ele teve esse cuidado, quando ele foi fazer planejamento e execução, enfim, todo esse experimento e aí ele teve esse cuidado e ao analisar todos os resultados, métricas que ele utilizou para analisar, enfim, todo o experimento. Então ele se preocupou com tudo isso e ao analisar todas essas questões é que você consegue ver: não teve nenhum processo ou fator que invalidou o experimento de alguma forma, então pode-se dizer que o mesmo é válido. Eu acho que está relacionado com tudo aquilo que é feito no experimento (E3).

A validade compreende, também, avaliação de resultados e dados ligados ao experimento controlado, a fim de se capacitar o experimentador para maior compreensão daquilo é necessário, pois, para a consecução de reais objetivos da pesquisa, distanciando a probabilidade do erro:

Uma parte importante no experimento é a parte de avaliação dos resultados, dos dados na verdade, então, se você não faz uma boa validação dos dados, uma avaliação correta dos dados, com certeza seu experimento não tem tanto impacto. Eu percebo que tem outros detalhes importantes do processo, mas o que eu vejo de mais comum assim de equívocos e do que realmente traz um impacto direto no resultado é a avaliação dos dados. O pessoal faz de qualquer jeito (E6).

Na qualidade de subsídio para a proposição de uma ferramenta que eleve o potencial da validade de experimentos controlados, o ValidEPlan constitui-se como uma ferramenta (mecanismo) que colabora para a efetividade da ação de identificação das ameaças e de riscos à validade experimental, bem como propicia um planejamento das fases mais articulado e sistematizado.

Acredito que a ferramenta tem boa contribuição na identificação das ameaças e de tratamentos mais adequados para riscos identificados. A ferramenta me ajudou identificar várias ameaças, por exemplo, em relação a se definir o treinamento das turmas por pessoas diferentes (INT-A29) (E1).

A ferramenta propõe, no caso ela faz baseado no trabalho de vocês, primeiro vocês fizeram levantamento de algumas ameaças a validade colocaram lá, na ferramenta, então pode ser que eu quando fosse fazer um determinado experimento, que eu não identificasse alguma daquelas ameaças que se levantou no planejamento do experimento e com a ferramenta ela vai auxiliar isso, então se eu coloquei lá todos os fatores, marquei no checklist as características do experimento, então a ferramenta já vai me trazer sugestões, digamos assim, de ameaças que podem acontecer no experimento, então acho que pode auxiliar bastante na questão da validade

do experimento e qualidade também consequentemente (E3).

Naturalmente, em se tratando de um projeto que deve atender a um fim específico no plano experimental controlado, é possível se observar lacunas e pontos de melhoria, a fim de tornar a ferramenta do ValidEPlan compatível com a complexidade dos ambientes experimentais e da multiplicidade de ameaças e fatores de risco, dentre os quais uma grande parte se forma da concepção, do desenvolvimento e relacionamentos interpessoais existentes.

Em princípio sim, mas talvez alguns, eu não diria nem alguns ajustes, mas ela deveria ser mais bem integrada com outras ferramentas, com outras fases do processo, não somente com a fase da especificação, porque acho que está focando muito exatamente em uma das fases do experimento, seria mais interessante que você conseguisse de forma mais ampla possível especificar todo o experimento lá dentro da ferramenta (E5).

Com certeza, porque assim, é muito comum que acabe esquecendo algum tipo de ameaça a validade, porque quando você vai escrever, você não vai ler ameaça por ameaça, para tentar identificar o que pode ou não ter acontecido no plano experimental ou durante o seu experimento, então, com certeza, ela vai ajudar a ter uma garantia de que você está colocando todas as ameaças ou pelo menos boa parte delas, no esforço maior, pro seu experimento, acho que nesse ponto ela ajuda bastante. Como existem várias ameaças, às vezes você nem conhece todas. Um detalhe que acho que coloque, em alguma parte da avaliação, do processo, eu coloquei que questão importante, é que ameaças à validade podem surgir, e aí, como se trata essas novas ameaças, como é que você consegue colocar na ferramenta, isso aí não consegui perceber, esse aí é talvez o ponto a se pensar, ou seja, como se trata nova ameaça? Surgiu uma nova ameaça na literatura, como a gente vai colocar isso na ferramenta (E6).

Ou seja, a partir da premissa de possíveis configurações e manutenções neste instrumento, apesar da sua significância para os estudos controlados em ES, do ponto de vista dos ambientes controlados, ainda assim torna-se válido e necessário gerenciar os recursos de identificação das ameaças com maior rigor e critérios, abrindo espaços para outras abordagens não consideradas durante a execução do programa, além de promover integração e coesão (mais intuitiva e conectada) com demais sistemas empregados para a mesma finalidade.

Penso que o checklist apresenta questionamentos amplos, atingindo boa completude em relação às ameaças à validade. Talvez algumas poucas questões possam ser excluídas (ou incorporadas a outras), até mesmo para deixar o checklist mais enxuto (E1).

Eu acho que quando vocês fazem o levantamento, não tem como, obviamente, levantar todas as ameaças que existem, mas é só com relação a isso. Eu acho que, no decorrer, depois que for concluído podem ser adicionadas outras ameaças que vocês ou mesmo outros identificarem. Acho que vai incrementar e melhorar o trabalho (E3).

A principal crítica é que a ferramenta ainda não está abrangente o suficiente, parece que está muito voltada para parte de ameaça, seria mais interessante que conseguisse especificar o experimento como um todo e que, baseado na especificação geral do experimento que as ameaças, elas fossem deduzidas indiretamente talvez, porque me pareceu que estava muito voltada, assim, a forma como a gente especificou as características do experimento já conduzia um pouco a responder algumas ameaças, até as próprias perguntas já induzia um pouco qual ameaça, se você já conhecesse um pouco, já induzia qual a ameaça ou quais ameaças aquela pergunta se tratava, ou seja, seria mais interessante especificar se ele indiretamente, assim, numa especificação genérica e fosse deduzida essas informações (E5).

Questionados se ao longo do emprego da ferramenta, quando ocorreu, de fato, a identificação de ameaças à validade do plano experimental controlado, observou alguns contrapontos que evidenciaram o atendimento à proposta do instrumento de identificação e de concepção das ações corretivas dos fatores de riscos perceptíveis.

Logo depois de preencher o checklist, ele já apareceu para classificar as ameaças, então, para mim, os fatores de risco evidenciados, ele já estava identificando o que poderia acontecer com o experimento, logo acho que é na segunda parte da ferramenta (E3).

Em linhas gerais, em relação a este questionamento (indagação científica) anterior, o que ficou claramente definido, por meio da ótica de especialistas, foi um estranhamento generalizado, definido em um momento de respostas quase vagas e razoavelmente superficiais, que não explicavam ou corroboram com a elucidação dos fenômenos da ameaça à validade nos experimentos em ES.

O *checklist*, por aduzir diversos pontos de risco e ameaças à validade, foi um dos pontos que também suscitou problemáticas na identificação real desse efeito nos experimentos controlados em ES, porém abrindo uma possibilidade e um espaço para adequações e melhorias à cada realidade da pesquisa.

Em relação ao relato da experiência do especialista na ferramenta com a etapa de classificação das ameaças e ações de controle, sob a perspectiva da narrativa episódica:

[...] a sensação que tive é que, eu acho que realmente é um momento que a pessoa precisa parar pra refletir de uma forma bem crítica cada um dos itens e a sensação que tive é que a ferramenta realmente faz com que você pare, aquela lista na realidade, não necessariamente a ferramenta, mais aquela lista que você dá, faz com que você pare e tenha que refletir sobre aquilo e eu acho que isso é bastante positivo. Em relação ao que está fazendo não é exatamente um experimento, mas é como se fosse um experimento. É uma avaliação. Eu senti um pouco cansado, porque o volume de coisas era muito grande, não sei se é porque eu já estava no fim do dia, já tinha feito bastante coisa e aquilo ali exigia que parasse e ficasse pensando, mas a dificuldade também tem muito a ver com o fato de eu estar respondendo pra um experimento que não tinha muito conhecimento, por mais que você tivesse passado o documento. Logo, acredito que isso talvez tenha influenciado um

pouco no cansaço, aí a sensação que eu tenho é que não foi tanto o checklist ou a ferramenta, se eu tivesse avaliando um experimento, uma coisa que eu tivesse fazendo agora, para mim, acho que talvez não tivesse sido tão cansativo. Só pra dividir essas duas coisas, porque tem a minha percepção sobre o negócio e em relação ao estar executando esse experimento (E2).

Utilizar a ferramenta na classificação foi uma das partes complicadas no sentido geral; foi onde a ferramenta foi instável, especialmente porque ela era muito verbosa, tipo assim, ela sugeria muitas ameaças e quando você sugere muita coisa já se perde um pouco assim e, muitas vezes, em algumas, até sugeria, eu não lembro agora, mas tinha algumas até que contradiziam as ameaças, tinham uma que dizia uma coisa e a outra dizia uma outra ameaça, sei lá, ameaça que era pequena demais a outra era grande demais, esse tipo de coisa na mesma situação. E como eram muitas ameaças para você classificar, isso deixaria também o trabalho bem cansativo, você ter que fazer cada uma, algumas ameaças são muito parecidas, sempre achava: essa aqui acho que já fiz lá em cima etc. Foram especialmente esses os momentos sobre a classificação, que eu mais destacaria (E5).

A ValidEPlan, embora passível de falhas e problemas, fatos absolutamente comuns a ferramentas experimentais, produziu entusiasmo e cumpriu seu fim de suscitar a reflexão mais consciente, determinada e racional acerca dos fatores de riscos e ameaças diversas à validade, confiabilidade e segurança de dados e informações oriundas de experimentos controlados em ES.

Eu acho que o processo levou a identificar as ameaças e as ações possíveis. Penso que este processo me levou a pensar em situações que provavelmente eu não teria pensado, principalmente da maneira exaustiva como ela faz e sistemática, porque você tem lá o checklist. Então realmente obriga você a pensar em todos os pontos (E2).

[...] a ferramenta auxilia bastante nessa identificação de ameaças a validade, por você já ter feito um trabalho de identificá-las, as ações de controle, enfim, todo o processo e trazer tudo isso para ferramenta, então acho que isso auxilia bastante quando a pessoa está fazendo o planejamento do experimento (E3).

[...] Sem dúvidas. Eram tantas as ameaças que você até ficava meio perdido, aí quando você a lista com suas formas gradativas de impacto, de magnitude, você tem uma noção. Porque se fosse pra você listar todas as ameaças, eu vou ter que tratar todas elas da mesma forma, você vai ler e isso pode acontecer, mais qual seria o impacto? Então fazer essa avaliação, em algum momento, depois você ter essa ideia da magnitude, realmente faz muita diferença (E6).

Por fim, a ferramenta ValidEPlan possibilita ao experimentador um novo horizonte a ser observado, percebendo os diversos riscos inseridos nos espaços experimentais, muitos dos quais o pesquisador não consegue, de uma certa forma, vislumbrar durante a fase de execução e também de desenvolvimento, seja pelo seu envolvimento contínuo, efeitos naturais, contato entre participantes ou mesmo pelo esgotamento físico e mental provenientes das tarefas e atividades experimentais, que

exigem dos experimentadores maior cuidado, atenção e exercício decisório.

6.3 AVALIAÇÃO EXPERIMENTAL POR ESTUDANTES EM ESE

Com o objetivo de analisar na prática a ferramenta ValidEPLan, apresentamos aqui um novo ciclo de avaliação a partir do ponto de vista de experimentadores iniciantes. Considerando que a proposta tem o objetivo de auxiliar pesquisadores experientes e principalmente iniciantes a planejar experimentos da área de ES, com foco na identificação e priorização das ameaças à validade e suas ações de mitigação, o grupo de alunos participantes constitui o público alvo elementar deste trabalho.

Foram analisados no viés desta avaliação experimental, experimento controlado, questionário e entrevistas, compreendidos a partir do ponto de vista de 20 discentes de mestrado e doutorado, matriculados na disciplina Engenharia de Software Experimental (IN1080), do CIn-UFPE, período 2018.2.

O planejamento de condução desse estágio da avaliação foi em função das atividades de cronograma instituído, desenvolvido entre setembro e dezembro de 2018, sendo compartimentado em três etapas principais (Tabela 13, abaixo).

Tabela 13 – Etapas do Planejamento da Avaliação com Estudantes

AVALIAÇÃO: ALUNOS DE ENGENHARIA DE SOFTWARE EXPERIMENTAL		
1ª Etapa	2ª Etapa	3ª Etapa
Questionário Demográfico	Experimento Controlado: - Completude da Tarefa Questionário de Avaliação da Ferramenta - Facilidade de uso - Percepção de satisfação - Percepção de aprendizagem	Entrevista Narrativa Episódica

Fonte: Elaboração própria (2019).

Em relação à **1ª etapa**, estipulou-se obter as informações sobre o perfil dos estudantes através de um formulário demográfico aplicado antes da realização das outras etapas de avaliação. O formulário foi usado também para informar aos participantes os objetivos da investigação, uso dos dados coletados e principalmente para descrição das características da amostra, verificação de experiências prévias e

ajustes quanto ao treinamento. Os dados coletados através desse instrumento também guiaram uma melhor atribuição dos participantes aos diferentes grupos e tratamentos durante o experimento.

No que compete à **2ª etapa** do estudo em voga, objetivou-se realizar um experimento controlado para analisar o uso na prática da ValidEPlan, avaliando o seu impacto no processo de planejamento e gestão de ameaças à validade de experimentos conduzidos pelos alunos, com respeito à corretude da tarefa de planejamento e gerência de ameaças à validade. Verificou-se, também, através de um questionário de avaliação coletar informações de cunho quantitativo ligados às opiniões sobre uso da ValidEPlan (objeto de avaliação da etapa), com o objetivo de verificar a facilidade de uso e percepção de satisfação e aprendizagem da ferramenta

A **3ª etapa** dessa fase executou o roteiro de entrevista narrativa episódica, este com o fim de coletar os depoimentos dos alunos (participantes), de maneira a analisar indícios de satisfação dos participantes, bem como, as percepções positivas e negativas sobre o uso da ferramenta, bem como apontar sugestões e descrever sobre suas experiências individuais.

6.3.1 ETAPA 1: Análise dos Dados Demográficos

O questionário usado para a coleta dos dados de informações demográficas foi conduzido com 20 alunos de pós-graduação (mestrado e doutorado), matriculados na disciplina Engenharia de Software Experimental. Dessa maneira, com base na identificação das questões, observaram-se os resultados percentuais abaixo relacionados, seguidos de análise teórico-conceitual, quando necessário.

Primeiramente, em relação ao sexo dos participantes, notou-se que, em maioria eram de homens (60%) em detrimento de participantes do sexo feminino (40%). Esse é um dado esperado pois cada vez mais tem-se observado, nas universidades, uma participação e presença em cursos da área de computação maior por parte de homens (MOREIRA; MATOS; REIS, 2014). No entanto, apesar da observação ser relevante, inclusive para suscitar pesquisas futuras para o cenário de investigação, não foi o cerne desta pesquisa identificar esses perfis no âmbito dos objetivos delimitados dessa avaliação.

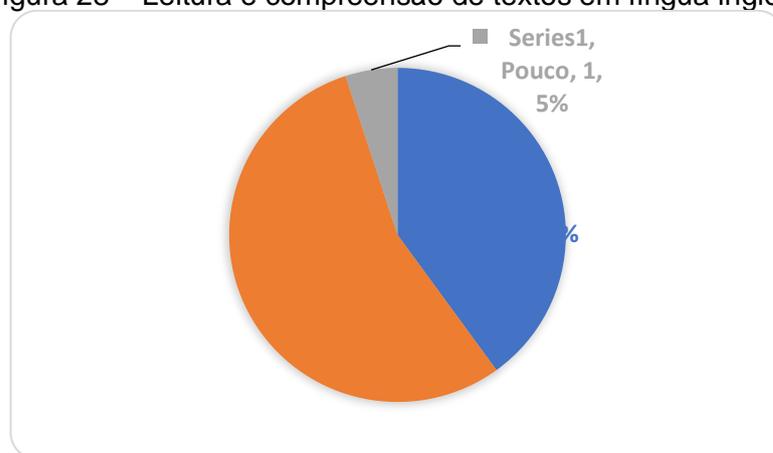
Com relação às idades dos participantes, observou-se um intervalo entre 22 e 46 anos, uma média de idade de 29 anos, evidenciando uma amostra com perfil

jovem, em plena fase de contribuições para o mercado ativo de trabalho e na proposição de soluções que viabilizem os problemas da sociedade atual.

Conforme relatado anteriormente, a ferramenta ValidEPlan foi escrita na língua inglesa para facilitar o acesso a pesquisadores de diversos países. Quando o idioma dos participantes é diferente do idioma do material experimental, pode existir um risco para o experimento realizado. Dessa forma, perguntamos aos participantes o nível de leitura e compreensão de textos em língua inglesa. O gráfico da Figura 28, abaixo, revelou que 55% leem e compreendem o idioma razoavelmente, 40% bem e apenas 5% pouco.

Evidentemente o aprendizado e compreensão da língua inglesa proporcionam um contato com novas culturas e novos conhecimentos e até mesmo um melhor posicionamento no mercado de trabalho. No entanto, para o escopo dessa etapa da avaliação, ter 95% dos participantes com aceitáveis habilidades de leitura e compreensão da língua inglesa foi suficiente pois a ferramenta ValidEPlan possui uma interface amigável, intuitiva, simples e fácil de utilizar, que possa ajudar a atender ao usuário de forma satisfatória.

Figura 28 – Leitura e compreensão de textos em língua inglesa

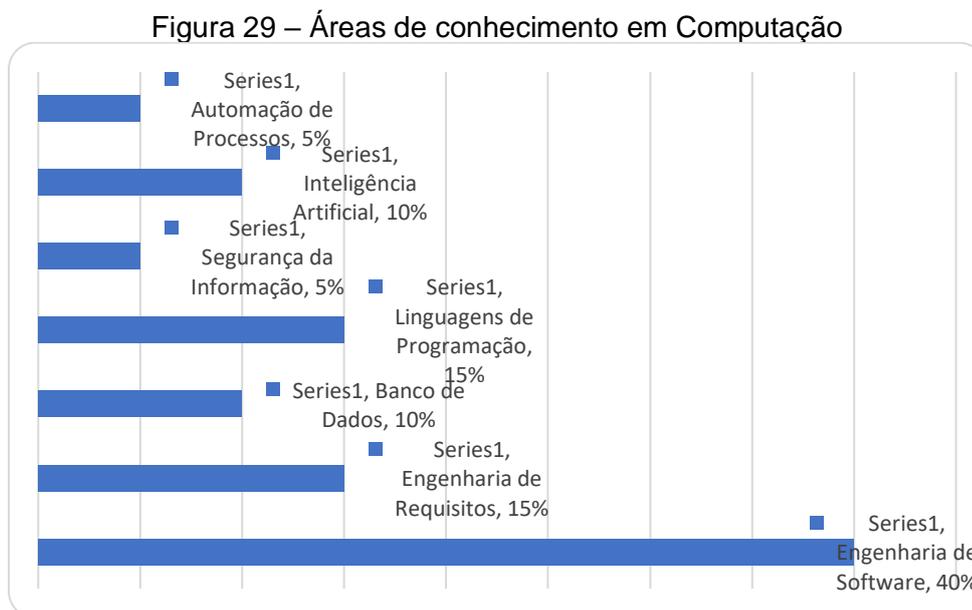


Fonte: Dados da pesquisa (2019).

Com respeito às áreas de interesse e conhecimento na computação por parte dos participantes verificamos que, sua grande maioria (40%) se interessam e possuem maior conhecimento em Engenharia de Software, conforme o gráfico da Figura 29, abaixo. Observamos, também, interesse e conhecimento em outras áreas como Engenharia de Requisitos (15%), Linguagens de Programação (15%), Banco de Dados (10%), Inteligência Artificial (10%), entre outras.

A importância da verificação de conhecimentos prévios se deu em virtude de

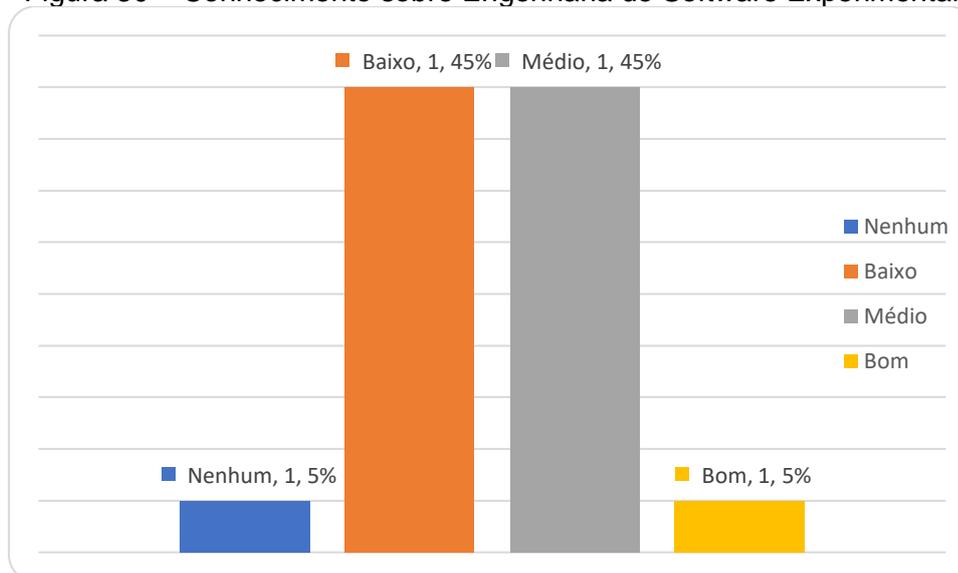
possíveis ajustes quanto ao treinamento prévio ao experimento controlado realizado.



Fonte: Dados da pesquisa (2019).

Conhecimentos prévios são as experiências e informações práticas e teóricas que temos guardadas em nossa mente e que podemos utilizar quando precisamos. Dessa forma, questionamos os participantes sobre nível de conhecimento ligado à Engenharia de Software Experimental. Obteve-se 45% com conhecimento médio, 45% possuem conhecimento baixo, 5% bom e 5% não possui nenhum conhecimento, conforme o gráfico da Figura 30.

Figura 30 – Conhecimento sobre Engenharia de Software Experimental

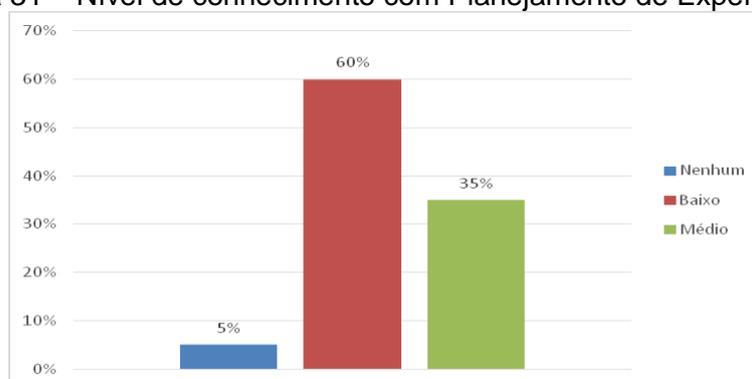


Fonte: Dados da pesquisa (2019).

Por sua vez, o gráfico da Figura 31, abaixo, mostra o nível de conhecimento relacionado ao planejamento de experimentos. Notou-se, que 60% dos estudantes participantes possuem um baixo conhecimento sobre os planos experimentais, 35% conhecimento médio e 5% baixo conhecimento.

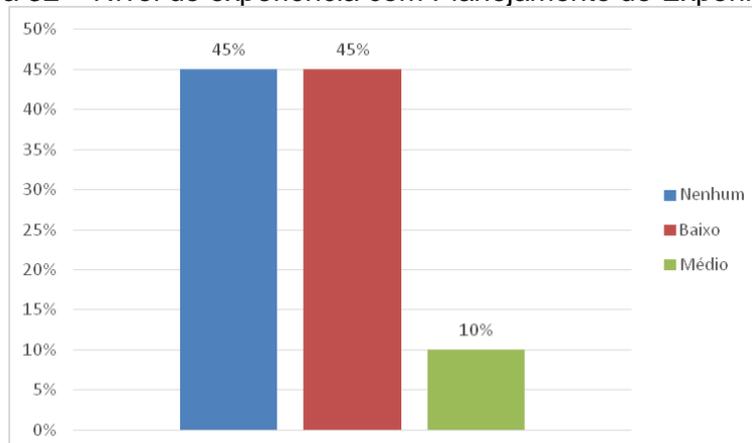
Conforme o gráfico da Figura 32, 45% dos estudantes não tinham nenhuma experiência com planos experimentais, assim como, 45% experiência baixa e 10% média. Esses dados são relevantes pois mostram que o público alvo dessa fase da avaliação, de fato era iniciante a planejar experimentos da área de ES.

Figura 31 – Nível de conhecimento com Planejamento de Experimentos



Fonte: Dados da pesquisa (2019).

Figura 32 – Nível de experiência com Planejamento de Experimentos



Fonte: Dados da pesquisa (2019).

Outras informações relacionadas ao nível de conhecimento dos participantes sobre Engenharia de *Software*, Linguagens de Programação, Desenvolvimento e Inspeção de *Software* foram coletadas, porém, serão relatadas quando discutirmos sobre os materiais e tarefas experimentais do experimento controlado, Etapa 2 desta avaliação.

De forma geral, todos os dados coletados que explicitam os níveis e áreas de

conhecimento por parte dos discentes, permitem-nos saber, por exemplo, se o conhecimento dos participantes é similar ao de pessoas do contexto real do estudo ou mesmo perceber se possuem conhecimento sobre as hipóteses experimentais, uma vez que essas informações de perfil dos participantes ajudam-nos a identificar e antecipar possíveis problemas à validade do experimento controlado realizado. Os dados do questionário também guiaram uma melhor atribuição dos participantes aos diferentes grupos e tratamentos.

6.3.2 ETAPA 2: Experimento Controlado

Para avaliar o impacto da proposta através da ferramenta ValidEPlan foi proposto este experimento que será apresentado ao longo das próximas seções. O estudo teve o intuito de analisar o uso na prática da ferramenta, avaliando o seu impacto no processo de planejamento e gestão de ameaças à validade de experimento conduzido por alunos matriculados em uma disciplina de ESE, com respeito à corretude das tarefas executadas e quantidade de ameaças e ações de controle identificadas.

Além disso, ao final do experimento, aplicamos um questionário com o objetivo de avaliar as percepções de facilidade, satisfação e aprendizagem dos participantes ao usar a ferramenta ValidEPlan.

6.3.2.1 Definição

Avaliar o uso da ferramenta web ValidEPlan com a execução de um experimento controlado. Confrontamos o plano experimental de um mesmo experimento realizado de maneiras diferentes: uma com o auxílio da ferramenta ValidEPlan e a outra utilizando a abordagem tradicional, de forma *ad hoc*. O objetivo do estudo foi estruturado segundo a abordagem GQM (*Goal, Question and Metric*) (VAN SOLINGEN *et al.*, 2002), descrita a seguir:

Analisar o uso da ferramenta ValidEPlan. **Com o propósito** de avaliar seu impacto relacionado à identificação e controle de ameaças à validade de um experimento da área de ES. **Com respeito** à forma que a ferramenta ajuda a examinar a corretude da tarefa de identificar ameaças, ações, relacionamentos e a ordem de prioridade, assim como, quantidade identificada de ameaças e ações de mitigação,

em comparação com um planejamento sem a ferramenta. **Sob o ponto de vista** de estudantes de pós-graduação de disciplina de Engenharia de Software Experimental. **No contexto** de planos experimentais para experimentos controlados em engenharia de software.

6.3.2.2 Questões de Pesquisa

A partir dos objetivos e definições anteriormente determinados para este experimento, definiram-se as seguintes questões de pesquisa:

- **QP1.** Uso da ferramenta ValidEPlan afeta a corretude da tarefa de identificar e analisar ameaças e ações de mitigação de um experimento controlado da área de ES, comparado ao uso das práticas *ad hoc*?
- **QP2.** O uso da ferramenta ValidEPlan permite ao experimentador identificar um maior número de ameaças e ações de mitigação ao planejar um experimento controlado da área de ES, comparado ao uso das práticas *ad hoc*?
- **QP3:** Qual é a percepção do participante relacionada à aprendizagem, satisfação e facilidade de uso quanto ao planejamento de experimento controlado da área de ES com a utilização da ferramenta ValidEPlan?

6.3.2.3 Preocupações Éticas

É importante relatar que esta pesquisa teve o cuidado com as questões éticas, garantindo os direitos dos participantes, sempre se pautando na Resolução 466/12 do Conselho Nacional de Saúde (CNS).

As informações gerais e específicas estão no Formulário de Consentimento (**Apêndice D**), que trata sobre permissão e uso dos dados capturados, formalização à participação no estudo, objetivos do estudo, pesquisadores, procedimentos, coleta de dados, confidencialidade de registros, riscos e desconfortos, custos e declaração de consentimento.

6.3.2.4 Hipóteses, Variáveis e Medidas

Para o referido experimento, foram definidas as seguintes hipóteses nulas e alternativas:

- H_{01} – não existe diferença estatística na corretude da tarefa (C_F) de identificar e gerenciar ameaças à validade de um experimento controlado da área de ES com ou sem (C_A) o uso da ferramenta ValidEPlan;

- H_{A1} – existe diferença estatística na corretude da tarefa (C_F) de identificar e gerenciar ameaças à validade de um experimento controlado da área de ES com o uso da ferramenta ValidEPlan;

$$H_{01}: C_F = C_A$$

$$H_{A1}: C_F \neq C_A$$

- H_{02} – não existe diferença estatística na quantidade (Q_F) de ameaças e ações de mitigação ao planejar um experimento controlado da área de ES, com ou sem (Q_A) o uso da ferramenta ValidEPlan;

- H_{A2} – existe diferença estatística na quantidade (Q_F) de ameaças e ações de mitigação identificadas ao planejar um experimento controlado da área de ES, com o uso da ferramenta ValidEPlan;

$$H_{01}: Q_F = Q_A$$

$$H_{A1}: Q_F \neq Q_A$$

O experimento foi conduzido de forma a tentar rejeitar as hipóteses nulas (H_{01} , H_{02}), demonstrando que existe diferença estatisticamente significativa ao usar a ferramenta ValidEPlan para examinar a corretude da tarefa e a quantidade de ameaças e ações de controle ao planejar um experimento controlado da área de ES, diferente da obtida sem o uso da ferramenta.

Visando responder às duas primeiras questões de pesquisa relativas à corretude da tarefa (QP1) e quantidade de ameaças e ações de controle (QP2), foram definidas duas variáveis dependentes e suas respectivas métricas:

- **Corretude da Tarefa:** Representada pelo percentual de itens identificados corretamente em relação ao número total de respostas.

- **Quantidade de ameaças e ações de controle:** Representada pelo número de ameaças e ações de mitigação encontrados pelos participantes.

Consideramos os “**Itens Identificados Corretamente**” em comparação com aqueles especificados por um documento de planejamento experimental, chamado Modelo de Referência, elaborado por quatro pesquisadores especialistas da área de ESE. Além dos itens de planejamento, o plano terá uma lista de ameaças à validade, ações de controle e seus respectivos relacionamentos. O documento construído será usado como parâmetro/gabarito de comparação da variável corretude da tarefa.

Para responder a terceira questão de pesquisa (QP3) foi elaborado um questionário para ser respondido pelos participantes do experimento, visando coletar as suas percepções de aprendizagem, satisfação e facilidade quanto ao uso da ferramenta ValidEPlan.

6.3.2.5 Materiais Experimentais e Tarefas

A finalidade desta seção é descrever os objetos, orientações e instrumentos de medida que foram utilizados para realização do experimento e o seu monitoramento (WOHLIN *et al.*, 2012).

A realização do experimento transcorreu em um dos laboratórios da pós-graduação do CIn-UFPE. Cada aluno participante utilizou um computador com acesso à internet e ao material necessário à realização do experimento. Entre os materiais disponíveis cada participante utilizou: a) Roteiro contendo os aspectos experimentais a serem seguidos; b) Slides utilizados durante o treinamento e; c) Descrição das unidades experimentais a serem utilizadas e d) Referências bibliográficas indicadas pela disciplina.

As unidades experimentais que receberam a aplicação dos tratamentos são definidas a seguir:

- **Unidade Experimental – UE1:** Realizar o planejamento de um experimento controlado para verificar se a escolha de uma determinada ferramenta lúdica para o ensino de programação (variável independente) afeta ou não o esforço e a efetividade da aprendizagem de alunos de uma disciplina de lógica de programação (variável dependente). O experimento a ser planejado é avaliar: (i) Se o tempo de execução de atividades práticas de lógica de programação, utilizando a ferramenta Alice é menor que utilizando a ferramenta Scratch e; (ii) Se a quantidade de acertos em atividades práticas de lógica de programação, utilizando a ferramenta Alice é menor que utilizando a ferramenta Scratch.

- **Unidade Experimental – UE2:** Realizar o planejamento de um experimento controlado para verificar se a escolha de determinado método de inspeção de software (variável independente) afeta ou não a efetividade da inspeção (variável dependente). O experimento a ser planejado é avaliar se a quantidade de defeitos encontrados em um documento de requisitos de software, utilizando o processo tradicional de inspeção de software *Ad-hoc* é menor que utilizando o processo de inspeção por *Checklist*.

Baseado na descrição das unidades experimentais UE1 e UE2, um grupo formado por quatro pesquisadores independentes, com experiência em engenharia de software experimental, criaram planos experimentais de referência (Modelos de Referência), seguindo o processo de experimentação baseado nos livros de Wohlin *et al.* (2012) e/ou Juristo e Moreno (2013).

Cada plano tinha também uma lista de ameaças à validade, ações de controle e seus respectivos relacionamentos. As categorias e critérios abordados nos Modelos de Referência, referentes à gestão das ameaças à validade, estão descritos na Tabela 14, os documentos completos estão disponíveis através do **Anexo A** e **Anexo B**.

Tabela 14 – Temáticas abordadas no Modelo de Referência

Itens	Critérios abordados
1	Ameaças à Validade Identificação e discussão das ameaças, vieses e possíveis limitações que possam influenciar os resultados do estudo.
2	Ações de mitigação Identificação e discussão das ações de controle aplicadas a cada uma das ameaças identificadas.
3	Relacionamentos entre ações e ameaças e ordem de prioridades Identificação de relacionamentos entre ações de mitigação e ameaças à validade que possam gerar novas ameaças, assim como, ordem de prioridade entre ameaças identificadas.

Fonte: Elaboração própria (2019).

As ameaças à validade e suas respectivas ações de controle, assim como os relacionamentos identificados pelos participantes em seus planos experimentais e que não constavam no Modelo de Referência, foram analisadas por pesquisadores independentes no intuito de decidir se iriam ou não as incluir ao referido modelo.

Para medir a variável dependente os participantes realizaram as seguintes tarefas: (i) construir planos experimentais baseados na descrição das unidades experimentais e seguindo o tratamento proposto; (ii) identificar as ameaças à validade e ações para controlá-las e; (iii) verificar se há relacionamentos de causa e consequências entre as ameaças e ações identificadas.

Ao final da execução do experimento, os participantes forneceram: (i) planos experimentais construídos com base na descrição das unidades experimentais e seguindo o tratamento proposto; (ii) lista com as ameaças e ações de controle identificadas e; (iii) lista com as ameaças e ações cujos relacionamentos entre ambos podem causar novas ameaças, indicando, por nível de prioridade, quais são as

ameaças que proporcionam maiores riscos à confiabilidade do estudo e as melhores ações de controle a serem utilizadas, de acordo com o contexto e objetivo do experimento.

Por fim, foi solicitado o preenchimento de um questionário de opinião onde os alunos participantes foram solicitados a responder questões fechadas que visaram caracterizar aspectos relativos à percepção do usuário quanto à facilidade, satisfação e aprendizagem por meio do uso da ferramenta ValidEPlan.

6.3.2.6 Seleção dos participantes

O experimento foi realizado por 20 alunos matriculados na disciplina de ESE do programa de pós-graduação do CIn da UFPE (2018.1). Com isso, a amostra utilizada foi não probabilística, ou seja, foi por conveniência. A participação no experimento contou como nota parcial da disciplina e foi realizada presencialmente nos laboratórios de pós-graduação do CIn, no período de agosto a outubro de 2018.

O público da disciplina, em geral, possui conhecimento em engenharia de software e engenharia de software experimental, porém com pouca experiência na realização de experimentos. No decorrer da disciplina o conhecimento dos alunos é desenvolvido e eles aprendem os princípios da área de ESE e os aplicam através do planejamento e execução de experimentos de seus próprios projetos de pesquisa.

É fundamental afirmar que os participantes tiveram a liberdade de solicitar que seus dados não fossem coletados ou utilizados para análise, sem qualquer prejuízo individual na disciplina.

6.3.2.7 Design Experimental

Com o objetivo de reduzir o efeito da variabilidade existente entre o nível de experiência dos participantes e a complexidade das unidades experimentais (UE1 e UE2) nas variáveis de saída, o experimento foi organizado segundo o modelo experimental Quadrado Latino (RYAN, 2009; JURISTO; MORENO, 2013).

Cada quadrado latino foi ordenado como uma matriz 2x2, com um fator e dois tratamentos. Linhas e colunas sofrerão variações e serão variáveis de bloqueio do estudo. Com isso, assegura-se que todos os participantes serão expostos a ambos os tratamentos e a aplicação de algum tratamento a cada unidade experimental.

O fator de tratamento do experimento será a tarefa de planejar experimentos da área de ES dirigido/orientado a validade de seus resultados. Para isso, os seguintes tratamentos serão utilizados: (i) Utilização da ferramenta ValidEPlan e (ii) Utilização de práticas *ad hoc* (livros, *checklists* e resumos).

Tabela 15 – Organização das Réplicas no Quadrado Latino

Participantes	UE1	UE2
P1 e P11 P3 e P13 P5 e P15 P7 e P17 P9 e P19	ValidEPlan (A)	<i>Ad hoc</i> (B)
P2 e P12 P4 e P14 P6 e P16 P8 e P18 P10 e P20	<i>Ad hoc</i> (B)	ValidEPlan (A)

Fonte: Elaboração própria (2019).

Conforme relatado, cada quadrado latino foi organizado como uma matriz dois por dois (duas linhas e duas colunas), como ilustra a Tabela 15 acima, contendo:

- Duas unidades experimentais (UE1 e UE2);
- Vinte participantes organizados em torno de 10 (dez) réplicas;
- Cada participante devendo analisar cada um dos tratamentos (*Ad-hoc* e ValidEPlan) de forma alternada.

6.3.2.8 Procedimentos

Antes da execução do experimento, os participantes foram submetidos à dinâmica da disciplina Engenharia de Software Experimental (IN1080), onde estudaram e discutiram os seguintes materiais: (i) Capítulos 3, 4, 5, 6, 7 e 9 do Livro “*Basics of software engineering experimentation*” (JURISTO; MORENO, 2013) e; (ii) Capítulo 8 do livro “*Experimentation in software engineering*” (WOHLIN *et al.*, 2012).

Os participantes também foram subordinados a um treinamento da ferramenta ValidEPlan. A dinâmica da disciplina assim como o treinamento, foram utilizados para nivelamento sobre o conhecimento de Engenharia de Software Experimental, planejamento de experimentos controlados da área de ES e sobre o uso da ferramenta

proposta.

Para execução do treinamento, utilizando o método tradicional (*ad hoc*) e a ferramenta ValidEPlan foi utilizado a seguinte unidade experimental: Realizar o planejamento de um experimento controlado para verificar se a escolha de determinada linguagem de programação (variável independente) afeta ou não o tempo de programação de um determinado algoritmo (variável dependente). O experimento planejado pelos participantes teve o objetivo de avaliar se o tempo de programação do algoritmo *MergeSort*, utilizando a linguagem de programação Java é menor que utilizando a linguagem de programação “C”.

Durante o treinamento, foram esclarecidas possíveis dúvidas dos participantes, a fim de realizar o experimento principal de forma adequada. O treinamento teve a duração de aproximadamente 2 (duas) horas e serviu para alinhar o entendimento dos conceitos adquiridos durante o estudo dos materiais teóricos da disciplina, além de permitir a realização de possíveis ajustes nos materiais experimentais que foram adotados no estudo.

A realização do experimento ocorreu em um dos laboratórios do Centro de Informática (CIn) da UFPE. Cada participante utilizou um computador, com acesso à internet e ao sítio onde está hospedado a ferramenta ValidEPlan. Como orientação para a realização do experimento cada participante utilizou: (i) uma descrição do experimento (roteiro) descrevendo aspectos experimentais a serem compreendidos e o cenário da ferramenta a ser explorado; (ii) os slides utilizados nos treinamentos; e (iii) as especificações das unidades experimentais a serem planejadas.

Cada participante teve um período de 30 minutos para ler o roteiro e a respectiva especificação da unidade experimental antes de iniciar o planejamento do experimento e posteriormente responder a questionários. Os dois grupos receberam materiais experimentais ao mesmo tempo. A ordem de execução e atribuição do tratamento a ser utilizado por cada participante foi definido por meio de aleatorização.

A execução do experimento foi acompanhada e guiada por dois pesquisadores. O cronograma com datas e atividades para a realização do experimento encontra-se na Seção 6.3.2.9.

6.3.2.9 Preparação da Execução do Experimento

A realização do experimento deu-se durante o período de desenvolvimento da disciplina de ESE, ou seja, no ínterim entre agosto e outubro de 2018. Dividimos essa dinâmica experimental em alguns marcos:

- Leitura e discussão dos capítulos 3, 4, 5, 6, 7 e 9 do Livro “*Basics of software engineering experimentation*” (JURISTO; MORENO, 2013) e capítulo 8 do livro “*Experimentation in software engineering*” (WOHLIN *et al.*, 2012), no período entre 13/08 e 17/09;
- Assinatura do Termo de Consentimento, apresentação dos objetivos do experimento e cronograma, realização dos sorteios das unidades experimentais, tratamentos e posição dos participantes em cada réplica do quadrado latino, em 20/09;
- Realização do treinamento dos participantes com a ferramenta ValidEPlan e preenchimento do questionário demográfico, em 24/09;
- Início da primeira etapa do experimento com o envio do material experimental por *e-mail* e apresentação da descrição da primeira Unidade Experimental (UE1) a ser utilizada, em 27/09;
 - Execução da primeira etapa (UE1), em 01/10;
 - Entrega da primeira etapa do experimento, em 04/10;
 - Início da segunda etapa do experimento com envio do material experimental por *e-mail* e apresentação da descrição da segunda Unidade Experimental (UE2) a ser utilizada, em 04/10;
 - Execução da segunda etapa (UE2), em 08/10;
 - Entrega da segunda etapa do experimento, em 11/10;
 - Aplicação do questionário de avaliação, da realização e do agendamento de entrevistas, em 15/10.

O *design* estatístico utilizado foi o quadrado latino (matriz 2x2), com 20 alunos participantes dispostos em 10 réplicas do quadrado, duas unidades experimentais (UE1 e UE2) e cada participante analisou cada um dos tratamentos (*Ad hoc* e ValidEPlan) de forma alternada.

Com a utilização da função randômica da ferramenta de planilha eletrônica Microsoft Excel, foram realizados sorteios para a aleatorização das posições dos participantes nas réplicas do quadrado latino, aplicação dos tratamentos e ordem de utilização das unidades experimentais. Para aleatorização das réplicas de quadrado

latino foram efetuados os seguintes sorteios:

- Posição dos participantes nas 10 (dez) réplicas;
- Ordem de colocação da unidade experimental a ser inserida na primeira coluna do quadrado latino;
- Ordem de atribuição do tratamento a ser utilizado pelo primeiro participante de cada réplica e assim por diante.

O sorteio dos participantes que ocuparam as 20 posições possíveis nas 10 (dez) réplicas de quadrado latino obedeceu às seguintes etapas:

1) Listamos o nome dos participantes em ordem alfabética em uma coluna da planilha eletrônica denominada “Participantes”;

2) Criamos uma coluna chamada “Sorteio”, contendo a expressão randômica – “=ALEATÓRIO()”. A função gerou um valor real aleatório entre zero e um;

3) Ordenamos os participantes pela coluna “Sorteio”, classificados pelos valores das células, pela ordem do menor valor para o maior valor;

4) Compomos cada quadrado por dois participantes conforme a ordem obtida, com os primeiros dois participantes, formando a primeira réplica e assim por diante.

O sorteio das posições das unidades experimentais igualmente respeitou um procedimento análogo, onde:

- Elencamos em uma planilha eletrônica os nomes das duas unidades experimentais (UE1 e UE2);

- Criamos uma coluna na planilha chamada de “sorteio”, contendo a expressão randômica “=ALEATÓRIO()” (esta função gerou um valor real aleatório entre zero e um);

- Ordenamos as unidades experimentais (UE1 e UE2) pela coluna “sorteio”;

- Usamos as unidades experimentais na ordem produzida pela função randômica, na primeira réplica do quadrado latino e na ordem oposta na segunda réplica, e assim por diante.

Aleatorização equivalente foi realizada para os tratamentos utilizados no experimento (ValidEPlan e *Ad hoc*). É importante destacar que garantimos que as alternativas possíveis de configurações do quadrado latino tivessem o mesmo número de sorteios. A Tabela 16 apresenta o resultado das aleatorizações.

Tabela 16 – Aleatorização dos Quadrados Latinos

	Participantes	Unidade Exp. 1	Unidade Exp. 2
Réplica 1	1	ValidEPlan	<i>Adhoc</i>
	2	<i>Adhoc</i>	ValidEPlan
Réplica 2	3	<i>Adhoc</i>	ValidEPlan
	4	ValidEPlan	<i>Adhoc</i>
Réplica 3	5	<i>Adhoc</i>	ValidEPlan
	6	ValidEPlan	<i>Adhoc</i>
Réplica 4	7	ValidEPlan	<i>Adhoc</i>
	8	<i>Adhoc</i>	ValidEPlan
Réplica 5	9	ValidEPlan	<i>Adhoc</i>
	10	<i>Adhoc</i>	ValidEPlan
Réplica 6	11	ValidEPlan	<i>Adhoc</i>
	12	<i>Adhoc</i>	ValidEPlan
Réplica 7	13	ValidEPlan	<i>Adhoc</i>
	14	<i>Adhoc</i>	ValidEPlan
Réplica 8	15	ValidEPlan	<i>Adhoc</i>
	16	<i>Adhoc</i>	ValidEPlan
Réplica 9	17	ValidEPlan	<i>Adhoc</i>
	18	<i>Adhoc</i>	ValidEPlan
Réplica 10	19	<i>Adhoc</i>	ValidEPlan
	20	ValidEPlan	<i>Adhoc</i>

Fonte: Elaboração própria (2019).

6.3.2.10 Resultados Alcançados

Por questão de impessoalidade, bem como, facilitação da análise dos dados, os participantes foram identificados nas tabelas sequencialmente como P1, P2, P3... P20. A Tabela 17 registra os resultados da correção de cada plano experimental entregue pelos participantes. A primeira coluna lista as réplicas do quadrado latino (variando de 1 a 10), a segunda lista os participantes (variando de 1 a 20), a terceira as unidades experimentais, a quarta lista os tratamentos utilizados. As demais listam o resultado para cada uma das temáticas avaliadas.

O processo de avaliação e confrontação dos planos experimentais desenvolvidos pelos participantes com o modelo de referência foi desempenhado pelo

pesquisador principal e revisado por um segundo pesquisador. A análise de possíveis conflitos foi tratada por dois pesquisadores e caso persistisse o desacordo, foi considerada a opinião de um terceiro pesquisador. Todo este processo foi realizado sob a supervisão do orientador desta tese de doutorado.

Os temas descritos e abordados pelo Modelo de Referência (ver Tabela 14), em comparação com os planos experimentais entregues ao final do experimento pelos participantes, foram avaliados como:

- **TOTALMENTE CORRETO (0)** → Significa que o participante discorreu sobre a temática e os critérios abordados de forma correta, ou seja, respondeu o esperado em relação ao Modelo de Referência definido por especialistas;
- **PARCIALMENTE CORRETO (1)** → Significa que o participante respondeu a temática de forma incompleta ao exposto no Modelo de Referência determinado pelos especialistas;
- **INCORRETO (2)** → Significa que o participante abordou a temática e os critérios de forma incorreta, isto é, o item tratado pelo aluno no plano experimental não estava descrito no Modelo de Referência estabelecido pelos especialistas.

Tabela 17 – Respostas obtidas a partir dos participantes

Réplica	Part.	UE	Tratamento	Pontos Abordados		
				Ameaças	Ações	Relacionamentos e Ordem de Prioridade
1	P1	1	ValidEPlan	0	0	1
	P2	1	<i>Adhoc</i>	0	0	0
2	P3	1	<i>Adhoc</i>	1	1	1
	P4	1	ValidEPlan	0	0	0
3	P5	1	<i>Adhoc</i>	1	2	2
	P6	1	ValidEPlan	0	0	0
4	P7	1	ValidEPlan	0	0	0
	P8	1	<i>Adhoc</i>	1	0	0
5	P9	1	ValidEPlan	0	0	0
	P10	1	<i>Adhoc</i>	0	0	0
6	P11	1	ValidEPlan	0	0	0
	P12	1	<i>Adhoc</i>	1	1	2
7	P13	1	ValidEPlan	0	0	0
	P14	1	<i>Adhoc</i>	1	1	1

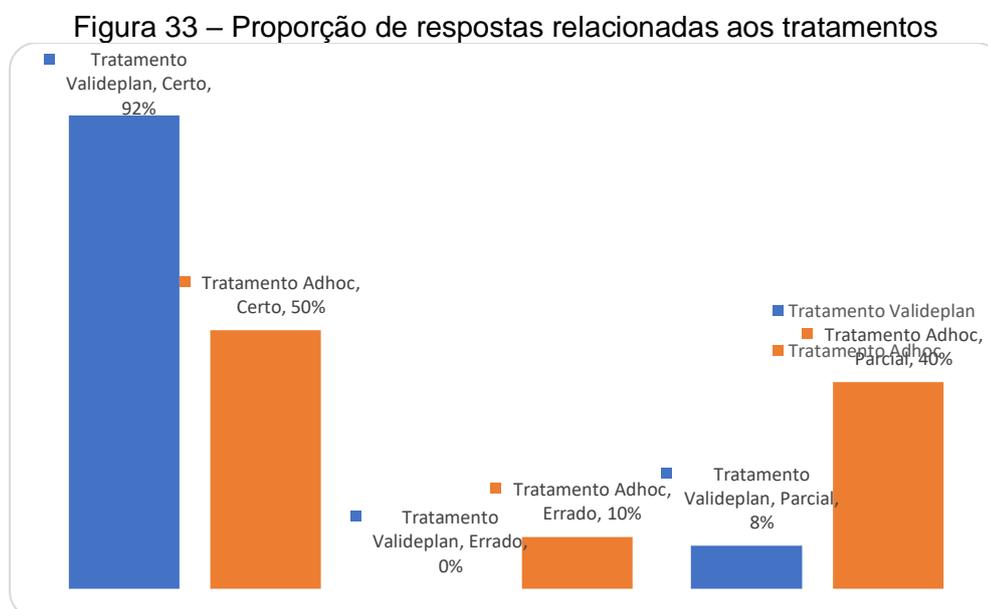
				Pontos Abordados		
Réplica	Part.	UE	Tratamento	Ameaças	Ações	Relacionamentos e Ordem de Prioridade
8	P15	1	ValidEPlan	0	0	0
	P16	1	<i>Adhoc</i>	1	1	0
9	P17	1	ValidEPlan	0	0	1
	P18	1	<i>Adhoc</i>	1	0	0
10	P19	1	<i>Adhoc</i>	1	1	2
	P20	1	ValidEPlan	1	1	1
1	P1	2	<i>Adhoc</i>	0	0	0
	P2	2	ValidEPlan	0	0	0
2	P3	2	ValidEPlan	0	0	0
	P4	2	<i>Adhoc</i>	0	0	0
3	P5	2	ValidEPlan	0	0	0
	P6	2	<i>Adhoc</i>	0	1	2
4	P7	2	<i>Adhoc</i>	0	1	0
	P8	2	ValidEPlan	0	0	0
5	P9	2	<i>Adhoc</i>	0	0	1
	P10	2	ValidEPlan	0	0	0
6	P11	2	<i>Adhoc</i>	1	1	1
	P12	2	ValidEPlan	0	0	0
7	P13	2	<i>Adhoc</i>	0	0	1
	P14	2	ValidEPlan	0	0	0
8	P15	2	<i>Adhoc</i>	0	0	0
	P16	2	ValidEPlan	0	0	0
9	P17	2	<i>Adhoc</i>	0	0	0
	P18	2	ValidEPlan	0	0	0
10	P19	2	ValidEPlan	0	0	0
	P20	2	<i>Adhoc</i>	1	1	2

Fonte: Elaboração própria (2019).

A Figura 33 exibe o gráfico de proporções de respostas totalmente corretas, parcialmente corretas e incorretas para o conjunto das temáticas abordadas pelo modelo de referência.

O tratamento ValidEPlan evidenciou 92% do total de respostas totalmente corretas contra, especificamente, 50% do tratamento *Ad hoc*. Em relação ao número

total de respostas totalmente incorretas, ValidEPlan teve 0% contra 10% de *Ad hoc*. Já as respostas às questões consideradas parcialmente corretas, o número total em ValidEPlan apontou 8% do total, e em *Ad hoc* esse número representou 40% do total.



Fonte: Dados da pesquisa (2019).

Quanto aos resultados individuais de cada uma das temáticas abordadas pelo modelo de referência, observamos através do gráfico da Figura 34 os seguintes resultados:

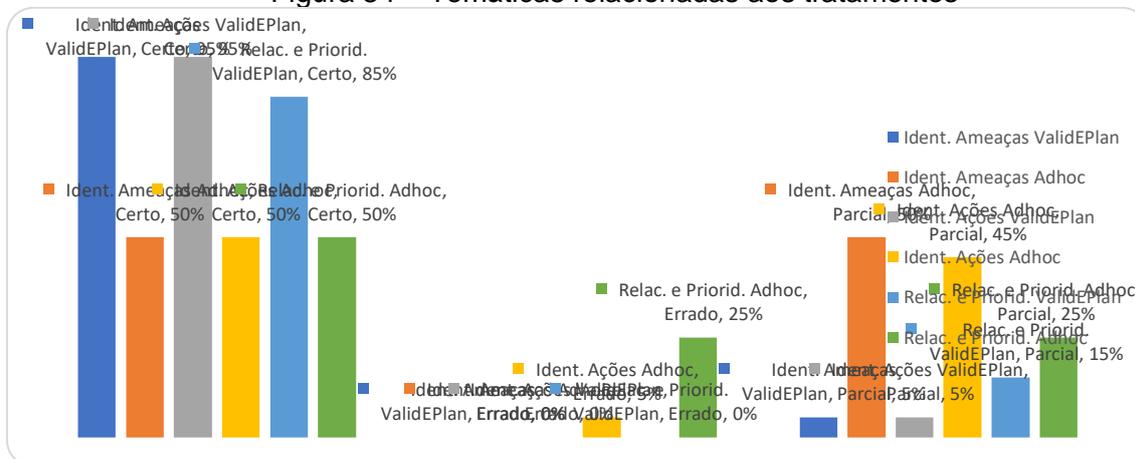
- **Relacionado à Identificação de ameaças à validade:** 95% das respostas totalmente corretas foram atribuídas ao tratamento com ValidEPlan enquanto 50% ao tratamento *Ad hoc*. Relacionado às respostas incorretas ambos os tratamentos obtiveram 0% das respostas. Quanto às respostas parcialmente corretas, 5% das respostas foram obtidas através de ValidEPlan e 50% *Ad hoc*.

- **Relacionado à Identificação de Ações de Controle:** Relativo às respostas totalmente corretas, 95% se referiam ao tratamento ValidEPlan enquanto 50% *Ad hoc*. Correspondente ao número de retornos definidos como incorretos 0% à ValidEPlan e 5% *Ad hoc*. Pertinente às temáticas definidas com respostas parcialmente corretas 5% foram atribuídas à ValidEPlan e 45% *Ad hoc*.

- **Relacionado à Identificação de Relacionamentos e definição de ordem de prioridades:** 85% das respostas totalmente corretas são referem ao tratamento ValidEPlan e 50% ao tratamento *Ad hoc*. Sobre as respostas incorretas, 0% foram indicadas à ValidEPlan e 25% *Ad hoc*. Quanto às respostas denominada como

parcialmente corretas, 15% foram do tratamento ValidEPlan e 25% referente ao tratamento *Ad hoc*.

Figura 34 – Temáticas relacionadas aos tratamentos



Fonte: Dados da pesquisa (2019).

A análise estatística foi realizada através do uso do pacote estatístico IBM SPSS (versão 2.0, IBM Corporation). O SPSS foi desenvolvido para ser empregado em inúmeras áreas da ciência e permite realizar uma variedade de análises (por exemplo: análises descritivas, análises inferenciais, multivariadas, gráficos, entre outros). Apesar da amostra utilizada ser de apenas 20 participantes, utilizamos uma significância estatística de $\alpha = 0,05$, considerada adequada para obtermos um melhor poder estatístico, não comprometendo a conclusão dos resultados (DYBÅ; KAMPENES; SJØBERG, 2006).

Para compreender quais os testes estatísticos seriam necessários para realizar o cálculo das diferenças de média, inicialmente verificamos o pressuposto de normalidade da amostra. Para isso, procedeu-se os testes de normalidade K-S *Lilliefors* e *Shapiro-Wilk* para os dados mostrados na Tabela 17, e os resultados exibidos através da Tabela 18. Considerando um nível de significância de 5%, percebe-se que os dois testes rejeitam a hipótese de normalidade, visto que o valor de Sig. (p), para todas as variáveis analisadas, é inferior a 0,05, revelando que a amostra apreciada não possui uma distribuição normal.

Tabela 18 – Testes de Normalidade

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estatística	Df	Sig. (p)	Estatística	df	Sig. (p)
Identif. de Ameaças – ValidEPlan	,538	20	,000	,236	20	,000
Identif. de Ameaças – Adhoc	,335	20	,000	,641	20	,000
Identif. de Ações – ValidEPlan	,538	20	,000	,236	20	,000
Identif. de Ações – Adhoc	,318	20	,000	,737	20	,000
Relac. e Prioridades – ValidEPlan	,509	20	,000	,433	20	,000
Relac. e Prioridades – Adhoc	,311	20	,000	,753	20	,000

a. *Lilliefors Significance Correction*

Fonte: Dados da pesquisa (2019).

Como os dados não seguem uma distribuição normal, a análise das médias foi realizada através de teste de hipótese estatística não paramétrica. Neste caso, foi utilizado o teste *Wilcoxon (Signed-Rank Test)*, um teste de hipótese estatística não paramétrica, usado para comparar amostras dependentes, também chamadas de emparelhadas ou relacionadas, obtidas em ocasiões distintas para avaliar se as classificações médias da população diferem (WOOLSON, 2007).

O teste comparou dois tratamentos ValidEPlan (T1) e *Adhoc* (T2) considerando a identificação de ameaças à validade, ações de controle, relacionamentos entre ações e ameaças e a ordem de prioridades.

Tabela 19 – Respostas relacionadas aos tratamentos

Respostas	Teste de Hipótese	Interpretação
TOTALMENTE CORRETO	Mean T1 = 2,7500 Mean T2 = 1,5000 Based in positive ranks Z = -3,022 Wilcoxon Signed Ranks Test, Asymp. Sig. (2-tailed) P-Value = 0,003	Podemos constatar que o teste de <i>Wilcoxon (signed-rank Test)</i> mostrou que as proporções de respostas totalmente corretas são estatisticamente diferentes, sendo MAIOR utilizando o tratamento ValidEPlan (T1), com valor de p menor que 0,05.
PARCIALMENTE CORRETO	Mean T1 = 0,2500 Mean T2 = 1,2000 Based in positive ranks Z	Podemos constatar que o teste de <i>Wilcoxon (signed-rank Test)</i> mostrou que as proporções das respostas

	= -2,795 Wilcoxon Signed Ranks Test, Asymp. Sig. (2- tailed) P-Value = 0,005	consideradas parcialmente corretas são estatisticamente diferentes, sendo MENOR com o uso do tratamento ValidEPlan (T1), com valor de p menor que 0,05.
INCORRETO	Mean T1 = 0,0000 Mean T2 = 0,3000 Based in positive ranks Z = -2,121 Wilcoxon Signed Ranks Test, Asymp. Sig. (2- tailed) P-Value = 0,034	Podemos constatar que o teste de <i>Wilcoxon (signed-rank Test)</i> mostrou que as proporções das respostas incorretas são estatisticamente diferentes, sendo MENOR com o uso do tratamento ValidEPlan (T1), com valor de p menor que 0,05.

Fonte: Dados da pesquisa (2019).

Os resultados obtidos, apresentados e interpretados através da Tabela 19 mostram que, para todos os casos de respostas, rejeitou-se a hipótese de nulidade e aceitou-se a alternativa. Ou seja, respondendo à primeira questão de pesquisa (**QP1**) deste experimento, para esta amostra, existe diferença estatística na corretude da tarefa de identificar e gerenciar ameaças à validade de um experimento controlado da área de ES com o uso da ferramenta ValidEPlan, comparado ao uso de práticas de planejamento *ad hoc*.

Destaca-se que, o número de respostas totalmente corretas aumentou e a taxa de erros diminuiu ao identificar ameaças, ações de mitigação, relacionamentos e ordem de prioridades, quando usando o tratamento ValidEPlan. No entanto, apesar desse benefício, percebemos que os participantes que usaram o tratamento *ad hoc* em seus planos experimentais identificaram uma quantidade maior de ameaças e ações específicas ao seu propósito de experimentação.

Apesar de possivelmente ter havido uma autoconfiança dos participantes que usaram o tratamento ValidEPlan e dessa forma não se esforçaram em identificar ameaças específicas ao seu experimento, interpretamos esse fato sendo absolutamente normal, visto que ameaças à validade são específicas e estão ligadas ao propósito de cada experimentação, o que mostra a necessidade de evolução da ferramenta no sentido de capturar mais informações de contextos experimentais específicos, além da realização de novos estudos experimentais para confirmar ou refutar tais fenômenos.

Ainda assim, durante a avaliação, a ferramenta ofereceu cobertura com maior completude dos principais riscos que poderiam afetar a validade dos resultados do experimento em questão, além de permitir a inserção manual de novas ameaças e ações de controle que não constavam entre as sugeridas pela ferramenta.

Tabela 20 – Respostas obtidas a partir dos participantes

Réplica	Part.	Unid. Exp.	Tratamento	Qtd. Ameaças	Qtd. Ações
Réplica 1	P1	UE1	ValidEPlan	44	86
Réplica 2	P4	UE1	ValidEPlan	37	82
Réplica 3	P6	UE1	ValidEPlan	46	65
Réplica 4	P7	UE1	ValidEPlan	8	5
Réplica 5	P9	UE1	ValidEPlan	44	85
Réplica 6	P11	UE1	ValidEPlan	42	40
Réplica 7	P13	UE1	ValidEPlan	54	50
Réplica 8	P15	UE1	ValidEPlan	36	48
Réplica 9	P17	UE1	ValidEPlan	48	141
Réplica 10	P20	UE1	ValidEPlan	46	48
Réplica 1	P2	UE2	ValidEPlan	52	103
Réplica 2	P3	UE2	ValidEPlan	46	50
Réplica 3	P5	UE2	ValidEPlan	60	63
Réplica 4	P8	UE2	ValidEPlan	41	94
Réplica 5	P10	UE2	ValidEPlan	29	35
Réplica 6	P12	UE2	ValidEPlan	34	50
Réplica 7	P14	UE2	ValidEPlan	41	25
Réplica 8	P16	UE2	ValidEPlan	29	30
Réplica 9	P18	UE2	ValidEPlan	45	96
Réplica 10	P19	UE2	ValidEPlan	41	60
Réplica 1	P1	UE2	<i>Adhoc</i>	10	21
Réplica 2	P4	UE2	<i>Adhoc</i>	10	16
Réplica 3	P6	UE2	<i>Adhoc</i>	6	8
Réplica 4	P7	UE2	<i>Adhoc</i>	8	9
Réplica 5	P9	UE2	<i>Adhoc</i>	14	29
Réplica 6	P11	UE2	<i>Adhoc</i>	3	4

Réplica	Part.	Unid. Exp.	Tratamento	Qtd. Ameaças	Qtd. Ações
Réplica 7	P13	UE2	<i>Adhoc</i>	7	8
Réplica 8	P15	UE2	<i>Adhoc</i>	8	9
Réplica 9	P17	UE2	<i>Adhoc</i>	13	21
Réplica 10	P20	UE2	<i>Adhoc</i>	4	3
Réplica 1	P2	UE1	<i>Adhoc</i>	9	7
Réplica 2	P3	UE1	<i>Adhoc</i>	10	14
Réplica 3	P5	UE1	<i>Adhoc</i>	5	6
Réplica 4	P8	UE1	<i>Adhoc</i>	9	9
Réplica 5	P10	UE1	<i>Adhoc</i>	25	25
Réplica 6	P12	UE1	<i>Adhoc</i>	4	4
Réplica 7	P14	UE1	<i>Adhoc</i>	6	9
Réplica 8	P16	UE1	<i>Adhoc</i>	3	5
Réplica 9	P18	UE1	<i>Adhoc</i>	20	37
Réplica 10	P19	UE1	<i>Adhoc</i>	10	13

Fonte: Dados da pesquisa (2019).

A Tabela 20, acima, segue o mesmo padrão da Tabela 17, no entanto, registra a quantidade de ameaças e as ações de mitigação que os participantes identificaram em seus planos experimentais, em detrimento do uso dos tratamentos ValidEPlan e *Adhoc*.

Tabela 21 – Testes de Normalidade

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estatística	df	Sig. (p)	Estatística	df	Sig. (p)
Quantidade de Ameaças ValidEPlan	,195	20	,046	,901	20	,043
Quantidade de Ameaças Ad hoc	,243	20	,003	,858	20	,007
Quantidade de Ações ValidEPlan	,156	20	,200	,967	20	,685
Quantidade de Ações Ad hoc	,261	20	,001	,860	20	,008
a. Lilliefors Significance Correction						

Fonte: Dados da pesquisa (2019).

Novamente, para entender quais testes estatísticos seriam necessários para realizar o cálculo das diferenças de média, verificamos se a distribuição dos dados

era normal ou diferente de uma distribuição normal. Para isso, procederam-se os testes de normalidade K-S *Lilliefors* e *Shapiro-Wilk* (RAZALI *et al.*, 2011) para os dados mostrados na Tabela 20, e os resultados exibidos através da Tabela 21.

Considerando um nível de significância de 5%, percebe-se que para as variáveis “Quantidade de Ameaças” usando o tratamento ValidEPlan e *Ad hoc*, bem como, a variável “Quantidade de Ações” utilizando o tratamento *Ad hoc*, os dois testes rejeitam a hipótese de normalidade, visto que o valor de Sig. (p), para estas variáveis analisadas, é inferior a 0,05. Apenas relacionado à variável “Quantidade de Ações”, utilizando o tratamento ValidEPlan, o valor de p é maior que 0,05 e, portanto, para esta variável a distribuição dos dados é normal.

Contudo, entende-se que, de forma geral, a amostra apreciada não possui uma distribuição normal e com isso, descarta-se a possibilidade de usar testes paramétricos. Os testes de normalidade apresentam baixo poder estatístico em amostras com $N < 30$ (valor da amostra menor que 30), então, para ter uma maior garantia na análise dos dados, optamos por usar testes não paramétricos (RAZALI *et al.*, 2011).

A análise das respostas referentes ao quantitativo de ameaças e ações de mitigação foi realizada por meio de teste de hipótese estatístico *Wilcoxon (Signed-Rank Test)*, que compara dados relacionados de amostra com distribuição diferente de normal (WOOLSON, 2003).

Tabela 22 – Teste de hipótese estatística Wilcoxon

Quantidade	Tratamento	Média	Z (Ranks)	Valor Sig. (p)
Ameaças à Validade	ValidEPlan	41,1500	-3,824	0,000
	<i>Ad hoc</i>	9,2000		
Ações de Controle	ValidEPlan	62,8000	-3,883	0,000
	<i>Ad hoc</i>	12,8500		

Fonte: Dados da pesquisa (2019).

Os resultados apresentados através da Tabela 22 acima, constatam que o teste de *Wilcoxon (signed-rank Test)* mostrou que a diferença entre quantidade de ameaças e ações de controle usando o tratamento ValidEPlan e *Ad hoc* é estatisticamente diferente, sendo MAIOR utilizando o tratamento ValidEPlan, com valores de $p = 0,000$ para ambas variáveis, ou seja, menor que 0,05.

Dessa forma, respondendo à segunda questão de pesquisa (QP2) deste

experimento, para esta amostra, o uso da ferramenta ValidEPlan permitiu ao experimentador identificar um maior número de ameaças e ações de mitigação ao planejar um experimento controlado da área de ES, comparado ao uso de práticas de planejamento *ad hoc*.

De forma geral, a ferramenta ValidEPlan cumpre o seu objetivo, no sentido de proporcionar, principalmente aos experimentadores menos experientes, um roteiro passo-a-passo com orientações que os ajudarão a não esquecer, negligenciar ou omitir aspectos relevantes ao desenvolvimento do seu plano experimental e ao mesmo tempo, através de um processo estruturado, prover meios para identificar, analisar e priorizar de forma semiautomática os principais riscos à validade de seu experimento.

6.3.2.11 Ameaças à Validade

A literatura define alguns procedimentos e taxonomia indicada para tratamento de ameaças à validade. Entre as indicações da literatura, podemos enfatizar a importância da identificação e tratamento destas ameaças ainda durante a etapa de planejamento do estudo (TEIXEIRA; FONSECA; SOARES, 2018)..

De acordo com a taxonomia proposta por Cook e Campbell (1979), os tipos de ameaças à validade de um experimento podem ser categorizados em: validade de conclusão, validade interna, validade de construção, e validade externa. Cada um dos quatro tipos está relacionado a uma questão metodológica na experimentação.

Para a gestão das ameaças à validade deste experimento, utilizamos o processo proposto PrioriTTVs e a ferramenta ValidEPlan como apoio a identificação, análise e priorização das principais ameaças à validade e suas respectivas ações de mitigação para cada risco identificado, durante a fase de planejamento do estudo. O conjunto de possíveis ameaças à validade foram organizadas por nome da ameaça, descrição, tipo de ameaça e suas respectivas ações de controle.

O mencionado grupo de ameaças e ações de controle foi classificado pela perspectiva da intensidade do impacto, urgência e tendência que estas poderiam causar aos resultados deste estudo. O quadro da Tabela 23, abaixo mostra as principais ameaças identificadas e como foram tratadas.

Tabela 23 – Principais ameaças identificadas através do processo PrioriTTVs

Validade de Conclusão: está relacionada à capacidade de tirar conclusões corretas e amplas sobre as relações entre os tratamentos usados no estudo e os resultados obtidos.		
Ameaça à validade	Descrição da Ameaça	Ações de controle
1. Tamanho pequeno da Amostra Impacto: 2 Urgência: 2 Tendência: 2 Magnitude: Moderada	Ocorre quando a amostra pode não representar a população alvo em termos de tamanho e com isso, o poder estatístico dos resultados do experimento pode ser baixo.	<ul style="list-style-type: none"> • Amostras pequenas são dificuldades conhecidas e difíceis de superar em ES. Os resultados do experimento foram considerados promissores e por isso, não conclusivos. Novas avaliações com amostras maiores devem ser realizadas.
2. Intervalo de confiança inadequado Impacto: 3 Urgência: 3 Tendência: 3 Magnitude: Muito Alta	Relacionada ao intervalo de confiança aplicado no teste estatístico. Se for inadequado, há um baixo poder estatístico, o que pode comprometer a conclusão dos resultados.	<ul style="list-style-type: none"> • Utilizamos o intervalo de confiança de $\alpha = 0,05$, considerado adequado para obtermos um melhor poder estatístico, não comprometendo a conclusão dos resultados.
Validade Interna: verifica se há algum relacionamento causal entre o tratamento e o resultado obtido, sem influência de outro fator que não é controlado ou medido.		
3. Amostra heterogênea. Impacto: 3 Urgência: 2 Tendência: 2 Magnitude: Muito Alta	Ocorre quando há diferenças entre os participantes referentes ao nível de experiência.	<ul style="list-style-type: none"> • Uso do Quadrado Latino como forma de reduzir o efeito da variabilidade existente do nível de experiência dos participantes. • Caracterizar a experiência do participante através de questionário de perfil • Nivelamento do conhecimento básico sobre ESE e planejamento de experimentos.
4. Treinamento conduzido por pessoas diferentes Impacto: 2	Ocorre quando os participantes são treinados por pessoas diferentes.	<ul style="list-style-type: none"> • Estabelecer procedimento padrão

Urgência: 2 Tendência: 2 Magnitude: Moderada	Assim, pode proporcionar informações diferentes para cada grupo do experimento.	pelos condutores durante o treinamento.
5. Participante pode adquirir conhecimento durante o experimento Impacto: 3 Urgência: 2 Tendência: 2 Magnitude: Muito Alto	Os participantes podem adquirir conhecimento utilizando o material experimental, aplicando os tratamentos, etc. Com isso, podem relembrar características do primeiro tratamento aplicado na primeira execução.	<ul style="list-style-type: none"> Utilização de duas diferentes unidades experimentais em nosso design, para reduzir esta ameaça. Para o efeito de maturação do conhecimento, atenuamos utilizando o quadrado latino.
6. Efeitos de fadiga podem afetar o desempenho do participante Impacto: 2 Urgência: 2 Tendência: 2 Magnitude: Moderado	Se o experimento é longo, os participantes podem se sentir cansados ou ficar entediados. Assim, a execução do experimento pode ser prejudicada.	<ul style="list-style-type: none"> Foi permitido um pequeno intervalo entre as execuções.
7. Comunicação entre os participantes durante os intervalos Impacto: 3 Urgência: 2 Tendência: 2 Magnitude: Muito Alta	Ocorre quando os participantes conversam durante os intervalos do experimento.	<ul style="list-style-type: none"> Foi solicitado aos participantes que não conversassem sobre o experimento durante o intervalo. Não foi permitido acesso ao material experimental durante o intervalo.
8. Idioma da ferramenta ValidEPlan diferente do idioma dos participantes. Impacto: 1 Urgência: 1 Tendência: 1 Magnitude: Muito Baixa	Ocorre quando o idioma do objeto de estudo é diferente do idioma fluente dos participantes.	<ul style="list-style-type: none"> Segundo o questionário de perfil, 95% dos participantes tinham aceitáveis habilidades de leitura e compreensão da língua inglesa, dessa forma, não se considerou relevante essa ameaça. Mesmo assim, assumiu-se o risco.
Validade do constructo: está relacionada ao grau em que as variáveis usadas no estudo conseguem medir com exatidão os conceitos teóricos que se quer medir.		
9. Influência do pesquisador principal Impacto: 3	Ocorre quando o pesquisador consciente ou	<ul style="list-style-type: none"> Para atenuar esse efeito, a execução e análise dos dados do

Urgência: 2 Tendência: 2 Magnitude: Muito Alta	inconscientemente influencia o resultado do estudo.	experimento foi acompanhada por dois pesquisadores.
10. Interação entre tratamentos Impacto: 3 Urgência: 3 Tendência: 3 Magnitude: Muito Alta	Acontece quando o participante usa mais de um tratamento durante o experimento. O primeiro tratamento pode influenciar o desempenho do participante durante a aplicação dos tratamentos seguintes.	<ul style="list-style-type: none"> Os tratamentos utilizados por cada participante foram em dias diferentes. A ordem de atribuição do tratamento a ser utilizado por cada participante foi definido por meio de aleatorização.
11. Viés na elaboração de documentos que servirão como parâmetro para o estudo. Impacto: 3 Urgência: 3 Tendência: 3 Magnitude: Muito Alta	Quando o pesquisador interessado elabora documentos que servirão como parâmetro de comparação de variáveis do estudo.	<ul style="list-style-type: none"> A elaboração do Modelo de Referência e o julgamento de suas ameaças, ações e relacionamentos foi executado por quatro pesquisadores (independentes) experientes em ESE.
Validade externa: está relacionada à capacidade de repetir o mesmo comportamento da pesquisa em outros grupos de participantes, além daqueles em que o estudo foi aplicado.		
12. Replicação do estudo em contextos similares Impacto: 2 Urgência: 2 Tendência: 2 Magnitude: Moderada	Acontece quando o experimentador não indica onde os dados brutos estão disponíveis, complicando a análise e/ou replicação por outros pesquisadores.	<ul style="list-style-type: none"> Foi indicado onde os dados brutos estão disponíveis para análise por outros pesquisadores, revisores ou auditores independentes.
13. Representatividade dos participantes Impacto: 2 Urgência: 2 Tendência: 2 Magnitude: Moderada	Ocorre quando a amostra dos participantes não representa necessariamente a população para a qual o pesquisador quer generalizar os resultados. A amostra pode não representar a população alvo em termos de conhecimento, experiência, habilidade, tamanho, dentre outros.	<ul style="list-style-type: none"> Foram selecionados pesquisadores iniciantes com pouca experiência com experimentos controlados. Os participantes possuem conhecimento similar do público alvo. Assume-se que a generalização é limitada.
14. Representatividade dos Artefatos	Ocorre quando os artefatos experimentais não	<ul style="list-style-type: none"> Os artefatos foram analisados por

Impacto: 2 Urgência: 2 Tendência: 2 Magnitude: Moderada	representam os artefatos do contexto real do estudo, em termos de complexidade e tamanho.	pesquisadores experientes em ESE.
--	---	-----------------------------------

Fonte: Elaboração própria (2019).

Após identificação das principais ameaças à validade e confiabilidade do planejamento experimental, assim como, escolha das respectivas ações de controle e saneamento, considerando cada caso e, também, os possíveis efeitos e reflexos da aplicação das medidas corretivas, segue a Tabela 24 com os possíveis relacionamentos entre ações de controle e ameaças à validade:

Tabela 24 – Possíveis relacionamentos entre Ações de Controle e Ameaças à Validade

Ameaça à validade	Ações de controle	Pode causar
3. Amostra heterogênea	<ul style="list-style-type: none"> • Uso de treinamento para nivelar as habilidades e conhecimento dos participantes 	<ul style="list-style-type: none"> • Ineficácia da ação por treinamento desbalanceado e insuficiente. Atenuamos essa ameaça balanceando o treinamento de acordo com as informações levantadas pelo questionário de perfil.
	<ul style="list-style-type: none"> • Caracterizar a experiência do participante através de questionário de perfil 	<ul style="list-style-type: none"> • Os participantes podem preencher o questionário de forma incorreta ou inverídica. Explicamos a importância do correto preenchimento dos dados aos participantes, ainda assim, assumimos o risco.
6. Efeitos de fadiga podem afetar o desempenho do participante	<ul style="list-style-type: none"> • Permitir pequenos intervalos entre as execuções 	<ul style="list-style-type: none"> • Comunicação entre os participantes durante os intervalos pode enviesar os resultados. Atenuamos o risco solicitando que não conversassem sobre o experimento durante o intervalo e não permitindo acesso ao material experimental durante o intervalo.
10. Interação entre tratamentos	<ul style="list-style-type: none"> • Foi estabelecido que os tratamentos seriam aplicados em dias diferentes 	<ul style="list-style-type: none"> • Comunicação entre os participantes durante os intervalos da aplicação dos tratamentos pode enviesar os resultados. Atenuamos com a utilização de

Ameaça à validade	Ações de controle	Pode causar
		diferentes unidades experimentais e o uso do design experimental quadrado latino.

Fonte: Elaboração própria (2019).

Para aumentar a confiabilidade nos resultados deste experimento identificamos preventivamente os principais riscos que poderiam surgir na execução e nos antecipamos planejando ações de mitigação (ver Tabela 23). No entanto, algumas ações de controle previstas poderiam desencadear outra ameaças e por isso, revemos o planejamento e adaptamos os procedimentos de execução do estudo, conforme mostrado através da Tabela 24.

6.3.3 Interpretação dos Dados do Questionário

Após a realização do experimento, os 20 participantes responderam a um questionário de avaliação quantitativo com o objetivo de avaliar suas percepções de facilidade, satisfação e aprendizagem ao usar a ferramenta ValidEPlan.

O instrumento de coleta de dados (**Apêndice D**) é composto por 21 afirmações construídas seguindo o modelo de escala aditiva tipo Likert (LIKERT, 1932), utilizada amplamente em respostas a questionários de pesquisa de opinião. Este instrumento foi elaborado inspirado no questionário desenvolvido pela pesquisadora Carine Heck (HECK *et al.*, 2017), cujo objetivo foi avaliar a utilização de uma ferramenta remota através da percepção e o grau de concordância de estudantes em relação à solução web proposta.

Os itens das respostas foram avaliados com pesos de 1 a 5. Os entrevistados expressaram seu nível de aceitação ou de rejeição a partir de uma escala com cinco valores numéricos e pontuações assim definidas:

- Concorda Totalmente: 5
- Concorda Parcialmente: 4
- Indiferente (Nem concorda nem discorda): 3
- Discorda Parcialmente: 2
- Discorda Totalmente: 1

Para fins de análise das respostas para as 21 afirmações do questionário,

foram categorizadas nas seguintes subescalas de percepção:

- **Aprendizagem:** Sinaliza por meio do emprego da ferramenta ValidEPlan se houve melhoria da aprendizagem dos participantes e se os conceitos que foram abordados durante a utilização da ferramenta foram melhor compreendidos.

- **Facilidade:** Percepção dos participantes quanto a facilidade de aprender a utilizar a ferramenta ValidEPlan, se houve problemas ao executar as ações e se os termos, palavras e conceitos usados na ferramenta são consistentes.

- **Satisfação:** Medida pela verificação da motivação dos participantes em aprender sobre o planejamento experimental após a utilização da ferramenta, bem como, se o participante recomendaria o uso da ferramenta a outros colegas e a utilizaria novamente em outros planejamentos experimentais.

De forma geral, os constructos aprendizagem, facilidade e satisfação não podem ser medidos e/ou avaliados diretamente de forma clara e precisa (sem erros). No entanto, podem ser medidos e/ou representados por indicadores. As respostas a essas indagações (indicadores), de forma combinada, fornece medida razoável dos construtos (no caso, a aprendizagem, facilidade e satisfação) para um indivíduo.

Com a finalidade de realizar análise de confiabilidade do questionário utilizado, ou seja, verificar o grau em que o total de indicadores das variáveis aprendizagem, facilidade e satisfação é preciso em suas mensurações, foi aplicado o coeficiente de consistência interna *Alfa* de *Cronbach* (MATTHIENSEN, 2010).

Quanto mais próximos de 1 (um) se encontrar o valor do *alfa*, melhor é a consistência interna dos itens avaliados. George e Mallery (2003) apontam alguns parâmetros para avaliação dos coeficientes de *Alfa* de *Cronbach* (TRAVASSOS; BARROS, 2006):

- Coeficiente *Alfa* > 0,9 → Excelente
- Coeficiente *Alfa* > 0,8 → Bom
- Coeficiente *Alfa* > 0,7 → Aceitável
- Coeficiente *Alfa* > 0,6 → Questionável
- Coeficiente *Alfa* > 0,5 → Pobre
- Coeficiente *Alfa* < 0,5 → Inaceitável

A Tabela 25 apresenta os valores de confiabilidade obtidos para o questionário aplicado, em sua totalidade (21 afirmações). O presente instrumento obteve um pressuposto psicométrico de *Alfa* de *Cronbach* de 0,81 que de acordo com critérios de recomendação de confiabilidade estimada pelo Alfa de Cronbach, é considerado

Bom (OPPENHEIM, 2001; MATTHIENSEN, 2010). Portanto, a partir dessa estimativa de fidedignidade, podemos afirmar que o questionário aplicado é consistente.

Tabela 25 – Estatística de Confiabilidade Alfa de Cronbach

Alfa de Cronbach	0,81
Alfa de Cronbach Baseado em Itens Padronizados	0,82
Número de Itens	21

Fonte: Dados da pesquisa (2019).

Os dados obtidos nos questionários foram reunidos de acordo com as três subescalas estabelecidas anteriormente (aprendizagem, facilidade e satisfação) e de acordo com a escala psicométrica de Likert foram apurados os escores médios (EM) para cada uma delas.

Para verificação se as percepções foram positivas ou negativas, através do EM, foram atribuídas as seguintes condições: valores menores que 3 representaram percepções desfavoráveis (negativas) e maiores que 3 percepções favoráveis (positivas). Pelas pontuações definidas, o valor 3 foi considerado “**indiferente**” ou “nem concorda nem discorda”. Os escores médios foram os seguintes:

- Aprendizagem: 4,66
- Facilidade: 3,97
- Satisfação: 4,61

6.3.3.1 Percepção de Aprendizagem

Para a percepção de aprendizagem foram formulados oito itens, cujas análises e escores médios estão representados na Tabela 26. O coeficiente de estimativa de fidedignidade estimada (Alfa de Cronbach) para questionamentos da subescala “Percepção de Aprendizagem” foi de 0,82. Já o escore médio alcançado para os oito itens foi de 4,67 (desvio padrão de 0,604 e variância de 0,387).

Tabela 26 – Análises e Escores para Aprendizagem

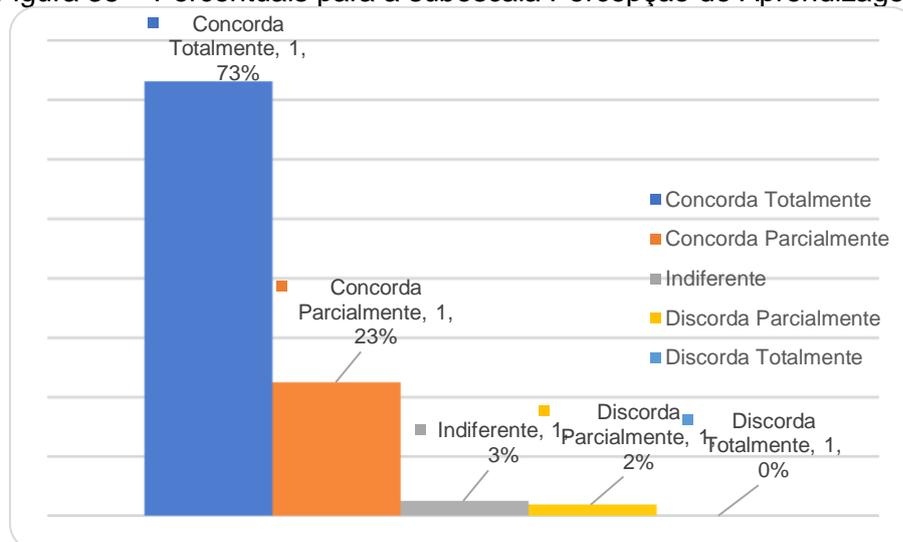
ID	Itens	Média	Desvio	Variância	Alfa
1	A ferramenta ajudou a planejar um experimento de uma melhor forma do que da maneira tradicional.	4,65	0,489	0,239	0,82
2	A ferramenta ajudou a identificar ameaças que não foram identificadas de forma manual.	4,9	0,307	0,095	

ID	Itens	Média	Desvio	Variância	Alfa
3	O uso da ferramenta contribuiu para minha aprendizagem sobre planejamento de experimentos.	4,55	0,759	0,576	
4	O uso da ferramenta contribuiu para minha aprendizagem sobre identificação e controle de ameaças à validade durante o planejamento experimental.	4,8	0,695	0,484	
5	O uso da ferramenta melhorou minha compreensão dos conceitos teóricos que foram abordados sobre planejamento de experimentos.	4,5	0,606	0,368	
6	O uso da ferramenta melhorou minha compreensão dos conceitos teóricos que foram abordados sobre identificação e controle de ameaças à validade.	4,7	0,571	0,326	
7	O uso da ferramenta ajudou a relacionar os conceitos estudados em sala de aula com a prática de experimentação.	4,55	0,604	0,366	
8	As habilidades adquiridas com o uso da ferramenta foram importantes para minha aprendizagem sobre planejamento de experimentos e identificação e controle de ameaças à validade.	4,7	0,801	0,642	
	Totais	4,67	0,604	0,387	

Fonte: Dados da pesquisa (2019).

O gráfico da Figura 35 apresenta os percentuais para as respostas dos participantes em relação aos oito itens de composição da subescala “Percepção de Aprendizagem”. Onde 96% dos respondentes concordaram parcialmente ou concordaram totalmente com as afirmações sobre a melhoria da aprendizagem a partir da utilização da ferramenta ValidEPlan.

Figura 35 – Percentuais para a subescala Percepção de Aprendizagem



Fonte: Dados da pesquisa (2019).

6.3.3.2 Percepção de Facilidade

Para a percepção de facilidade foram formulados seis itens, cujas análises e escores médios estão representados na Tabela 27. O coeficiente de estimativa de confiabilidade estimada (Alfa de Cronbach) para os questionamentos da subescala “Percepção de Facilidade” foi de 0,70. Já o escore médio alcançado para os seis itens foi de 3,98 (desvio padrão de 0,820 e variância de 0,715).

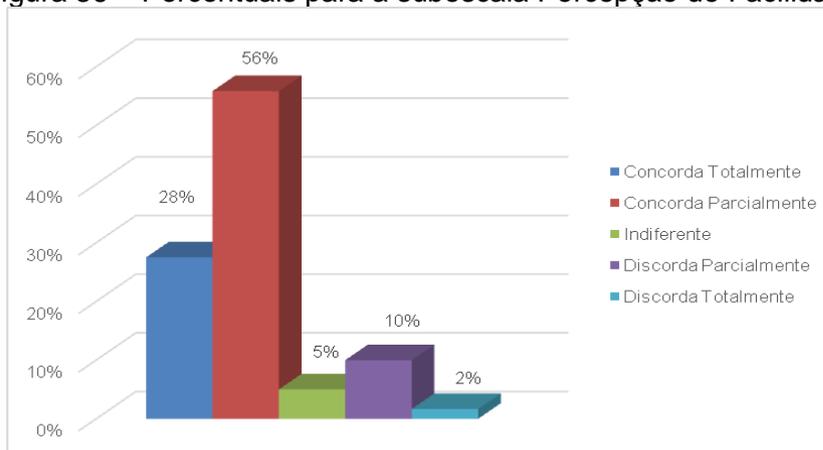
Tabela 27 – Análises e Escores para Facilidade

ID	Itens	Média	Desvio	Variância	Alfa
9	Foi fácil aprender a utilizar a ferramenta.	3,95	0,825	0,682	0,70
10	Foi fácil usar a ferramenta.	4,20	0,695	0,484	
11	A execução de uma tarefa na ferramenta leva a um resultado previsível.	3,40	0,994	0,989	
12	Os termos, palavras e conceitos usados na ferramenta são consistentes.	4,40	0,753	0,568	
13	Não encontrei problemas para executar as ações que pretendia realizar utilizando a ferramenta.	3,40	1,140	1,305	
14	As informações explicativas contidas na ferramenta contribuíram para manuseá-la.	4,50	0,512	0,263	
Totais		3,98	0,820	0,715	

Fonte: Dados da pesquisa (2019).

O gráfico da Figura 36 apresenta os percentuais para as respostas dos entrevistados em relação aos seis itens de composição da subescala “Percepção de Facilidade”. 84% dos respondentes concordaram parcialmente ou concordaram totalmente com as afirmações sobre facilidade de utilizar a ferramenta ValidEPlan.

Figura 36 – Percentuais para a subescala Percepção de Facilidade



Fonte: Dados da pesquisa (2019).

6.3.3.3 Percepção de Satisfação

Para a percepção de satisfação foram formulados sete itens, cujas análises e escores médios estão representados na Tabela 28. O coeficiente de estimativa de confiabilidade estimada (Alfa de Cronbach) para os questionamentos da subescala “Percepção de Facilidade” foi de 0,70. Já o escore médio alcançado para os sete itens foi de 3,98 (desvio padrão de 0,820 e variância de 0,715).

Tabela 28 – Análises e Escores para Satisfação

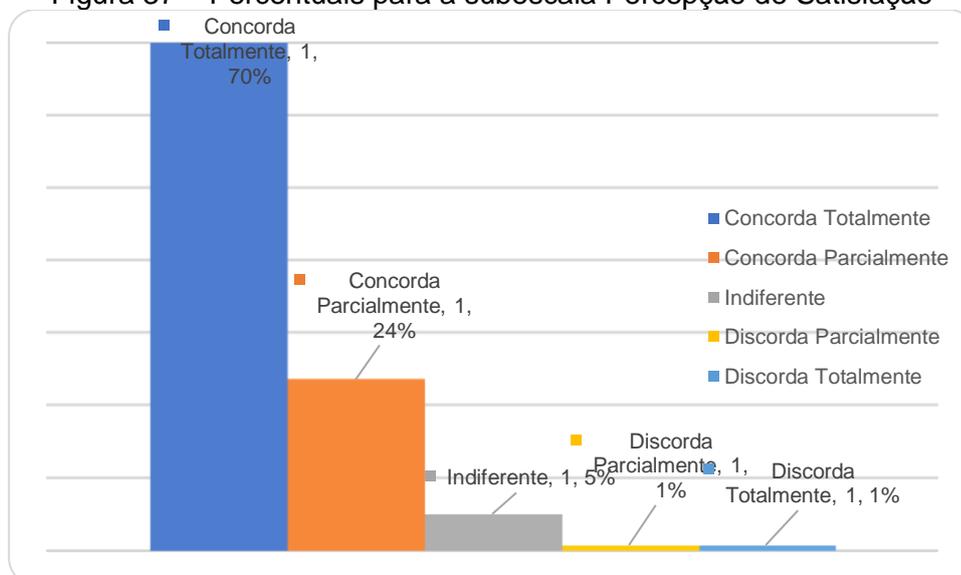
ID	Itens	Média	Desvio	Variância	Alfa
15	De forma geral, fiquei satisfeito com a ferramenta.	4,75	0,444	0,197	0,71
16	A ferramenta foi relevante e útil para a tarefa de planejamento de experimentos e identificação e controle de ameaças à validade.	4,80	0,410	0,168	
17	O uso da ferramenta proporcionou maior motivação em aprender sobre planejamento de experimentos e identificação e controle de	4,40	0,882	0,779	

	ameaças à validade.			
18	Eu indicaria o uso da ferramenta para outros pesquisadores?	4,80	0,410	0,168
19	O uso da ferramenta melhora a comunicação com outros colaboradores e/ou experimentadores.	3,90	1,020	1,042
20	Eu usaria a ferramenta novamente para o planejamento de experimentos.	4,70	0,470	0,221
21	Eu usaria a ferramenta novamente para identificação e controle de ameaças à validade.	4,95	0,223	0,050
	Totais	4,61	0,551	0,375

Fonte: Dados da pesquisa (2019).

O gráfico da Figura 37 apresenta os percentuais para as respostas dos alunos em relação aos sete itens de composição da subescala “Percepção de Satisfação”. 94% dos respondentes concordaram parcialmente ou concordaram totalmente com as afirmações sobre a motivação dos participantes em aprender sobre o planejamento experimental após a utilização da ferramenta ValidEPlan.

Figura 37 – Percentuais para a subescala Percepção de Satisfação



Fonte: Dados da pesquisa (2019).

Respondendo à terceira questão de pesquisa (QP3) desse experimento, para essa amostra, a interpretação dos dados dos questionários mostrou que há uma notável percepção positiva por parte dos participantes, relacionada à aprendizagem,

satisfação e facilidade de uso quanto ao planejamento de um experimento controlado da área de ES com a utilização da ferramenta ValidEPlan.

6.3.4 ETAPA 3: Entrevista Narrativa Episódica

Durante a terceira etapa da avaliação com os alunos, foi aplicado o roteiro de entrevista semiestruturada (**Apêndice E**) contendo seis questionamentos, com o fim de coletar os depoimentos dos participantes, de maneira a analisar indícios referentes à aceitação, suporte à identificação e priorização de ameaças à validade e funcionalidade da ferramenta ValiDEPlan, bem como apontar sugestões e descrever sobre experiências individuais. Por vezes a resposta dada pelo entrevistado não abordou diretamente a pergunta e com isso, deu abertura ao pesquisador refazer a questão de forma mais clara e objetiva do que a inicialmente planejada.

A metodologia de entrevista narrativa episódica, aqui empregada, teve como objetivo obter a percepção explicativa de resultados observados durante o uso da ferramenta ValidEPlan, onde o participante pôde manifestar seus pontos de vista ou a exposição dos efeitos deste fenômeno. Para tal, foram utilizados para a coleta dos dados as seguintes ferramentas: Guia de Entrevista, Gravador de Áudio, nos casos em que a entrevista não pôde ser feita presencialmente, utilizamos o aplicativo *Skype*. Quanto aos procedimentos, foi informado aos participantes que:

- O áudio da entrevista seria gravado;
- Todas as informações obtidas são sigilosas e seus dados pessoais não serão identificados em nenhuma fase/etapa desta pesquisa.
- Os dados são guardados em local seguro e a divulgação dos resultados será feita de forma a não identificar os participantes.
- Não há resposta certa ou errada e que as respostas deveriam ser de forma sincera e espontânea.
- De preferência, evitasse responder às perguntas apenas com “sim” ou “não” e que tentasse lembrar de situações que pudessem exemplificar/justificar sua resposta.

Os dados foram manipulados pelo autor desta tese, supervisionado por dois revisores, orientadores da pesquisa. Uma planilha eletrônica foi usada para facilitar a codificação do conteúdo e análise das respostas às questões.

Considerando as respostas dos alunos, para fins de sigilo e anonimato, identificados por meio de códigos (A1, A2, A3, A4, A5... A20), realizou-se a entrevista e, de maneira complementar, para melhor tratamento das informações, seguimos as etapas de análise de conteúdo proposta por Bardin: (i) análise prévia, (ii) exploração de conteúdo, e (iii) tratamento dos resultados, inferência e interpretação.

Na etapa de análise prévia, foi realizada uma leitura inicial das respostas às perguntas da entrevista no intuito de organizar os dados que posteriormente foram analisados. Na fase de exploração de conteúdo, os textos foram estudados com maior acuidade no intuito de definir as unidades de registro e unidades de contexto. Nesta etapa, encontraram-se 32 temas (unidades de registro) e em seguida os agrupamos em torno de cinco categorias de análise (unidades de contexto).

Na etapa de tratamento e de interpretação dos resultados, os dados foram sintetizados e analisados de acordo com eixos temáticos, ligados a uma das cinco categorias de análise.

Tabela 29 – Frequência dos Temas relacionados à qualidade dos resultados

Categoria de Análise	Eixos Temáticos	% Freq.
1. Qualidade e/ou Validade dos Resultados	Considerações sobre ameaças que só foram identificadas através da ferramenta ValidEPlan	33,3%
	Ponderações sobre revisão do plano experimental a partir das ameaças e ações identificadas	11,1%
	Argumentos sobre maior número de ameaças e ações identificadas entre os tratamentos usados no experimento	11,1%
	Opiniões sobre a diferença do plano experimental elaborado com tratamento ValidEPlan e <i>ad hoc</i>	22,2%
	Observações sobre ameaças genéricas e/ou abrangentes	5,6%
	Reflexões sobre os itens considerados necessários e indispensáveis ao planejamento de um experimento	2,8%
	Alegações sobre a ferramenta no sentido de não permitir ao experimentador esquecer,	13,9%

Categoria de Análise	Eixos Temáticos	% Freq.
	negligenciar ou omitir informações ao plano experimental	

Fonte: Elaboração própria (2019).

A Tabela 29 apresenta a categoria, temas identificados e as frequências das citações analisadas, referente à qualidade e/ou validade dos resultados. Em um contexto experimental em que se evitam vieses e erros amostrais, torna-se decisivo promover direcionamentos para a observância dos fatores de risco à validade dos experimentos em ES, objetivando a identificação e priorização de ameaças à validade existentes e que de fato, possam interferir na qualidade e/ou validade dos resultados dos estudos.

Relacionado a isto, os principais argumentos dos respondentes dizem respeito ao fato de que, comparando os dois tratamentos usados, muitas das ameaças só puderam ser identificadas através da ferramenta ValidEPlan (Freq. 30,3%). Os entrevistados justificaram, também, que há clara diferença entre plano experimental realizado através da ferramenta ValidEPlan e *ad hoc* (Freq. 22,2%), reproduzindo-se o seguinte:

No caso ajuda, até porque na questão das ameaças, na hora de identificar ameaças como “o participante se comportar diferente quando está sendo observado”, não teria sido detectado de forma tradicional e com a ferramenta ela acabou apresentando essa ameaça também e outra questão foi a de “alguns participantes recebem premiação e outros não” (A1).

Ela conseguiu mostrar umas ameaças que eu não iria imaginar, porque muita coisa interfere no experimento e sinceramente algumas das ameaças lá eu nunca teria pensado, então ela ajudou muito em relação a isso. E muitas das ameaças eu achei muito certa vamos dizer assim, eu estava pensando nas minhas ameaças bem específicas só das minhas hipóteses, no meu contexto, nada muito abrangente, porque a aplicação mostrou coisas que eu não tinha imaginado, interferências de terceiros que eu não teria percebido sozinha, ela ajudou em relação a isso. Uma das ameaças que o plano mostrou foi conversa entre os participantes, interferências se uma experiencia se eles poderiam trocar experiências e eu fiquei: sim, mais isso vai interferir? Eu não consegui enxergar isso, depois que a ferramenta mostrou foi que eu analisei e realmente isso pode interferir muito, porque eu não tinha pensado nessa perspectiva, se a pessoa compartilhar essas informações no efeito de aprendizagem, eu não tinha pensado nisso, eu sabia que poderia ter mais eu não tinha pensado que realmente poderia interferir na opinião do outro, eu não tinha pensado nesse lado. (A4)

A quantidade de ameaças que ela consegue detectar foi muito maior em relação a quando eu mesmo tive que tirar de cabeça e escrever as ameaças à validade quando eu fui fazer ad hoc e a forma como ela fica analisando, vendo como você marcou as opções, como você caracterizou o plano pra ela identificar automaticamente as ameaças fez com que depois eu olhasse coisas do plano, que eu revisasse, melhorasse um pouco para poder analisar

de novo, ir um pouco mais a fundo, então você consegue ficar iterando várias vezes para revisar o plano enquanto na ad hoc não seria tão simples fazer essa iteração. (A8)

Assim, uma padronização benéfica também é obtida neste caso, pois a ferramenta sugere, de acordo com uma lógica interna, uma sequência de ameaças que deve ser analisada. Então, principalmente as ameaças que aparecem menos na prática, a ferramenta me ajudou a identificar, coisa que dificilmente aconteceria em um experimento projetado ad hoc. Assim, a ferramenta auxiliou tanto na qualidade, como validade do experimento. (A11)

Promoveu uma melhor qualidade porque a ferramenta sugeriu algumas ameaças à validade que, pela forma tradicional, eu não teria percebido, e acabou fazendo com que eu pensasse em ações de controle para cada ameaça. (A13)

Na verdade a ferramenta garante que pelo menos os fatores que são previsíveis ou possibilidades dentro de todo esse contexto, as principais já estão sendo vistas, pode ser que, em algum experimento específico, a ferramenta não contemple, mas em geral ela acaba cercando, vamos dizer assim, com contramedidas e classificação de ameaças as principais que podem ocorrer durante um experimento. Então devido a isso com certeza os dados e todo o experimento vai ter uma certa veracidade maior do que se eu fizesse de uma forma, vou dizer assim mais empírica, tradicional, baseado no conhecimento aberto, vou dizer dessa forma. (A17)

Nas reproduções acima, citamos algumas das observações dos participantes relacionadas à melhoria dos planos experimentais quando são elaborados através da metodologia empregada na ferramenta ValidEPlan.

Outros argumentos foram utilizados pelos participantes como justificativa à temática, por exemplo: comentários de que a ferramenta ajuda o experimentador a não esquecer, negligenciar ou omitir informações ao plano experimental (Freq. 13,9%), oferece suporte à revisão do plano experimental a partir das ameaças e das ações identificadas pela ferramenta (Freq. 11,1%) e comparada ao tratamento *ad hoc* proporciona identificar maior número de ameaças e ações de mitigação (Freq. 11,1%).

Tabela 30 – Frequência dos Temas relacionados a Identificação de Ameaças à Validade

Categoria de Análise	Eixos Temáticos	% Freq.
2. Identificação de Ameaças e melhores Ações de mitigação	Aspectos relacionados ao surgimento de novas ameaças a partir da aplicação de ações de controle	13,8%
	Pontos de vista relacionados à identificação de ameaças anteriormente negligenciadas e/ou esquecidas.	27,6%
	Reflexões sobre mudança de percepção e/ou	20,7%

Categoria de Análise	Eixos Temáticos	% Freq.
	comportamento após a identificação de ameaças	
	Opiniões sobre a inserção de novas ameaças e ações de controle que não foram identificadas pela ferramenta	3,4%
	Sugestão de ameaças e ações de controle consideradas específicas ao plano experimental	10,3%
	Ponderações sobre as ações de controle que de fato ajudam a mitigar os problemas	24,1%

Fonte: Elaboração própria (2019).

A Tabela 30 apresenta a categoria, temas identificados e as frequências das citações analisadas, referentes à identificação de ameaças e melhores ações de mitigação. Neste aspecto, para que conclusões corretas sejam extraídas de um experimento, os pesquisadores devem identificar e tratar adequadamente os fatores de risco que podem fazer com que os resultados do estudo sejam mal interpretados, comprometendo a confiabilidade e validade e, assim, tenhamos resultados úteis e aplicáveis (DIAS-NETO, 2004; MAFRA; TRAVASSOS, 2006; FELDT; MAGAZINIUS, 2010; NETO; CONTE, 2013).

Relacionado a isto, os principais argumentos dos entrevistados dizem respeito ao fato de que, comparando os dois tratamentos usados, a ferramenta ValidEPlan ajudou o pesquisador lembrar de ameaças e ações anteriormente negligenciadas e/ou esquecidas durante a execução do tratamento tradicional (Freq. 27,6%).

Os entrevistados justificaram, também, que as ações de controle sugeridas pela metodologia empregada na ferramenta de fato ajudaram a mitigar problemas identificados (Freq. 24,1%), reproduzindo-se o seguinte:

A ferramenta ajudou bastante porque além de identificar ameaças que você normalmente, por algum motivo não identificou, ela ainda lhe dá as opções de controle, ou seja, o usuário não tem que pensar muito e também descreve uma ação de controle. Ela já lhe dá as opções e o mais interessante é que ela acaba, dependendo do que você escolher da ação de controle, ela acaba lhe avisando se existe o perigo de introduzir uma nova ameaça ao experimento. (A1)

Bom, a ferramenta ajudou sim, quando ela sugere uma ameaça, ela faz você refletir um pouco sobre aquele ponto e as sugestões são legais. Por exemplo tem uma lá que era sobre fadiga dos participantes aí tinha lá uma sugestão que era pra você dizer o que fazer, tinha sugestões do que fazer, aí tinha lá descanso, dá um descanso pros participantes, aí assim, a sacada que ele dá é já sugerir outra ameaça que vem daquela tua ação de controle, então os

participantes iriam se comunicar aí essa outra ameaça já tinha algumas outras formas de você combater ai é legal que ele trabalha um pouco isso ai. (A3)

As ações de controle que ela mostrava eram realmente coisas bem diretas para resolver o meu problema, sugestões bem diretas, conseguiu me mostrar a solução que eu conseguiria obter. (A4)

Ela realmente ajudou. No total só no meu plano foram 29 ameaças e com a ajuda das ações sugeridas pelo sistema ficou bem mais fácil a definição das ações em si a serem tomadas. (A5)

Achei bastante interessante a funcionalidade da ferramenta, confesso que algumas ações de controle eu não tinha pensado, mas o fato de eu ter colocado alguma coisa lá que ele disse: “olha, a gente vai ter que acrescentar no plano isso aqui e aquilo outro” ela faz automaticamente, ela acrescenta algumas informações, eu não tinha pensado definitivamente daquela forma, então isso pra mim foi bastante interessante. (A9)

No caso das ações a se tomar para mitigar as ameaças identificadas, a ferramenta ajudou bastante pois ela identifica e sugere uma linha de raciocínio a ser tomada para a mitigação das ameaças encontradas. Novamente, por ser um processo sistemático, ele auxilia a que nenhum passo do processo seja esquecido ou negligenciado, auxiliando principalmente o experimentador inexperiente. (A11)

Tinha um parâmetro lá, eu não me lembro exatamente, era pra controlar a quantidade de ações de controle que você queria dependendo do nível de criticidade, não sei a palavra que eu uso, qual o mais crítico mais ou menos, as ações de controle apareciam e eu controlei aquele negócio, tipo quando você começa a considerar muitas, qualquer tipo de ameaças independentemente do nível delas, ou seja, ameaças que tinham um nível crítico bem baixo ele começava a gerar muitas ações de controle desnecessária, eu achava, mas quando você começa a regular esses parâmetros para tipo, filtrar essas ações de controle de nível mais baixo sobravam mais ações de controle que eram mais interessantes para o uso. (A18)

Outros aspectos sobre a identificação de ameaças e ações de mitigação pela ferramenta ValidEPlan também foram citados pelos respondentes. Por exemplo, um aspecto importante que é a mudança de percepção e/ou comportamento do usuário após a identificação de ameaças (Freq. 20,7%), bem como, a citação de argumentos relacionados ao surgimento de novas ameaças a partir da aplicação de ações de controle (Freq. 13,8%).

Tabela 31 – Frequência dos Temas relacionados às percepções da ferramenta

Categoria de Análise	Eixos Temáticos	% Freq.
3. Percepções da ferramenta ValidEPlan: Aspectos positivos	Aspectos referentes ao suporte à identificação de ameaças e ações de controle	25,6%
	Concepções sobre o controle e sistematização do processo de gestão de ameaças à validade	7,0%

Categoria de Análise	Eixos Temáticos	% Freq.
	Argumentos relacionados à ferramenta como um guia, roteiro e fluxo para pesquisadores iniciantes.	30,2%
	Opiniões referentes ao suporte à priorização (classificação) das ameaças e ações de mitigação	9,3%
	Considerações sobre relatórios de ameaças e ações de controle	7,0%
	Posições sobre a usabilidade, organização e padronização da ferramenta	16,3%
	Reflexões sobre a disponibilidade e funcionalidade da ferramenta web.	4,7%

Fonte: Elaboração própria (2019).

A Tabela 31 mostra a frequência dos temas relacionados aos aspectos positivos da compreensão dos usuários que utilizaram a ValidEPlan. O principal argumento dos entrevistados descreve a ferramenta como um guia, roteiro e fluxo para pesquisadores iniciantes (Freq. 30,2%).

Os alunos também justificaram como ponto positivo da ferramenta o fato desta ser um importante suporte à identificação de ameaças e ações de controle (Freq. 25,6%), reproduzindo-se o seguinte:

Eu acho que de principal, ela é um excelente guia principalmente pra quem está começando. Só o fato de você ter um roteiro ali, ter etapas pra você cumprir elas te ajudam a organizar esse processo de construir o experimento, principalmente para quem tem dificuldade de dar os passos iniciais. Então em vez de você sair pra pesquisar um monte de coisas pra construir o roteiro do experimento e das ameaças, ela já está ali com as sessões, já tem as explicações, já tem as referências, tudo pra você só sentar ali e começar a escrever. (A3)

Ela tem um guia, ela me mostra um guia do que eu preciso fazer, mesmo eu não sabendo exatamente como o meu experimento vai ser no final ele vai lá me dando sugestões de ameaças e aquelas considerações que me ajudam a tomar conhecimento do que eu preciso fazer para fazer um bom plano. (A4)

A ferramenta em si, no geral está bem organizada, segue um fluxo que é comumente seguido na preparação de um plano experimental, possui os elementos que visam ajudar o experimentador na preparação do plano além das sugestões de ameaças e ações de controle sugeridas. (A5)

Acho que essa parte das ameaças, a questão também de ela agilizar o processo, querendo ou não quando você vai fazer bota em uma planilha, bota em um arquivo ai fica um monte de arquivos diferentes, ali está tudo

aglutinado, tudo está bem estruturado, também dá um norte pra pessoa, eu acho que não exige um nível tão aprofundado de experimentador, mas também não acho que ela é básica. Mas assim, um ponto positivo é que ela não exige tanto conhecimento prévio. (A6)

Eu confesso que a dificuldade de uso é zero e existe uma pergunta na avaliação final que é: “ela ajuda pessoas que são inexperientes com experimentos a iniciar um experimento?” totalmente, eu não vi dificuldade absoluta em usar, acho que as explicações são muito tranquilas do ponto de vista do entendimento... Achei a ordem também superintuitiva, a ordem que é apresentada a estrutura você consegue montar na cabeça o plano, a estrutura do plano, o esqueleto digamos assim, eu conseguia no próprio treinamento, na sala de aula, eu consegui perceber mesmo em branco eu já consegui ver a estrutura dele na minha cabeça, e dizer assim: no primeiro eu errei a ordem das sessões ou a sessão tal que teve o número incorreto, então eu tive um conforto enorme com o uso, quando você pergunta sobre percepções positivas, na minha concepção é sempre mais, dificuldade zero. (A9)

Eu achei ela de fácil uso, não é nada muito complicado nada que precise de muito treinamento em si, é bem intuitiva e outro fato é que ela ajuda muito quem nunca teve contato ou quem teve pouco contato ou até quem teve muito também, com elaboração de planos experimentais. Então ela ajuda muito nesse ponto de construir um plano experimental, de servir de guia, de propor dicas legais e outro fato importante é no tocante as ameaças e as ações de controle, o fato dela propor inúmeras ai, que a gente como experimentador não consegue, as vezes, nem vislumbrar esses diversos cenários que podem ocorrer, a ferramenta auxilia muito nisso ai. (A12)

As percepções positivas foram que a ferramenta me ajudou bastante a definir um planejamento de experimento de software me apresentando já um fluxo utilizado na literatura e me apresentou diversas ameaças à validade que eu não tinha me atentado pela falta de experiência em planejamento de experimento de software. (A13)

Sem dúvida o que tem de positivo é essa identificação meio que automática ou com base nas informações que a gente coloca naquela caixinhas “radiobottom” ou “checkbox”, nas questões mais objetivas quando a gente está construindo o planejamento e a partir daí gerar as possíveis ameaças que estão presentes no estudo, isso é sem dúvidas é uma coisa que eu considero que foi muito positiva. (A14)

Eu fiquei bastante satisfeita com a utilização da ferramenta por ela ajudar e fornece um suporte à criação do plano e a identificar as ameaças, então isso foi bastante gratificante porque eu tive como saber o que eu estava fazendo, se estava correto, se estava dentro do que se esperava para aquele tópico e o que se esperava também dentro de um plano, por exemplo, tinha tópicos que inicialmente eu me questioneei se teria, mais que quando eu fui realizando o passo a passo da ferramenta eu vi que ia estar lá e ficou bastante completo. (A16)

A definição das ameaças, então selecionar, definir quais ameaças podem existir durante o experimento e também as contramedidas para poder sanar essas ameaças. Então o que eu daria uma nota aí de 0 a 100%, a primeira parte de contemplar todos os tópicos isso é completamente essencial, mas o principal de todos é a questão das ameaças e as contramedidas para poder sanar essas ameaças. (A17)

A ferramenta, definitivamente, contribui com quem está iniciando com planejamento de experimentos, mas, além de tudo, é uma ferramenta útil para

identificação de ameaças e suas ações. Se a ferramenta fosse limitada apenas às seções do planejamento, avaliaria como sendo uma ferramenta para iniciantes, mas com o conjunto completo, é útil para qualquer um envolvido com experimentos. (A20)

Nas representações acima, aponta-se algumas observações dos participantes relacionadas aos aspectos positivos da ferramenta ValidEPlan. Outros argumentos também foram utilizados pelos participantes como justificativa a essa temática, por exemplo: posições sobre a usabilidade, organização e padronização da ferramenta (Freq. 16,3%), bem como, opiniões referentes a suporte à priorização (classificação) das ameaças e ações de mitigação (Freq. 9,3%).

Tabela 32 – Frequência dos Temas relacionados às percepções da ferramenta

Categoria de Análise	Eixos Temáticos	% Freq.
4. Percepções da ferramenta ValidEPlan: Aspectos negativos	Opiniões sobre a visualização das seções textuais serem longas e cansativas	34,3%
	Aspectos relacionados à funcionalidade da ajuda da ferramenta	5,7%
	Considerações sobre o feedback oferecido pela ferramenta ao usuário	22,9%
	Tópicos relacionados ao backup e alteração da estrutura do plano experimental	5,7%
	Ponderações sobre salvar automaticamente o plano experimental	14,3%
	Alegações sobre relatórios da ferramenta	17,1%

Fonte: Elaboração própria (2019).

A Tabela 32 acima, mostra a frequência dos temas relacionados a aspectos negativos relacionados ao entendimento dos usuários que utilizaram a ValidEPlan. O principal argumento contraproducente dos entrevistados é em relação a visualização das seções textuais serem longas e cansativas (Freq. 34,3%).

Outro aspecto contraposto pelos entrevistados foi relacionado ao feedback oferecido pela ferramenta ao usuário (Freq. 22,9%), reproduzindo-se o seguinte:

Algumas seções ficaram muito longas, cansativas. Onde, algumas vezes, perdíamos até as respostas no formulário pelo baixo tempo da sessão web. (A2)

Acho que ficou um pouco cansativa na hora que você vai dar o input das informações, eu acho, é como estou falando, eu não estou criticando a metodologia, talvez foi a ferramenta, o resultado final dela. A sacada é boa mas a ferramenta pode dar uma

melhorada, por exemplo, eu achei massa hoje esse negócio de sair arrastando as coisas pra umas cores diferentes e tal, aquele negócio de drive and drop ficou bem bacana aquilo ali, eu acho que poderia ter alguma coisa nesse sentido lá na ferramenta, por que estou dizendo isso? Vou dar um exemplo prático, a gente abre uma tela que tem 20 perguntas de caixa de texto e área, campos abertos pra você digitar ai você bota um avançar lá e tem mais 20, isso é um pouco exaustivo, talvez se vocês pudessem fazer uma, é como estou falando, eu acho que isso ai já é uma crítica com relação a usabilidade mesmo da ferramenta, talvez se vocês pudessem fazer uma interação diferente em cada transição de tela, explorar mais um pouco a interface gráfica, para não ficar um coisa tão repetitiva, porque já é cansativo de natureza porque é muita informação para passar. Mas aí não obstante é também a questão de ter muitos campos para informar e tudo da mesma maneira, eu achei muito cansativo. (A6)

O fato de você não poder utilizar tabela, o fato de você não ter uma visão geral, de ser na caixinha de texto e eu também tive um pouco de dificuldade porque eu não sabia que tinha que terminar de marcar todas as seleções para ir para próxima fase, isso me atrasou um pouquinho. (A7)

Tem alguns blocos que tem uma relação entre eles e a gente acaba trocando entre eles frequentemente e como tu escolhesse um site onde esses blocos ficam utilizando o modelo tipo sanfona, eu tinha que ficar fechando um e abrindo outro o tempo todo para ficar conferindo as informações entre eles. (A8)

Uma coisa que eu fiz, quando a gente fala em usabilidade e ai a gente vai um pouco para as heurísticas de Nielsen que é dar o feedback ao usuário, eu não sei se eu não vi ou não tem, eu ficava abrindo as caixinhas para ver quantas perguntas tinham no próximo item para eu calcular o tempo, eu sou muito criteriosa em relação a estruturar o tempo da minha vida, então eu sei exatamente quanto tempo eu vou gastar nisso e aquilo outro, então eu ficava medindo ali, eu ia abrindo o mais Zinho para poder expandir lá os campos para saber quantas perguntas tinham no questionário porque eu não via o andamento, como tem na parte de ameaças, você está em 10 de 50, eu não vi isso na parte do planejamento do experimento então eu acho que eu senti um pouco de falta disso. E com relação ao feedback ao longo do plano, “você está em x perguntas do total de y perguntas” acho que ficava interessante também e um review, acertar o revisor, acho que é interessante. No geral acho que é percepções dessa forma, mas de uso, mas eu acho que é só acostumar. (A9)

No caso dos problemas percebidos com a ferramenta, alguns pontos importantes de serem levantados: a intensidade das letras na parte de completção das caixas de diálogo são diminuídas. As letras também foram percebidas como pequenas, e o resultado disto foi uma certa dificuldade para se ler. (A11)

As minhas percepções negativas, elas repousam sobretudo na questão da usabilidade e da interface com o usuário, essa coisa de muitos campos para serem preenchidos com texto livre, um formulário imenso que nunca acaba torna a atividade dolorosa, torna a atividade bastante chata e em alguns momentos a vontade é você sair preenchendo todos os campos o mais rápido possível, só para sair daquele formulário. Então assim, tem esses pontos de interface com o usuário, de feedback ao usuário, da questão da visualização dos dados e da sequência de preenchimento pesa bastante negativamente a ponto de eu considerar que tinha feito tudo errado ou que eu não sabia o que eu estava fazendo direito, justamente por não ter um controle, não ter a visão do todo, mais esses defeitinhos mesmo. (A14)

Outros aspectos negativos relacionados ao entendimento dos usuários que

utilizaram a ValidEPlan também foram citados pelos respondentes. Por exemplo, alegações sobre problemas encontrados nos relatórios da ferramenta (Freq. 17,1%), sobre salvar automaticamente o plano experimental enquanto este estava sendo elaborado (Freq. 14,3%), assim como, aspectos contrários à funcionalidade da ajuda da ferramenta (Freq. 5,7%).

Tabela 33 – Frequência dos Temas relacionados aos comentários e oportunidades

Categoria de Análise	Eixos Temáticos	% Freq.
5. Comentários e oportunidades: Aspectos gerais	Considerações sobre o oferecimento de exemplos reais de preenchimento de planos experimentais com referências e dicas de ajuda na ferramenta	15,6%
	Reflexões sobre repensar a comunicação textual de apresentação e visualização dos textos da ferramenta	18,8%
	Opiniões sobre ter todo o controle do plano experimental durante a utilização da ferramenta	6,3%
	Alegações sobre o uso da metodologia empregada na ferramenta para a aprendizagem de pesquisadores iniciantes	40,6%
	Ponderações sobre a redução de esforço do pesquisador proporcionada pela ferramenta	12,5%
	Argumentos sobre a indiferença e o efeito prejudicial à aprendizagem	6,3%

Fonte: Elaboração própria (2019).

A Tabela 33 apresenta a categoria de comentários e aspectos gerais, relacionados ao experimento realizado, aprendizagem dos participantes, bem como ao uso da ferramenta ValidEPlan. Relacionado a isto, os principais argumentos dos entrevistados dizem respeito ao uso da metodologia empregada na ferramenta para aprendizagem de pesquisadores iniciantes (Freq. 40,6%), reproduzindo-se, pois:

De forma geral, aumentou meu conhecimento na atividade de planejamento de experimentos. (A2)

Ela ajudou também no aprendizado porque quando eu fui fazer da forma ad hoc e meio que na hora eu consertei muitas coisas que já vieram à mente, as sessões, as coisas que tinham que ter nas sessões já vieram a mente, eu usaria a ferramenta para elaborar um experimento no meu doutorado

tranquilamente, eu até prefiro, vai ajudar bastante. (A3)

Eu nunca tinha visto o experimento em si, nunca tinha tido contato nenhum e foi muito bom porque em maio eu já tinha lido alguns artigos, visto, mas eu nunca tinha vivenciado, eu achei muito bom que tanto a questão do ad hoc quanto a ferramenta conseguiu propor como eu poderia fazer realmente na minha pesquisa do mestrado que até agora eu não tenho conhecimento de como vai ser a pesquisa, mas pelo menos agora eu consigo saber como eu posso usar o experimento de verdade. (A4)

É uma área que eu sempre tive muito interesse de me aprofundar e eu tinha um conhecimento básico, cheguei na disciplina e aprendi muita coisa mesmo, acho que essa abordagem de fazer leitura e comentar em sala eu achei arretado isso e tipo, eu aprendi muita coisa. Eu achava que sabia muito, agora estou terminando a disciplina e vendo que eu não sei tanto assim. Mas eu acho que o conhecimento a partir do uso da estrutura usada na ferramenta aumentou com certeza. (A6)

Eu quando fiz o questionário demográfico eu até me avaliei que meu conhecimento para um planejamento seria baixo, agora eu já acredito que eu conhecimento médio para planejar um experimento. (A7)

Sobre o aprendizado eu percebi mais quando eu fui para parte ad hoc que não dava para desconiliar o que eu já tinha feito porque quando eu fui utilizando a ferramenta eu fui amadurecendo os conceitos, então eu meio que acabava seguindo mentalmente o mesmo conceito, mas acabei olhando um outro template e seguindo esse template fazendo uma mescla dos dois, mas deu para memorizar bem, associar bem como escrever um plano, mesmo que manualmente, depois que a gente usa a ferramenta. Então eu acho que ela contribui bastante pro aprendizado. (A12)

Ajuda, principalmente quanto a aprendizagem do pesquisador iniciante, eu mesmo que nunca tive esse contato ela me ajudou, contribuiu bastante. (A19)

De maneira geral, gostei bastante de utilizar a ferramenta e fixar os conteúdos visto na disciplina. Espero que seja finalizada e divulgada em breve no CIn. (A20)

Os entrevistados também sugeriram repensarmos a comunicação textual de apresentação e visualização dos textos da ferramenta (Freq. 18,8%), reproduzindo-se o seguinte:

Como o planejamento pode ser realizado por várias pessoas de um time, talvez repensar a comunicação textual dentro da ferramenta poderia gerar algum ganho. Dividir as seções em seções com menores quantidades de questões. (A2)

Melhoria para mim eu acho que seria em questão de informação textual, tipo, eu tive uma dificuldade eu demorei muito tempo na tela e quando fui salvar eu perdi o que eu tinha feito, ele não informa tempo, é uma questão de usabilidade informar que aquela tela pode expirar em tanto tempo essas questões de usabilidade em si. (A4)

Como sugestões dar a continuidade do aperfeiçoamento da ferramenta. Não existe outra ferramenta parecida no mercado que seja do conhecimento que existe bastante perspectiva para que esta ferramenta se torne um importante paradigma no estudo do projeto de experimentos. Também como sugestão gostaria de citar os problemas mencionados na resposta da questão anterior,

quais sejam a intensidade da fonte e as cores e a correção do bug detectado na parte de remédios às ameaças. (A15)

Primeiro organizar essas sessões para ver se realmente você encontra partes repetidas, coisas que você acaba repetindo ao longo da escrita do planejamento experimental e a segunda é ter alguma forma de adicionar anexos, adicionar tabelas e adicionar subseções caso o planejador queria adicionar, tipo ter formas de escrever textos mais estruturados na ferramenta. (A18)

Nas representações, apontamos algumas das observações dos participantes relacionadas aos comentários e aspectos gerais, relacionados ao experimento realizado, aprendizagem dos participantes, bem como ao uso da ferramenta ValidEPlan.

Outras sugestões e argumentos também foram utilizados pelos participantes, por exemplo: considerações sobre oferecimento de exemplos de preenchimento de planos experimentais com referências e um maior número de dicas de ajuda na ferramenta (Freq. 15,6%), assim como, ponderações sobre a redução de esforço do pesquisador proporcionada pela ferramenta (Freq. 12,5%).

6.4 RESUMO

Este é o capítulo de avaliação do processo PrioriTTVs, bem como, da ferramenta ValidEPlan. Apresentamos o planejamento, execução, análise e interpretação dos dados referentes a estudos experimentais com diferentes objetivos e perspectivas a partir do ponto de vista de 26 participantes iniciantes e especialistas em ESE.

Os resultados das avaliações realizadas trazem evidências científicas que o processo e a ferramenta propostos são viáveis e possuem um bom nível de completude, facilidade de uso, percepção de satisfação e aprendizagem. No entanto, se faz necessário a realização de mais estudos de avaliação, com amostras probabilísticas, a fim de que possamos fazer fortes inferências para a população-alvo e com isso os resultados possam ser generalizados.

7 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

“Quem alcança seu ideal, vai além dele”.

Friedrich Nietzsche.

O trabalho apresentou o processo PrioriTTVs e uma ferramenta *web* chamada ValidEPlan que semiautomatiza o processo proposto, ambos para o suporte à identificação, análise e priorização de ameaças à validade e sugestão de ações de mitigação no planejamento de experimentos da área de ES.

Este capítulo de considerações finais, examina as questões de pesquisa investigadas na tese (Seção 7.1), faz uma revisão das contribuições (Seção 7.2), indica as limitações do trabalho (Seção 7.3), oferece direcionamentos em relação a estudos futuros (Seção 7.4) e ao final apresenta um resumo (Seção 7.5).

7.1 ANÁLISE DAS QUESTÕES DE PESQUISA DA TESE

Esta seção apresenta e analisa respostas às questões de investigação desta tese e menciona conclusões com base em evidências científicas observadas durante o desenvolvimento da pesquisa.

Questão 1: Quais são os procedimentos realizados por pesquisadores para identificar e controlar as ameaças à validade ao planejar experimentos controlados em ES? Qual a importância desse processo?

Como estratégia metodológica de busca à resposta da primeira questão de pesquisa, realizamos um *Survey Research* com 115 experimentadores experientes que planejaram e executaram experimentos controlados da área de ES em suas pesquisas, o qual apresentamos seus resultados no Capítulo 3. Através deste survey, descobrimos que, segundo os participantes, a principal estratégia apontada por pesquisadores para gerir ameaças à validade são os *checklists*.

Os *checklists* são listas de verificação apresentadas de forma que os experimentadores verifiquem se algumas das ameaças elencadas nesta relação possam ocorrer em seus experimentos.

Ao perguntarmos aos participantes se usavam guias, listas de verificação, ferramentas de software, ou outras estratégias para identificar e controlar ameaças à

validade de experimentos, 62.11% responderam que sim. Destes, 50% afirmaram que usaram como suporte a essa temática o livro “*Experimentation in software engineering*” de Wohlin *et al.* (2012), onde encontra-se uma lista que contém sete ameaças à validade de conclusão, 13 ameaças à validade interna, dez ameaças à validade de constructo e três ameaças à validade externa. Os demais entrevistados informaram que utilizaram outras referências conforme mostra a Figura 4.

É importante mencionar que as estratégias utilizadas para identificação e controle de ameaças à validade apontadas pelo *survey*, bem como, mencionadas na literatura, não exploram mecanismos que fornecem apoio ao experimentador no sentido de ajudá-lo à priorizar as ameaças mais graves e com isso, implementar ações de controle mais assertivas (TEIXEIRA; FONSECA; SOARES, 2018).

Em geral, os participantes do *survey* enfatizaram o fato de que pouco se sabe e falha-se muito sobre o controle de ameaças à validade de nossos experimentos e muitos pesquisadores continuam a tomar decisões baseadas em crenças e não em fatos, negligenciando o processo de identificação e priorização dos fatores de riscos de seus experimentos.

Quanto à importância do processo de identificação e controle de ameaças à validade ao planejar experimentos, a maioria dos entrevistados do *survey* (78,26%) consideraram extremamente importante identificar e analisar previamente as ameaças à validade de experimentos da área de ES e declararam que identificam e controlam as ameaças à validade para:

- Divulgar à comunidade as possíveis vieses e limitações do experimento;
- Ajustar o plano experimental e;
- Adequar o processo às etapas de execução do experimento.

Questão 2: Quais são as recomendações indicadas por pesquisadores para melhorar a validade de resultados de experimentos da área de ES?

Ainda sobre a estratégia metodológica *Survey Research* e respondendo ao segundo questionamento da pesquisa, os resultados corroboraram a importância de melhorar a confiabilidade e validade das conclusões resultantes de experimentos da área de ES a partir da adoção das seguintes recomendações:

- Identificar e priorizar ameaças à validade e suas respectivas ações de mitigação durante o planejamento do experimento;
- Refazer ou ajustar as etapas do plano experimental a partir das ameaças à

validade e ações de mitigação identificadas;

- Utilizar guias, *checklists* e ferramentas de software que auxiliem o processo de identificação e controle de ameaças à validade;
- Identificar previamente ações de controle que possam causar novas ameaças ao estudo;
- Identificar ordem de prioridade entre tipos de ameaças baseado no propósito de cada experimento, evitando decisões equivocadas ao priorizar uma ameaça em detrimento de outra.

Questão 3: Como o planejamento de experimentos da área de ES pode ser apoiado quanto ao processo de identificação e priorização das ameaças à validade?

Em resposta à terceira questão de pesquisa, definimos e modelamos o PrioriTTVs, um processo que se destina a apoiar os experimentadores a priorizar ameaças à validade e as respectivas ações de mitigação ao planejar experimentos controlados em ES, bem como, um instrumento de caracterização, utilizado para capturar os atributos de um experimento com base em um conjunto de perguntas objetivas, respondidas por cada pesquisador (Capítulo 4).

Também foi desenvolvida uma ferramenta web para viabilizar o processo anteriormente proposto de forma semiautomática aos experimentadores. A ferramenta chama-se ValidEPlan, uma solução web útil ao planejamento de experimentos de ES orientada à validade de seus resultados. A ferramenta ValiDEPlan oferece suporte a identificação, priorização e controle das ameaças à validade já na fase de planejamento, a fim de reduzir os potenciais riscos e problemas que podem ocorrer desde o planejamento, execução e reporte do experimento (Capítulo 5).

Questão 4: Quais são os benefícios e as limitações da abordagem proposta quanto ao processo de identificação e priorização das ameaças à validade?

Com o objetivo de avaliar na prática o processo PrioriTTVs, bem como, a ferramenta ValidEPlan, foram utilizadas abordagens distintas como survey, experimento controlado, questionários e entrevistas a partir da percepção de iniciantes e especialistas em ESE. O processo e ferramenta propostos foram avaliados através de diferentes objetivos, tais como, completude, facilidade de uso, percepção de satisfação e aprendizagem (Capítulo 6).

De acordo com as avaliações experimentais realizadas com os **Pesquisadores Especialistas em ESE**, identificamos os principais **Benefícios** ao processo e a

ferramenta propostos:

- Relevância do processo. 83,3% dos especialistas concordaram plenamente sobre a importância do instrumento utilizado no processo de avaliação das práticas de controle em investigação com experimentos controlados, ao passo que 16,7% concordam parcialmente com este recurso.

- Concordância com os procedimentos adotados. 100% dos especialistas concordaram com a abordagem e critérios utilizados no processo para identificação de ameaças e ações de controle;

- Utilidade da ferramenta. 83,3% dos especialistas consideraram a solução extremamente útil ao processo de análise e priorização de ameaças e ações de controle e 16,6% apenas como útil;

- Aspectos gerais da ferramenta. Classificaram como adequada (mediana 4,5), fácil (mediana 5), satisfatória (mediana 4) e consistente (mediana de 5), numa escala linear de cinco pontos (1 a 5);

De acordo com as avaliações experimentais realizadas com os **Pesquisadores Iniciantes em ESE**, identificamos principais **Benefícios** ao processo e ferramenta propostos.

- Corretude da Tarefa. Ao comparar-se a utilização da ferramenta ValidEPlan com a forma tradicional para identificar ameaças, ações, relacionamentos e definir ordem de prioridade entre ameaças, evidenciou-se 92% do total de respostas corretas atribuídas à ferramenta, contra especificamente, 50% do tratamento *Ad hoc*, demonstrando diferença estatística na corretude da tarefa;

- A ferramenta permite a inserção manual de novas ameaças e ações de controle que não constam entre as sugeridas pela ferramenta;

- Diferença estatisticamente significativa entre quantidade de ameaças e ações de controle usando a metodologia empregada através da ferramenta ValidEPlan comparado ao tratamento *Ad hoc*;

- Notável percepção positiva por parte dos participantes, relacionado à aprendizagem (96%), satisfação (94%) e facilidade (96%) de uso quanto ao planejamento de um experimento com a utilização da ferramenta ValidEPlan;

- Considerações positivas dos participantes relacionados à ferramenta como um guia, roteiro e fluxo para pesquisadores iniciantes.

De acordo com as avaliações experimentais realizadas com os **Pesquisadores**

Especialistas em ESE, identificamos as principais **Limitações** ao processo e ferramenta propostos.

- Falta de integração facilitada com outras ferramentas;
- Necessidade de novas revisões estruturais, coesão e coerência textuais a fim de se evitarem ambiguidades ou dúvidas ligadas ao procedimento de análise;
- Melhorar a padronização de terminologias usadas;
- Necessidade de melhorias de usabilidade da ferramenta.

De acordo com as avaliações experimentais realizadas com os **Pesquisadores Iniciantes** em ESE, identificamos as principais **Limitações** ao processo e a ferramenta propostos.

- Ao usar o tratamento *ad hoc* em seus planos experimentais, os participantes identificaram uma quantidade maior de ameaças e ações específicas ao seu propósito de experimentação;
- Necessidade de evolução da ferramenta no sentido de conseguir capturar mais informações específicas dos planos experimentais;
- Opiniões negativas dos participantes sobre visualização das seções textuais da ferramenta serem longas e cansativas;
- Considerações negativas sobre o *feedback* oferecido pela ferramenta ao usuário.

De forma geral, a estratégia viabiliza, principalmente a experimentadores menos experientes, acesso a um processo estruturado, que possa ajudar no suporte à identificação e priorização de ameaças a validade e suas respectivas ações de mitigação. Apesar dos benefícios e contribuições apontadas por pesquisadores iniciantes e especialistas, pretendemos estender nossa abordagem para gerenciar ameaças específicas para diferentes contextos de experimentos em ES.

7.2 REVISÃO DAS CONTRIBUIÇÕES

As principais contribuições desta pesquisa são relacionadas ao:

- **Processo PrioriTTVs (Capítulo 4):**
 - Reduzir o esforço exigido a experimentadores iniciantes ou experientes ao gerenciar ameaças à validade e ações de mitigação em seus experimentos;

- Fornecer um *checklist* que possa ajudar a capturar informações sobre os principais aspectos de um plano experimental;
 - Apoiar o processo decisório de priorizar ameaças de acordo com o objetivo do plano experimental realizado.
- **Ferramenta ValidEPlan (Capítulo 5):**
- Fornecer suporte semiautomatizado ao planejamento de experimentos controlados da área de ES, dirigido/orientado à validade dos seus resultados;
 - Contribuir para a melhoria do ensino e aprendizagem da ESE, através do uso de uma ferramenta que possa ajudar a relacionar os conceitos estudados em sala de aula com a prática da experimentação;
 - Fornecer um repositório online de ameaças à validade e ações de mitigação, organizado de acordo com as categorias de ameaças.
- **Avaliação com Pesquisadores Especialistas (Capítulo 6):**
- Planejamento e realização de um *Survey Research* para coleta de dados aplicado aos praticantes de experimentos controlados da área de ES com o propósito exploratório de: (i) compreender as ações dos pesquisadores ao realizar os procedimentos para controle das ameaças à validade de experimentos controlados e; (ii) verificar a percepção entre as práticas reais de tratamento de ameaças a validade e as orientações e recomendações existentes na literatura;
 - Planejamento e realização de estudo experimental que objetivou-se avaliar os procedimentos para a identificação e a classificação das ameaças à validade e as suas respectivas ações para controle, além do mapeamento e classificação de ameaças à validade, por meio da verificação dos seguintes objetos: a) instrumento de caracterização de um experimento; b) Processo PrioriTTVs para a priorização e o controle de ameaças à validade e suas respectivas ações de controle e; c) Ferramenta ValidEPlan;
 - Planejamento e execução de questionário estruturado cujo objetivo foi coletar informações de cunho quantitativo e qualitativo ligados às percepções positivas e negativas da ferramenta ValidEPlan

(objeto de avaliação) e, também, coletar informações qualitativas (cognitivas e comportamentais) sobre aceitação do referido instrumento no que diz respeito à percepção da facilidade do seu emprego para identificação das ameaças à validade e sua utilidade nas ações de controle dos fatores de risco para os experimentos controlados;

- Planejamento e execução de roteiro de entrevista narrativa, este com o fim de coletar os depoimentos dos especialistas, de maneira a analisar indícios de funcionalidade, aceitação e a satisfação da ferramenta ValiDEPlan e do Processo PrioriTTVs.

▪ **Avaliação com Pesquisadores Iniciantes (Capítulo 6):**

- Planejamento e execução de experimento controlado para analisar o uso na prática da ferramenta ValidEPlan, avaliando o seu impacto no processo de planejamento e gestão de ameaças à validade de experimentos conduzidos pelos alunos, com respeito à correção da tarefa de planejamento e gerência de ameaças à validade, bem como, verificar o número de ameaças e ações de mitigação ao planejar um experimento controlado da área de ES, comparado ao uso das práticas *ad hoc*;
- Planejamento e execução de questionário estruturado de avaliação para coletar informações de cunho quantitativo e qualitativo ligadas às opiniões sobre o uso da ValidEPlan, com o objetivo de verificar aprendizagem, satisfação e facilidade de uso quanto ao planejamento de um experimento controlado da área de ES com a utilização da ferramenta proposta;
- Planejamento e execução de roteiro de entrevista narrativa episódica, este com o fim de coletar os depoimentos dos alunos, de maneira a analisar indícios de satisfação dos participantes, bem como, as percepções positivas e negativas sobre a facilidade de uso da ferramenta, bem como apontar sugestões e descrever sobre suas experiências individuais.

7.3 LIMITAÇÕES DO TRABALHO

O trabalho de pesquisa apresentado nesta tese possui limitações, dentre as quais podemos destacar:

- O processo PrioriTTVs e a ferramenta ValidEPlan oferecem suporte limitado ao contexto de planejamentos experimentais gerais, com isso, é preciso estender nossa abordagem para gerenciar ameaças e ações de mitigação específicas para diferentes contextos de experimentos em ES;
- A variedade de planos experimentais usados como unidades experimentais no experimento controlado (Seção 6.3.2) é apenas um subconjunto existente do domínio da ES e por isso, há necessidade de ampliar para abranger um maior número de subdomínios e contextos da área de ES;
- Integração da ferramenta ValidEPlan com outras ferramentas e ambientes de apoio à condução de experimentos da área de ES;
- Necessidade de novas revisões ao processo proposto, no sentido de melhoria da coesão e coerência textuais a fim de se evitarem ambiguidades ou dúvidas ligadas ao procedimento de análise;
- A avaliação realizada por especialistas do mesmo grupo de pesquisa que o pesquisador principal desta tese pode tender a resultados mais positivos, mesmo que inconscientemente;
- Por não possuírem experiência ampla em ESE, apesar de serem um dos perfis de usuário foco do estudo, os resultados demonstrados pela análise dos dados do experimento controlado realizado com experimentadores iniciantes trazer diferentes percepções quanto a avaliação da facilidade, satisfação e aprendizagem quanto ao uso da metodologia utilizada na ferramenta ValidEPlan.

7.4 TRABALHOS FUTUROS

Apesar dos benefícios e contribuições mencionadas, prevemos alguns desafios a serem superados em nossa proposta. As ameaças à validade são muito específicas e estão ligadas diretamente ao propósito e contexto experimental de cada estudo. Dessa forma, é difícil ter um banco de dados único e completo de ameaças e ações de mitigação para cada conjuntura específica de experimentação.

Para mitigar este e outros problemas, sugerimos como trabalho futuro:

- Avaliar o processo e ferramenta propostos através de experimentos controlados em contextos específicos da área de ES (por exemplo, experimentos

envolvendo usabilidade ou prática ágil de programação em par, etc.), assim como, com amostras maiores;

- Realizar um estudo de mapeamento sistemático para identificar procedimentos, guias e metodologias de apoio ao reporte de ameaças à validade e suas respectivas ações de mitigação em ES e em outras áreas;
- Realizar um estudo de mapeamento sistemático para identificar ameaças à validade e ações de controle de experimentos para diferentes contextos em ES, orientados a humanos e a tecnologia;
- Utilizar tecnologias de processamento de linguagem natural para análise das características dos experimentos e implementar melhorias ao processo;
- Realizar amplo estudo qualitativo (entrevistas) com especialistas em ESE para um maior entendimento sobre como planejam estudos orientados a humanos e a tecnologia, bem como, gerenciam as ameaças à validade e ações de mitigação;
- Evoluir a ferramenta ValidEPlan com base nas sugestões de melhoria identificadas nas avaliações realizadas, assim como, integrá-la a outras ferramentas e ambientes de apoio à condução de experimentos da área de ES;
- Ampliar as contribuições do processo e ferramenta propostos no sentido de identificar, classificar e priorizar ameaças à validade e ações de mitigação para outros tipos de estudos (por exemplo: *surveys*, estudos de caso, dentre outros).
-

7.5 RESUMO

Chegamos ao final do capítulo de considerações finais e trabalhos futuros desta tese. A pesquisa aqui apresentada tratou do desafio que é reportar resultados de experimentos da área de ES confiáveis e válidos.

Propomos um processo estruturado chamado PrioriTTVs, como também desenvolvemos ferramenta web para viabilizar o processo de forma semiautomática aos pesquisadores, cujo objetivo foi apoiá-los, principalmente os iniciantes, na identificação e priorização de ameaças à validade e suas correspondentes ações de controle no contexto geral de um planejamento experimental em ES.

Tanto o processo quanto a ferramenta foram avaliados através de survey, experimento controlado, questionários e também entrevistas na perspectiva de 26 participantes iniciantes e experientes. Os resultados das avaliações realizadas trazem evidências científicas que o processo e a ferramenta propostos são viáveis e possuem

um bom nível de completude, facilidade de uso, percepção de satisfação e aprendizagem.

REFERÊNCIAS

- ALMEIDA, A. *et al.* Mecanismos para guiar estudos empíricos em engenharia de software: mapeamento sistemático. *In: EXPERIMENTAL SOFTWARE ENGINEERING LATIN AMERICAN WORKSHOP (ESELAW), 8., 2011. Proceedings [...]. [S. l.]: ESELAW, 2011.*
- ANDERSSON, C. *et al.* An experimental evaluation of inspection and testing for detection of design faults. *In: 2003 INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING, 2003, [s. l.]. Proceedings [...]. [S. l.]: IEEE, 2003. p. 174-184.*
- BARDIN, L. **Análise de conteúdo.** Lisboa: Editora Edições, 2009.
- BARROS, M.; DIAS-NETO, A. Desenvolvendo abordagem sistemática para avaliação dos estudos experimentais em search based software engineering. *In: Workshop de Engenharia de Software Baseada em Buscas-WESB, 2., 2011, São Paulo. Anais [...]. São Paulo: WESB, 2011. p. 49_56.*
- BASILI, V.; SELBY, R.; HUTCHENS, D. Experimentation in software engineering. **IEEE Transactions on software engineering**, [s. l.], n. 7, p. 733-743, 1986.
- BAUER, M.; GASKELL, G. **Pesquisa qualitativa com texto, imagem e som: um manual prático.** Rio de Janeiro: Editora Vozes Limitada, 2002.
- BORGES, A., *et al.* Support mechanisms to conduct empirical studies in software engineering: a systematic mapping study. *In: INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING, 19., 2015. Proceedings [...]. [S. l.]: ACM, 2015. p. 22-29.*
- BOX, George; HUNTER, W. Coautor; STUART, J. Coautor. **Statistics for experimenters: an introduction to desing, data analysis, and model building.** New Jersey: John Wiley & Sons, 1978.
- CAMPBELL, D. Experimental and quasi-experimental designs for research on teaching. **Handbook of research on teaching**, [s. l.], v. 5, p. 171-246, 1963.
- CARVER, J.; VANVOORHIS, J.; BASILI, V. Understanding the impact of assumptions on experimental validity. *In: INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING, 2004, [s. l.]. Proceedings [...]. [S. l.]: IEEE, 2004. p. 251-260.*
- CHAVA, N.; NACHMIAS, D. **Research Methods in the social science.** 5 ed. New Yourk: St Martin's Press, 1996.
- CONTE, T., *et al.* Usability evaluation based on Web design perspectives. *In: INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT, 1., 2007, [s. l.]. Proceedings [...]. [S. l.]: IEEE, p. 146-155, 2007.*
- COOK, T.; CAMPBELL, D. **Design and analysis issues for field settings.** McNally, Chicago: Rand, 1979.

CORBIN, J.; STRAUSS, A. **Basics of qualitative research: Techniques and procedures for developing grounded theory**. California: Sage publications, 2014.

CORTÉS, M., *et al.* Utilização de Estudos Empíricos em Engenharia de Software. *In: COMPUTER ON THE BEACH*, 2018, [s. l.]. **Anais [...]**. [S. l.: s. n.], 2018. p. 472-481.

CRESWELL, J. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto**. Porto Alegre: Artmed, 2010.

DAVIS, F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. **MIS quarterly**, [s. l.], v. 13, n. 3, p. 319-340, 1989.

DAYCHOUW, M. **40 Ferramentas e Técnicas de Gerenciamento**. Rio de Janeiro: Braspot, 2007.

DIAS-NETO, A. *et al.* Infrastructure for SE Experiments Definition and Planning. *In: EXPERIMENTAL SOFTWARE ENGINEERING LATIN AMERICAN WORKSHOP (ESELAW'04)*, 1., 2004, [s. l.]. **Proceedings [...]**. [S. l.]: ESELAW, 2004.

DIESTE, O. *et al.* Quantitative determination of the relationship between internal validity and bias in software engineering experiments: Consequences for systematic literature reviews. *In: INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT*, 2011, [s. l.]. **Proceedings [...]**. [S. l.]: IEEE, p. 285-294, 2011.

DYBÅ, T.; DINGSØYR, T. Strength of evidence in systematic reviews in software engineering. *In: ACM-IEEE international symposium on Empirical software engineering and measurement*, 2., 2008, [s. l.]. **Proceedings [...]**. [S. l.]: ACM, 2008. p. 178-187.

DYBÅ, T.; KAMPENES, V.; SJØBERG, D. A systematic review of statistical power in software engineering experiments. **Information and Software Technology**, [s. l.], v. 48, n. 8, p. 745-755, 2006.

EASTERBROOK, S., *et al.* Selecting empirical methods for software engineering research. **Guide to advanced empirical software engineering**, Springer, London, p. 285-311, 2008.

EUSTÁQUIO, Geraldo. Tentar e falhar é, pelo menos, aprender. Não chegar a tentar é sofrer a inestimável perda do que poderia ter sido. **Dicocitations**, França, 30 maio 2019.

FALCAO, L., *et al.* An analysis of software engineering experiments using human subjects. *In: ACM/IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT*, 2015, [s. l.]. **Proceedings [...]**. [S. l.]: IEEE, 2015. p. 1-4.

FELDT, R.; MAGAZINIUS, A. Validity threats in empirical software engineering research-an initial survey. **Seke**, [s. l.], p. 374-379, 2010.

FONSECA, L. **An instrument for reviewing the completeness of experimental plans for controlled experiments using human subjects in software engineering**. 2016. 460 f. Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, Recife, 2016.

FONSECA, L.; TEIXEIRA, E.; SOARES, S. ValiDEPlan: Validity-Driven Software Engineering Experiments Planning Tool. *In: CONFERÊNCIA BRASILEIRA DE SOFTWARE: TEORIA E PRÁTICA*, 10., 2019, [s. l.]. **Anais Estendidos [...]**. [S. l.]: SBC, 2019. p. 102-107.

FREIRE, M. **Formalização de experimentos controlados em engenharia de software**. 2015. 216 f. Tese (Doutorado) – Pós-Graduação em Sistema e Computação, Universidade Federal do Rio Grande do Norte, Natal, 2015.

GABRIEL, P. **Software languages engineering: experimental evaluation**. 2010. 125 f. Tese (Doutorado) – Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, 2010.

GARVEY, P.; LANSLOWNE, Z. Risk matrix: an approach for identifying, assessing, and ranking program risks. **Air Force Journal of Logistics**, [s. l.], v. 22, n. 1, p. 18-21, 1998.

GEORGE, D.; MALLERY, P. **SPSS for Windows step by step: a simple guide and reference**. Boston: La Sierra University, 2003.

GÜNTHER, H. Pesquisa qualitativa versus pesquisa quantitativa: esta é a questão. **Psicologia: teoria e pesquisa**, [s. l.], v. 22, n. 2, p. 201-210, 2006.

GUROV, D.; TRAVASSOS, G.; AMARAL, E. **Introdução à engenharia de software experimental**. São Paulo: Editora Sistemas, 2002.

HAIR, J.; ANDERSON, R.; TATHAM, R. **Análise multivariada de dados**. Tradução de AS Sant'anna e A. Cloves Neto. v. 5. Porto Alegre: Bookman, 2005.

HECK, C. *et al.* **Integração de tecnologia no ensino de física na educação básica: um estudo de caso utilizando a experimentação remota móvel**. 2017. 133 f. Tese (Doutorado) – Programa de Pós-Graduação em Tecnologia da Informação e Comunicação, Universidade Federal de Santa Catarina, Florianópolis, 2017.

HENDERSON, V. *et al.* Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. **PLoS medicine**, [s. l.], v. 10, n. 7, p. e1001489, 2013.

JEDLITSCHKA, A.; CIOLKOWSKI, M. **Guidelines for Empirical Work in Software Engineering**. v. 53. USA: Fraunhofer Institute for Experimental Software Engineering, 2005.

JEDLITSCHKA, A.; CIOLKOWSKI, M.; PFAHL, D. Reporting experiments in software engineering. **Guide to advanced empirical software engineering**, Springer, London, p. 201-228, 2008.

JØRGENSEN, M., *et al.* Incorrect results in software engineering experiments: How to improve research practices. **Journal of Systems and Software**, [s. l.], v. 116, p. 133-145, 2016.

JUDD, C.; HOYLE, R.; HARRIS, M. **Research methods in social relations**. [S. l.]: Thomson Learning, 2002.

JURISTO, N.; MORENO, A. **Basics of software engineering experimentation**. [S. l.]: Springer Science & Business Media, 2013.

KITCHENHAM, B. *et al.* An Evaluation of Quality Checklist Proposals-A participant-observer case study. **EASE**, [s. l.], p. 167, 2009.

KITCHENHAM, B., *et al.* Preliminary guidelines for empirical research in software engineering. **IEEE Transactions on software engineering**, [s. l.], v. 28, n. 8, p. 721-734, 2002.

KITCHENHAM, B.; PFLEEGER, S. L. Principles of survey research. Parts 1 to 6. **ACM SIGSOFT Softw. Eng. Notes**, [s. l.], v. 28, p. 24-27, 2003.

KITCHENHAM, B.; PICKARD, L.; PFLEEGER, S. Case studies for method and tool evaluation. **IEEE software**, [s. l.], v. 12, n. 4, p. 52-62, 1995.

KO, A.; LATOZA, T.; BURNETT, M. A practical guide to controlled experiments of software engineering tools with human participants. **Empirical Software Engineering**, [s. l.], v. 20, n. 1, p. 110-141, 2015.

LIKERT, R. A technique for the measurement of attitudes. **Archives of psychology**, [s. l.], v. 40, 1932.

LIMA, V.; NETO, A.; EMER, M. Investigação Experimental E Práticas Ágeis: Ameaças À Validade De Experimentos Envolvendo A Prática Ágil Programação em Par. **Revista Eletrônica De Sistemas de Informação**, [s. l.], v. 13, n. 1, p. 1, 2014.

LOZARES, C.; VERD, J. La entrevista biográfico-narrativa como expresión contextualizada, situacional y dinámica de la red socio-personal. **Redes – Revista hispana para el análisis de redes sociales**, [s. l.], v. 15, n. 2, p. 95-125, 2008.

MAFRA, S.; TRAVASSOS, G. Estudos Primários e Secundários apoiando a busca por Evidência em Engenharia de Software. **Relatório Técnico ES**, Rio de Janeiro, v. 687, n. 06, 2006.

MALDONADO, M. *et al.* A framework for software engineering experimental replications. *In*: IEEE INTERNATIONAL CONFERENCE ON ENGINEERING OF COMPLEX COMPUTER SYSTEMS (ICECCS 2008), 13., 2008, [s. l.]. **Proceedings** [...]. [S. l.]: IEEE, 2008. p. 203-212.

MATTHIENSEN, A. **Uso do Coeficiente Alfa de Cronbach em Avaliações por Questionários**. Boa Vista, RR: Embrapa Roraima-Documentos (INFOTECA-E), 2010.

MEIRELES, M. **Ferramentas administrativas para identificar, observar e analisar problemas**. v. 2. São Paulo: Arte & Ciência, 2001.

MIAN, P.; TRAVASSOS, G.; ROCHA, A. A computerized infrastructure for supporting experimentation in software engineering. *In: EXPERIMENTAL SOFTWARE ENGINEERING LATIN AMERICAN WORKSHOP–ESELAW*, 2., 2005. **Proceedings** [...]. [S. l.: s. n.], 2005.

MOHER, D. *et al.* Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. **Controlled clinical trials**, [s. l.], v. 16, n. 1, p. 62-73, 1995.

MONTGOMERY, D. **Design and analysis of experiments**. New Jersey: John Wiley & Sons, 2017.

MOREIRA, J.; MATTOS, G.; REIS, L. Um Panorama da Presença Feminina na Ciência da Computação. *In: REDOR*, 18., 2014, Recife. **Anais [...]**. Recife: UFRPE, 2014.

MUYLAERT, C., *et al.* Entrevistas narrativas: um importante recurso em pesquisa qualitativa. **Revista da Escola de Enfermagem da USP**, São Paulo, v. 48, p. 184-189, 2014.

NAKAGAWA, E., *et al.* **Revisão sistemática da literatura em Engenharia de Software: teoria e prática**. Rio de Janeiro: Elsevier Brasil, 2017.

NETO, A.; CONTE, T. A conceptual model to address threats to validity in controlled experiments. *In: INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING*, 17., 2013, [s. l.]. **Proceedings [...]**. [S. l.]: ACM, 2013. p. 82-85.

OPPENHEIM, A. **Questionnaire design, interviewing and attitude measurement**. UK: Bloomsbury Publishing, 2001.

PETERSON, R. A meta-analysis of Cronbach's coefficient alpha. **Journal of consumer research**, [s. l.], v. 21, n. 2, p. 381-391, 1994.

PFLEEGER, S. Experimental design and analysis in software engineering. **Annals of Software Engineering**, [s. l.], v. 1, n. 1, p. 219-253, 1995.

PIMENTEL, M. Um Pesquisador em Computação em Busca de um Modo de Fazer Pensar Pesquisas em Informática na Educação. **Revista Brasileira de Informática na Educação**, [s. l.], v. 26, n. 1, p. 51-69, 2018.

PRESS, D. **Systems engineering fundamentals**. Virginia: Defense Systems Management College Press, 2001.

RAZALI, N., *et al.* Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. **Journal of statistical modeling and analytics**, [s. l.], v. 2, n. 1, p. 21-33, 2011.

ROSS, D. Structured analysis (SA): A language for communicating ideas. **IEEE Transactions on software engineering**, [s. l.], n. 1, p. 16-34, 1977.

RYAN, T. **Estatística moderna para engenharia**. Rio de Janeiro: Elsevier Brasil, 2009.

SARAMAGO, S., *et al.* Planejamento de experimentos e otimização de sistemas mistos. **Revista Eletrônica da Faculdade de Matemática**, Uberlândia, n. 11, p. 231-246, 2008.

SELEME, R.; STADLER, H. **Controle da qualidade: ferramentas essenciais**. Paraná: Editora Ibplex, 2008.

SIEGMUND, J.; SIEGMUND, N.; APEL, S. Views on internal and external validity in empirical software engineering. *In: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING*, v. 1., 37., 2015, [s. l.]. **Proceedings [...]**. [S. l.]: IEEE Press, 2015. p. 9-19.

SINGH, R. *et al.* Prioritizing threats to patient safety in rural primary care. **The Journal of Rural Health**, [s. l.], v. 23, n. 2, p. 173-178, 2007.

SPSS. **IBM SPSS statistics for Windows, version 20.0**. v. 440. New York: IBM Corp, 2011.

SUTTON, A.; HIGGINS, J. Recent developments in meta-analysis. **Statistics in medicine**, [s. l.], v. 27, n. 5, p. 625-650, 2008.

TEIXEIRA, E.; FONSECA, L.; SOARES, S. Threats to validity in controlled experiments in software engineering: what the experts say and why this is relevant. *In: BRAZILIAN SYMPOSIUM ON SOFTWARE ENGINEERING*, 32, 2018, Brasil. **Proceedings[...]**. Brasil: ACM, 2018. p. 52-61.

TRAVASSOS, G.; BARROS, M. Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. *In: WORKSHOP ON EMPIRICAL SOFTWARE ENGINEERING THE FUTURE OF EMPIRICAL STUDIES IN SOFTWARE ENGINEERING*, 2., 2003, [s. l.]. **Proceedings [...]**. [S. l.: s. n.]: 2003. p. 117-130.

VAN SOLINGEN, R., *et al.* Goal question metric approach. *In: MARCINIAK, J. J. Encyclopedia of software engineering*. [S. l.: s. n.]: 2002.

WAZLAWICK, R. **Metodologia de pesquisa para ciência da computação**. Rio de Janeiro: Elsevier Brasil, 2017.

WOHLIN, C. *et al.* **Experimentation in software engineering**. [S. l.]: Springer Science & Business Media, 2012.

WOOLSON, R. Wilcoxon signed-rank test. *In*: D'AGOSTINO, R. B.; SULLIVAN, L.; MASSARO, J. **Wiley encyclopedia of clinical trials**. New Jersey: John Wiley & Sons, 2007. p. 1-3.

APÊNDICE A – AN EXPERIMENTAL PLANS CHARACTERIZATION CHECKLIST FOR CONTROLLED EXPERIMENTS IN SOFTWARE ENGINEERING

This checklist purpose is to capture information from characteristics of a given controlled experiment in Software Engineering, created based on the experimental completeness plan and review defined by Fonseca (2016). Depending on the answers, an algorithm searches for threats to validity in a database, generating a list of potential threats and their mitigation actions. The checklist is divided into eight categories and forty-three questions as follows.

1. Goals

This category allows researchers to review their experimental plan regarding the study goals, the research questions, the choice of the controlled experiment as the most appropriate research technique to be used, and concerns about ethical aspects.

2. Hypotheses, Variables, and Measurements

This group of items revises the relationship between the hypotheses, variables and measurements and the research goals.

3. Participants

This section involves concerns about the human subjects since recruitment strategy, the process described, remind to collect important information from the participants, concerns about the population which participants are drawn, sample size, until the way in how to deal with them.

4. Experimental Materials and Tasks

This set of items is focused on materials and tasks which should be used in the controlled experiment.

5. Experimental Design

It makes the researchers check if the experiment design chosen is the most appropriate if the treatments are well defined if the randomization is well described, and if an appropriate blinding procedure should be applied to reduce bias.

6. Procedure

This section allows reviewing the procedure section including an adequate description of the controlled experiment context, training, pilot, and timeline of the experiment.

7. Data Collection and Data Analysis

This category includes items regarding data collection and analysis procedures. Also, this section presents concerns about statistical methods.

8. Document

This category is focused on the general writing of the experimental plan regarding its audience suitability, and its facility to read.

Each item in the instrument presents a set of things to consider, in the form of questions that lead experimenters and reviewers

of experiments to think systematically about the important components that should be contained in the experimental plan.

Each item is associated with a response scale: “Yes”, “Partially”, “No”, or “Not Applicable”.

- “Yes” means the experimental plan could not be improved anymore for that item.
- “Partially” means that the experimental plan has described certain things, but it could be improved.
- “No” means the item is not described at all.
- “Not Applicable” means that the item is not applicable for the experimental plan.

CATEGORY 1 – GOALS

1. Will you interact with participants during the experiment execution?

CATEGORY 2 – HYPOTHESES, VARIABLES, AND MEASUREMENTS

2. Will participants have prior knowledge of the experiment's hypotheses?
3. Is there a guide, method, or protocol for data measurement?
4. Do the defined metric(s) **not** represent what they intend to measure?
5. Will only one metric be used to answer the research questions?
6. Is the confidence interval **not** used to draw conclusions?

CATEGORY 3 – PARTICIPANTS

7. Will there be participants with different levels of experience, or skills performing in any tasks during the experiment?

8. Will there be participants with different gender, personality, or any other characteristics?
9. Is the participants native language different from the experimental material language?
10. Will any participant handle more than one treatment?
11. Will the participants receive any reward for participating in the experiment?
12. Had the participants been involved in similar studies in the past?
13. Could there be any lack of motivation for participants to perform the study?
14. Will the participants have the option to leave the experiment at any stage of its execution?
15. Will there be communication between participants during the experiment?
16. Will there be data collection from participants via characterization and/or a demographic questionnaire?
17. Could the participants not represent the target population?
18. Is the selected sample size small?
19. Could events that influence participants' performance occur in the experimental environment?

CATEGORY 4 – EXPERIMENTAL MATERIALS AND TASKS

20. Could the experimental artefacts and procedures (instruments, materials, technologies, tools, tasks, and so on.) not represent those used in the real context?
21. Are the experimental procedures (instruments, materials, technologies, tools, and so on.) different for any group of participants?
22. Is there a requirements document among the experimental artefacts of the experiment?
23. If the experiment context is inspection, will it be necessary to insert defects in the artefacts?
24. Are the tasks that will be performed by the groups distinct?
25. Will the tasks be carried out by researchers?

CATEGORY 5 – EXPERIMENTAL DESIGN

26. Could the treatments not represent the treatments used in the real context?

- 27. Could the use of a treatment influence other treatments used in the experiment?
- 28. Will each group use only one experimental artefact, or task, or treatment?

CATEGORY 6 – PROCEDURE

- 29. Could the environment in which the experiment will be conducted not represent the actual context of the study?
- 30. Will the experiment be performed in different experimental environments?
- 31. There will be breaks between experiments sections?
- 32. Will the training is given to the participants and/or groups be given by different people in different moments?
- 33. Could any of the participants not understand or do not have sufficient knowledge of the experimental material?
- 34. Could the pilot study not be sufficient to point out possible problems with the execution of the experiment?
- 35. Do the experiment activities have a time constraint for their execution?
- 36. Are the events and the presentation order of the experimental material for the participants described?

CATEGORY 7 – DATA COLLECTION AND ANALYSIS PROCEDURE

- 37. Could the data collected be incomplete, accurate, and/or reliable?
- 38. Do you plan to include observing researchers during the experiment?
- 39. Will any set of procedures, protocols, or methods be used for data analysis?
- 40. Will any procedure be used to resolve possible data analysis disagreements?
- 41. Are the procedures and criteria that support the statistical rigour of the study described or referenced?

CATEGORY 8 – DOCUMENT

- 42. If the experiment is a replication, is the experimental package modified from the original study?
- 43. Do you plan to present or indicate where the raw data from your experiment will be available?

APÊNDICE B – ROTEIRO DE TAREFA: CENÁRIO 3 – IDENTIFICAÇÃO E CLASSIFICAÇÃO DE AMEAÇAS À VALIDADE

VALIDEPLAN – VALIDITY-DRIVEN SOFTWARE ENGINEERING EXPERIMENTS PLANNING TOOL © 2018

Em navegador de internet acesse o seguinte endereço: <https://valideplan.cin.ufpe.br>

1. Crie uma conta na ferramenta usando a opção **SIGN UP**. Forneça seu nome e endereço de e-mail e em seguida receberá um e-mail com os detalhes (login/senha) da conta criada.
2. Conforme instruções do e-mail, faça login na ferramenta **ValidEPlan**.
3. Preencha suas informações pessoais na opção Profile que fica no menu esquerdo.
4. Em seguida, acesse a área principal da ferramenta (*Dashboard*), clique no terceiro item (cenário 3) a saber **<Identify and Prioritize Threats>** cujo ícone de identificação é uma ferramenta.
5. Preencha os campos obrigatórios: Nome (**Draft Name**) e Descrição de seu plano experimental (**Experiment Description**) usando como inspiração um plano experimental que deverá estar aberto em um editor de documentos (doc/pdf) conforme instruções anteriores. Clique na opção **SAVE**.
6. Criada a representação de seu Plano Experimental na ferramenta, o usuário será redirecionado automaticamente para a tela de caracterização (**Experiment Characteristics**). Um *checklist* será aberto e deverá ser integralmente respondido de acordo com as informações do seu plano experimental que deverá estar aberto em um editor de documentos (doc/pdf) conforme o item de instruções dessa avaliação.

Observação 1: As opções de resposta do *Checklist* são **Yes** (quando houver concordância da questão apresentada com seu plano experimental), **No** (quando não houver concordância com a questão apresentada) e **N/A** (quando a questão não se aplica ao contexto de seu plano experimental). O usuário pode salvar o seu progresso a qualquer momento.

Observação 2: Se por algum motivo você sair da tela de preenchimento do *checklist*, você poderá acessá-lo na opção **Experimental Drafts** do menu esquerdo da ferramenta. Ao visualizar o nome do seu plano, clique na opção *Threats and control actions* (símbolo de uma engrenagem, da coluna *actions*).

7. Ao concluir o preenchimento do checklist e clicar em **SAVE**, o usuário será questionado se deseja classificar as ameaças à validade sugeridas para o seu experimento. Clique em **YES** e será redirecionado para a tela de classificação de ameaças.

Observação 3: Na tela de classificação de ameaças serão listadas todas as ameaças sugeridas ao seu plano experimental e indicar que você precisa classificá-las em uma escala de 1 a 3, considerando a intensidade dos seguintes critérios: **Impact** (impacto que a ameaça poderá causar aos resultados do experimento), **Urgency** (urgência de resolução da ameaça) e

Trend (tendência de evolução da situação de risco identificada). O usuário pode salvar o seu progresso a qualquer momento.

8. Após o entendimento das informações anteriores, clique em **CLASSIFY THREATS** e realize a classificação de cada ameaça à validade sugerida ao seu plano experimental.

Observação 4: Informações sobre o entendimento da ameaça apresentada ou dos critérios a serem classificados poderão ser obtidas através do sinal de interrogação em cada um dos itens.

9. Após classificar **TODAS** as ameaças e clicar em **SAVE**, o usuário será questionado se deseja definir e classificar as ações de controle sugeridas para mitigar as ameaças do seu plano experimental. Clique em **YES** e será redirecionado para a tela de definição e classificação de ações de controle (**Control Actions**).
10. Após ler as instruções da tela de classificação de ações de controle, clique em **DEFINE ACTIONS**.

Observação 5: A tela de classificação de ações de controle informa que as ameaças classificadas anteriormente poderão ser filtradas/agrupadas de acordo com uma escala de magnitude/prioridade (muito baixa, baixa, moderada, alta e muito alta). Nesta tela, serão exibidas, individualmente, cada ameaça, junto às suas ações de controle. O usuário poderá selecionar as ações de controle desejadas para o seu plano experimental e classificá-las utilizando uma escala de importância que vai de “**Not Important**” até “**Extremely Important**”. As ações classificadas com maior nível de importância serão as que terão maior assertividade ao mitigar as ameaças sugeridas ao seu plano experimental.

11. Após concluir a etapa de definição e classificação das ações de controle, clique em **SAVE**.
12. O usuário será redirecionado para o espaço de trabalho de seu plano experimental (**Workspace**). Neste espaço o usuário deverá visualizar os relatórios (**Reports**) de ameaças gerados para o seu plano experimental. A saber: relatório de ameaças sugeridas, relatório de ameaças classificadas e relatório de ameaças e ações de controle.
13. Após esta etapa, o fluxo é encerrado, podendo o usuário criar um novo experimento ou editar cada uma das etapas anteriores.
14. Ao concluir essa etapa da avaliação experimental da ferramenta **ValidEPlan**, clique no link abaixo e preencha um questionário sobre a avaliação realizada: <https://goo.gl/forms/ERnnp3boxNDODTSJ2>

APÊNDICE C – ROTEIRO DE ENTREVISTA NARRATIVA EPISÓDICA – ESPECIALISTAS EM ENGENHARIA DE SOFTWARE EXPERIMENTAL

Participantes: Especialistas em Engenharia de Software Experimental (Grupo ESEG)

Objetivo: Coletar depoimentos dos participantes de modo a analisar indícios de funcionalidade, aceitação e satisfação da ferramenta ValiDEPlan.

Objeto: ValiDEPlan | Validity-Driven Software Engineering Experiments Plan Tool.

Ferramentas: Guia de Entrevista, Gravador de Áudio e Skype

Aviso aos participantes

- O áudio da entrevista será gravado;
- Sua participação é voluntária;
- Todas as informações obtidas serão sigilosas e seus dados pessoais não serão identificados em nenhuma fase/etapa desta pesquisa.
- Os dados serão guardados em local seguro e a divulgação dos resultados será feita de forma a não identificar os participantes.
- Não há resposta certa ou errada. Responda com sinceridade e espontaneidade as perguntas.
- De preferências, evite responder apenas com “sim” ou “não”. Tente lembrar-se de situações que possam exemplificar/justificar sua resposta.

Referente ao processo proposto, o papel das ameaças à validade e da ferramenta "ValidEPlan"

1. Para você, qual o papel das ameaças à validade relacionado à qualidade dos resultados de experimentos controlados?
2. Ao que você associa quando pensa em validade de um experimento?
3. Para você, a ferramenta "ValidEPlan" promove de alguma forma uma melhor qualidade e/ou validade dos resultados do Plano Experimental que você utilizou na avaliação? Você poderia falar mais sobre quais ameaças a ferramenta o ajudou a identificar?

Referente às atividades de identificação e classificação de ameaças e ações de controle

4. Durante a avaliação do processo proposto para controle de ameaças à validade:
 - a. Qual a sua principal crítica ao checklist utilizado para identificar ameaças à validade? Você poderia dar alguma sugestão para a melhoria desse procedimento?
 - b. Para você, os critérios (impacto, urgência e tendência) utilizados para classificar as ameaças à validade são efetivos para essa atividade? Você poderia dar alguma sugestão para a melhoria desse procedimento?
 - c. Relacionado à forma que utilizamos para organizar as ameaças à validade (escala de magnitude | Muito alto, alta, moderada, baixa, muito baixa), qual a sua opinião? Você poderia dar uma sugestão de melhoria?
5. Durante a utilização da ferramenta, quando ocorreu de fato a identificação de ameaças à validade de seu plano experimental? Você poderia falar mais sobre essa situação?

6. Você poderia relatar sua experiência na ferramenta com a etapa de classificação das ameaças e ações de controle? Fale sobre uma situação que represente isso?
7. De acordo com sua experiência ao utilizar a ferramenta, você acredita que de fato houve a identificação de possíveis ameaças e a indicação de quais delas devem ter prioridade ao serem tratadas? Você poderia falar mais sobre essa situação?
8. Para você, ao realizar os procedimentos para classificar as ações de controle, a ferramenta lhe ajudou a identificar quais ações são mais assertivas para mitigar as ameaças? Você poderia falar mais sobre essa situação?

Referente aos relatórios disponíveis

9. Para você, os relatórios da ferramenta ajudam no entendimento de quais ameaças que foram sugeridas?
10. Você lembra de quais ameaças que causaram maior impacto ao seu plano experimental e quais são as ações de controle mais assertivas?
11. Você consegue indicar uma melhor forma de representar essas informações de saída da ferramenta em formato de relatórios?
12. De acordo com sua experiência, fique a vontade pra falar sugerir e/ou comentar sobre o processo usado para identificar e classificar ameaças e ações de controle.
13. De acordo com sua experiência, fique a vontade pra falar sugerir e/ou comentar sobre sua experiência em usar a ferramenta ValidEPlan.

APÊNDICE D – FORMULÁRIO DE CONSENTIMENTO DO PARTICIPANTE

Título do Estudo

Avaliação comparativa entre a ferramenta ValidEPlan e a abordagem *Ad hoc* para planejar, identificar e classificar ameaças à validade em experimentos controlados da área de Engenharia de Software.

Instruções Gerais

É importante que leia cuidadosamente as informações contidas neste formulário.

Este formulário de consentimento fornece-lhe todas as informações sobre o estudo, tais como objetivo, procedimento, coleta de dados, privacidade, custos, riscos e informações adicionais. Assim que tiver compreendido o estudo, será solicitado que assine e date este formulário.

Caso necessite de maiores esclarecimentos sobre algum item mencionado aqui, ou precise de informações que não foram incluídas, por favor, pergunte aos experimentadores.

Antes de ser informado a respeito do estudo, é importante que tome conhecimento do seguinte:

1. A sua participação se dá em virtude deste estudo/experimento ser requisito parcial da avaliação da disciplina Engenharia de Software Experimental – IN1080;
2. Você poderá solicitar que seus dados não sejam utilizados para análise, a qualquer momento, por qualquer razão, sendo que todos os dados coletados até o referido momento serão descartados.

É preciso entender a natureza da sua participação e dar o seu consentimento por escrito. Sua assinatura indicará que você entendeu todas as informações referentes a sua participação e que concorda em participar.

Objetivo do Estudo

Analisar o uso na prática da ferramenta de planejamento experimental e controle de ameaças à validade ValidEPlan. Com o propósito de avaliar seu impacto no processo de aprendizagem relacionado ao planejamento de um experimento controlado da área de Engenharia de Software.

Pesquisadores

Eudis Oliveira Teixeira é estudante de doutorado do centro de informática (CIN) da Universidade Federal de Pernambuco (UFPE), e este estudo faz parte da sua pesquisa de doutoramento. Seu orientador e co-orientadora são respectivamente os professores Sergio Castelo Branco Soares e Liliane Sheyla da Silva Fonseca.

Procedimentos

Os participantes serão submetidos à dinâmica da disciplina Engenharia de Software Experimental (IN1080), onde serão estudados e discutidos os seguintes materiais: (i) Capítulos 5, 6, 7 e 9 do Livro “Basics of software engineering experimentation” e; (ii) Capítulo 8 do livro “Experimentation in software engineering”. Os participantes também serão subordinados a um treinamento da ferramenta ValidEPlan. A dinâmica da disciplina assim como o treinamento, serão utilizados para nivelamento sobre o conhecimento de Engenharia de Software Experimental, planejamento de experimentos controlados da área de ES e sobre o uso da ferramenta proposta.

Coleta de Dados

Os dados serão coletados através de questionários específicos e a entrega dos documentos de planejamento experimental.

Caráter confidencial dos registros

As informações obtidas a partir de sua participação neste estudo serão mantidas estritamente confidenciais. Qualquer material será referenciado somente por um identificador.

Meramente para registro do trabalho, você deverá fornecer seu nome e assinar este formulário de consentimento. No entanto, todos os resultados apresentados na tese ou publicações científicas serão anônimos, seu nome não fará parte dos dados.

Riscos e/ou Desconfortos

Não há possibilidade de riscos ou desconfortos associados com a sua participação no estudo.

Custos

Você não terá nenhum gasto com a sua participação no estudo.

Declaração de Consentimento

Declaro que tive tempo suficiente para ler e entender as informações contidas neste formulário de consentimento antes de assiná-lo. Os objetivos e o procedimento foram explicados, bem como o que será requerido a mim como participante. Também recebi respostas para todas as minhas dúvidas.

Compreendo que apesar da minha participação no estudo ser parte da avaliação da disciplina Engenharia de Software Experimental, sou livre para solicitar que meus dados não sejam utilizados para análise, sem aplicação de qualquer penalidade. Confirmando também que recebi uma cópia deste formulário de consentimento. Dou meu consentimento para participar deste estudo.

Participante

Data

Atesto que expliquei cuidadosamente a natureza e o objetivo deste estudo. Acredito que o participante recebeu todas as informações necessárias, que foram explicadas em uma linguagem adequada e compreensível.

Eudis Teixeira

Data

APÊNDICE E – ROTEIRO DE ENTREVISTA NARRATIVA EPISÓDICA COM PESQUISADORES INICIANTE

Objetivo: Coletar depoimentos dos participantes de modo a analisar indícios de funcionalidade, aceitação e satisfação da ferramenta ValiDEPlan.

Ferramentas: Guia de Entrevista, Gravador de Áudio e/ou Skype

Aviso aos participantes

- O áudio da entrevista será gravado;
- Todas as informações obtidas serão sigilosas e seus dados pessoais não serão identificados em nenhuma fase/etapa desta pesquisa.
- Os dados serão guardados em local seguro e a divulgação dos resultados será feita de forma a não identificar os participantes.
- Não há resposta certa ou errada. Responda com sinceridade e espontaneidade as perguntas.
- De preferência, evite responder apenas com “sim” ou “não”. Tente lembrar-se de situações que possam exemplificar/justificar sua resposta.

Perguntas

1. Para você, de fato a ferramenta ajudou a planejar um experimento de uma melhor forma do que da maneira tradicional? Você poderia falar mais sobre essa percepção?
2. Para você, a ferramenta "ValidEPlan" ajudou a construir um planejamento experimental com maior qualidade e/ou validade dos resultados? Você poderia falar mais sobre quais ameaças e ações de controle a ferramenta o ajudou a identificar e classificar, diferente das que foram identificadas da maneira tradicional?
3. Para você, a ferramenta lhe ajudou a identificar possíveis ameaças e definir quais ações são mais eficazes para mitigá-las? Você poderia falar mais sobre essa situação?
4. Quais foram suas percepções positivas relacionadas ao uso da ferramenta ValidEPlan?
5. Quais foram suas percepções negativas relacionadas ao uso da ferramenta ValidEPlan?
6. Relacionado ao uso da ferramenta ValidEPlan, quais são suas sugestões/opportunidades e/ou comentários?
7. De acordo com sua participação no experimento, fique à vontade pra falar sugerir e/ou comentar sobre sua experiência em usar a ferramenta ValidEPlan.

**APÊNDICE F – QUESTIONÁRIO DE AVALIAÇÃO DA FERRAMENTA
VALIDEPLAN**

Legenda				
CT (5)	CP (4)	IN (3)	DP (2)	DT (1)
Concorda totalmente	Concorda parcialmente	Indiferente	Discorda parcialmente	Discorda totalmente

	Questões	CT	CP	IN	DP	DT
Percepção da Aprendizagem	A ferramenta ajudou a planejar um experimento de uma melhor forma do que da maneira tradicional.					
	A ferramenta ajudou a identificar ameaças que não foram identificadas de forma manual.					
	O uso da ferramenta contribuiu para minha aprendizagem sobre planejamento de experimentos.					
	O uso da ferramenta contribuiu para minha aprendizagem sobre identificação e controle de ameaças à validade durante o planejamento experimental.					
	O uso da ferramenta melhorou minha compreensão dos conceitos teóricos que foram abordados sobre planejamento de experimentos.					

	O uso da ferramenta melhorou minha compreensão dos conceitos teóricos que foram abordados sobre identificação e controle de ameaças à validade.					
	O uso da ferramenta ajudou a relacionar os conceitos estudados em sala de aula com a prática de experimentação.					
	As habilidades adquiridas com o uso da ferramenta foram importantes para minha aprendizagem sobre planejamento de experimentos e identificação e controle de ameaças à validade.					
Percepção de Facilidade	Foi fácil aprender a utilizar a ferramenta.					
	Foi fácil usar a ferramenta.					
	A execução de uma tarefa na ferramenta leva a um resultado previsível.					
	Os termos, palavras e conceitos usados na ferramenta são consistentes.					
	Não encontrei problemas para executar as ações que pretendia realizar utilizando a ferramenta.					
	As informações explicativas contidas na ferramenta contribuíram para manuseá-la.					

Percepção de Satisfação	Em geral, estou satisfeito com a ferramenta.				
	A ferramenta foi relevante e útil para a tarefa de planejamento de experimentos e identificação e controle de ameaças à validade.				
	O uso da ferramenta proporcionou maior motivação em aprender sobre planejamento de experimentos e identificação e controle de ameaças à validade.				
	Eu indicaria o uso da ferramenta para outros pesquisadores.				
	O uso da ferramenta melhora a comunicação com outros colaboradores e/ou experimentadores.				
	Eu usaria a ferramenta novamente para o planejamento de experimentos.				
	Eu usaria a ferramenta novamente para identificação e controle de ameaças à validade.				

ANEXO A – DESCRIÇÃO DA UNIDADE EXPERIMENTAL – UE1

Descrição da primeira Unidade Experimental (UE1) apresentado aos alunos que participarão do experimento: Realizar o planejamento de um experimento controlado para verificar se a escolha de uma determinada ferramenta lúdica para o ensino de programação (variável independente) afeta ou não o esforço e a efetividade da aprendizagem de alunos de uma disciplina de lógica de programação (variável dependente). O experimento a ser planejado é avaliar: (i) Se o tempo de execução de atividades práticas de lógica de programação, utilizando a ferramenta Alice é menor que utilizando a ferramenta Scratch e; (ii) Se a quantidade de acertos em atividades práticas de lógica de programação, utilizando a ferramenta Alice é menor que utilizando a ferramenta Scratch.

PLANO EXPERIMENTAL – UE1

Introdução

A partir da análise de trabalhos da área de informática na educação, encontramos a indicação de dificuldades pertinentes à aprendizagem de lógica de programação, relacionadas a leitura e interpretação dos enunciados, capacidade de abstração das informações (e.g. Piva Jr. e Freitas 2010) e dificuldades com a lógica matemática, sintaxe e estruturas das linguagens de programação.

Este plano tem a finalidade de avaliar na prática se a escolha de uma determinada ferramenta lúdica para o ensino de programação afeta ou não o esforço e a efetividade da aprendizagem de alunos de uma disciplina de lógica de programação. Dessa forma, foi planejado um experimento controlado, de acordo com o modelo proposto por Wohlin (2012) e Juristo (2013). Neste sentido, o experimento consiste em quantificar e analisar o tempo de execução e quantidade de acertos de atividades de lógica de programação, de duas maneiras diferentes: Uma com o suporte da ferramenta Alice e a outra com o apoio da ferramenta Scratch.

Q1. Definição

O objetivo deste plano experimental foi estruturado segundo a abordagem GQM (Goal, Question, and Metric), Solingen (2002), descrito a seguir:

Analisar as ferramentas lúdicas para o ensino de programação Alice e Scratch. **Com o propósito de** comparar as duas ferramentas. **Com respeito ao** tempo de execução e quantidade de acertos de atividades de lógica de programação. **Sob o ponto de vista** dos alunos de computação. **No contexto do** ensino e aprendizagem de uma disciplina de lógica de programação.

Q2. Questão de Pesquisa

O experimento visa responder à seguinte pergunta de pesquisa:

- O uso da ferramenta Alice é melhor do que a ferramenta Scratch, relacionado ao esforço e a efetividade da aprendizagem de alunos quando se realiza atividades de lógica de programação?

Para responder a essa questão, será medido o tempo de execução e a quantidade de acertos em atividades práticas de lógica de programação.

Q3. Preocupações Éticas

É importante relatar que esta pesquisa terá o cuidado com as questões éticas, garantindo os direitos dos participantes. As informações estarão num Termo de Consentimento, que tratará sobre participação no estudo, objetivos do estudo, pesquisadores, procedimentos, coleta de dados, sigilo e privacidade dos registros, riscos e/ou benefícios, autonomia e ressarcimento e indenização. A participação no experimento é espontânea e os participantes terão acesso a todas as informações sobre o estudo e poderão solicitar sua desistência a qualquer momento e/ou que seus dados não sejam utilizados para análise.

Q4. Hipóteses

Na definição do nosso objetivo, expressamos que gostaríamos de comparar as ferramentas Alice e Scratch quanto ao tempo de execução e quantidade de acertos de atividades de lógica de programação.

Considere:

- QA-A como sendo a quantidade de acertos de atividades de lógica de programação utilizando a ferramenta Alice.
- QA-S como sendo a quantidade de acertos de atividades de lógica de programação utilizando a ferramenta Scratch.
- TA e TS como sendo o tempo médio para execução de atividades corretas de lógica de programação utilizando a ferramenta Alice e Scratch, respectivamente.

Hipóteses para a compreensão de respostas corretas:

$$H_0: QA-A = QA-S$$

Não existe diferença na proporção de respostas corretas utilizando as ferramentas Alice e Scratch.

$$H_1: QA-A \neq QA-S$$

Existe diferença na proporção de respostas corretas utilizando as ferramentas Alice e Scratch.

Hipóteses para o tempo de compreensão de respostas corretas:

$$H_0: TA = TS$$

Não existe diferença no tempo de compreensão de respostas corretas utilizando as ferramentas Alice e Scratch.

$$H_1: TA \neq TS$$

Existe diferença no tempo de compreensão de respostas corretas utilizando as ferramentas Alice e Scratch.

Q5. Variáveis

Para examinar as hipóteses apresentadas acima serão utilizadas as seguintes variáveis dependentes e suas respectivas métricas:

Percentual de respostas corretas (corretude): Esta variável representa o número de respostas corretas em relação ao número total de respostas.

Tempo gasto para compreensão (eficácia): Esta variável representa o tempo gasto para compreensão e execução das atividades.

Q6. Participantes

Serão convidados a participar do experimento todos os alunos da disciplina de Lógica de Programação do curso de graduação em Ciência e Engenharia da Computação do Centro de Informática (CIn) da Universidade Federal do Pernambuco (UFPE). O

experimento será realizado presencialmente no laboratório de informática do curso. Dessa forma a amostra será não probabilística, ou seja, por conveniência.

De forma geral, o público da disciplina deverá possuir conhecimentos básicos em lógica de programação. O conhecimento dos alunos é, em geral, desenvolvido como parte da própria disciplina, onde eles aprendem os fundamentos da área. Dessa forma, o experimento deverá ser executado na segunda metade da disciplina. Os participantes terão a liberdade de solicitar que seus dados coletados não sejam utilizados para análise, sem nenhum prejuízo na disciplina, porém, a participação é obrigatória. O perfil dos participantes será identificado através de um questionário demográfico antes da realização do experimento.

Q7. Materiais Experimentais

Aqui serão descritos os principais objetos, orientações e instrumentos de medida que serão utilizados para a execução do experimento. A realização do experimento se dará em um laboratório de computação e cada participante utilizará um computador (**objeto**), que estará com as ferramentas Alice e Scratch (**tratamentos**) devidamente instaladas. Como **orientação** cada participante utilizará um roteiro com as atividades e ordem em que deverão ser realizadas e os slides utilizados durante o treinamento. Como instrumento de medida será aplicado um questionário (planilha eletrônica) a cada participante com o objetivo de coletar as respostas das atividades práticas de lógica de programação e que mede automaticamente o tempo que o usuário usou para solucionar as atividades em milissegundos. Cada participante teve um período de 30 minutos para ler as especificações antes de iniciar a solucionar as atividades propostas pelas unidades experimentais.

Serão definidas duas unidades experimentais que receberão a aplicação dos tratamentos. As unidades consistirão em atividades práticas de lógica de programação, envolvendo os seguintes conteúdos: Noções de Lógica, algoritmo, estrutura de decisão, estrutura de repetição, constantes e variáveis.

Com o objetivo de obter parâmetros de comparação das variáveis dependentes coletadas, será criado um gabarito das unidades experimentais que são as atividades práticas de lógica de programação que receberão a aplicação dos tratamentos. Esse gabarito será validado pelo professor da disciplina.

As seguintes tarefas serão realizadas pelos participantes:

- Desenvolver as atividades práticas de lógica de programação, usando a ferramenta Alice e Scratch.
- Preencher o questionário de coleta de dados com as respostas das atividades práticas de lógica de programação.

Todos os materiais necessários à execução do experimento serão disponibilizados aos participantes no momento da execução do experimento, conforme cronograma estabelecido. Após a execução do experimento serão aplicados questionários de opinião para avaliar a avaliação experimental, bem como, as percepções positivas e negativas dos participantes sobre a facilidade de uso das ferramentas.

Q8. Treinamento

O treinamento será realizado em três seções distintas, com duas horas cada parte do treinamento, ministrado pelo experimentador. O primeiro momento será a explanação básica sobre os conceitos sobre os conteúdos de lógica de programação que serão usados para definição das unidades experimentais, necessários para contextualizar e balizar os participantes quanto ao experimento. No segundo momento será realizado

o treinamento dos participantes em relação ao uso das ferramentas Alice e Scratch. Serão apresentadas as funcionalidades das ferramentas.

O terceiro e último momento do treinamento será realizado a partir de uma atividade prática de lógica de programação, criada especificamente para este momento, utilizando ambas as ferramentas. Com o objetivo de evitar influência do treinamento na execução do experimento, este será realizado em duas oportunidades, a primeira logo após o sorteio da primeira unidade experimental a ser utilizada e a segunda antes do início do trabalho com a segunda unidade experimental.

Nesta etapa do experimento, os pesquisadores poderão esclarecer possíveis dúvidas dos participantes, a fim de ser possível realizar o experimento de forma adequada. Servirá também para alinhar o entendimento dos conceitos envolvidos no experimento e permitir a realização de ajustes nos materiais que serão adotados no estudo. Além de fornecer informações sobre a média de tempo necessário para a execução do experimento.

Q9. Design Experimental

Em virtude das diferentes experiências e habilidades dos participantes, assim como da influência das diferenças de complexidade das unidades experimentais que serão usadas, o design experimental proposto é quadrado latino 2x2, ilustrado pela figura a seguir:

	Und Exp (UE1)	Und Exp (UE2)
Grupo P1	Tratamento A	Tratamento S
Grupo P2	Tratamento S	Tratamento A

Os participantes serão divididos em dois grupos (P1 e P2), ambos aplicarão os tratamentos (A e S), utilizando as unidades experimentais (UE1 e UE2). De forma randômica, serão realizados sorteios para a aleatorização das posições dos participantes no quadrado latino, ordem de aplicação dos tratamentos e ordem de utilização das unidades experimentais.

Q10. Procedimentos

Os procedimentos para execução do experimento se darão a partir das seguintes atividades: a) Aprendizagem dos conceitos de lógica de programação pela dinâmica natural da disciplina e; b) treinamento dos participantes em relação ao uso das ferramentas Alice e Scratch;

O experimento deverá ser presencial, realizado no laboratório de informática do curso onde todos os computadores terão a mesma configuração, acesso à internet e as ferramentas Alice e Scratch devidamente instaladas. Será disponibilizado para cada participante um roteiro de execução do experimento, com a ordem de aplicação de cada material experimental, as especificações das unidades experimentais e a planilha de coleta de dados. A execução do experimento terá duração estimada de três horas para cada unidade experimental, em dias alternados. Os dois grupos receberão os materiais ao mesmo tempo.

Como se trata de um modelo de referência, as duas unidades experimentais que receberão a aplicação dos tratamentos deverão ser definidas pelos experimentadores em nível de detalhes suficientes, pois dependendo de quais

atividades/problemas/algoritmos os participantes terão que resolver, podem haver ameaças ou viés nessas escolhas. Por exemplo, pode ser que uma das plataformas (tratamentos) seja mais propícia um determinado tipo de atividade/problemas/algoritmo e isso possa favorecer-la. Conforme definido anteriormente, consistirão em atividades práticas de lógica de programação, envolvendo conteúdos de: Noções de Lógica, algoritmo, estrutura de decisão, estrutura de repetição, constantes e variáveis.

Q11. Coleta e Análise de Dados

O instrumento de coleta de dados será uma planilha eletrônica com o objetivo de obter as respostas das atividades práticas de lógica de programação e medir o tempo de execução das atividades em milissegundos.

Serão utilizados procedimentos estatísticos exploratórios, testes de normalidade e de hipóteses. Para isso, será utilizada a ferramenta “R” para execução das análises estatísticas, e uma significância estatística de $\alpha = 0,05$. A ferramenta “R” é um ambiente de desenvolvimento integrado, para cálculos estatísticos e geração de gráficos.

A análise exploratória será inicialmente realizada através da organização dos dados, de forma que facilite a interpretação e análise das informações, em seguida, a representação visual dos dados amostrais será feita através de técnicas gráficas como o histograma e o *box-plot*, recursos que permitem investigar a simetria e estatísticas dos dados como média, moda e mediana.

Após a verificação da normalidade e/ou homogeneidade das variâncias coletadas, através dos testes de *Shapiro-Wilk* e *Lilliefors*, serão definidos se usaremos testes estatísticos paramétricos a exemplo do *Paired t-test*, ou testes estatísticos não paramétricos a exemplo do teste de *Wilcoxon*. Ao reportar os resultados do experimento em questão, iremos indicar onde os dados brutos serão disponibilizados, o que facilitará a análise por outros pesquisadores e/ou revisores e conseqüentemente possíveis replicações do estudo.

Q12. Ameaças à Validade e Ações de Controle

Para este plano experimental, o conjunto de possíveis ameaças à validade estão organizadas por nome da ameaça, descrição, tipo de ameaça e suas respectivas ações de controle, conforme tabela abaixo:

Ameaça à validade / Tipo	Descrição da Ameaça	Ações de controle
15. Coleta de dados incompleta ou incorreta (Interna).	Os participantes poderão coletar dados em uma primeira sessão e não nas sessões seguintes. Também pode ocorrer quando o participante não preenche o formulário de coleta de dados da maneira correta.	<ul style="list-style-type: none"> Descartar dados incompletos ou incorretos. Fornecer um exemplo durante de como preencher e coletar os dados.
16. Interação entre tratamentos (Constructo)	Ocorre quando usamos mais de um tratamento. Com isso, o primeiro tratamento pode influenciar o desempenho do participante durante a	<ul style="list-style-type: none"> Usar tratamentos que não influenciam o desempenho do outro.

	aplicação dos tratamentos posteriores.	
17. Ordem de apresentação do material experimental para os participantes (Interna).	Essa ameaça ocorre quando a ordem de apresentação do material experimental pode afetar o desempenho dos participantes.	<ul style="list-style-type: none"> • Para evitar possível influência, cada participante receberá um roteiro e ordem de aplicação de cada material experimental.
18. Os participantes se comportam de maneira diferente quando observados (Constructo).	Os participantes podem ter seu desempenho afetado quando observados por outras pessoas durante o experimento.	<ul style="list-style-type: none"> • Um dos objetivos do estudo treinamento é proporcionar adaptação dos participantes ao ambiente experimental.
19. Falta de motivação aos participantes (Interna).	Diz respeito à falta de motivação ao realizar o experimento por parte dos participantes.	<ul style="list-style-type: none"> • Informar aos participantes que os tratamentos utilizados oferecerão aprendizagem a eles. • O experimento é parte obrigatória da disciplina de lógica de programação.
20. Treinamento Insuficiente (Interna).	Ocorre quando o treinamento não fornece todas as informações necessárias para os participantes realizarem o experimento, comprometendo-o.	<ul style="list-style-type: none"> • Permitir que os participantes façam perguntas
21. Familiaridade com o material experimental (Interna).	Quando o participante possui pouca (ou nenhuma) familiaridade com o material experimental, a execução do experimento pode ser comprometida.	<ul style="list-style-type: none"> • Fornecer familiaridade aos participantes através do treinamento.
22. Amostra Heterogênea (Conclusão)	Quando por exemplo, podem existir diferenças entre os participantes relacionadas à experiência e habilidades. Tais diferenças podem causar impacto nos resultados do experimento.	<ul style="list-style-type: none"> • Utilizar um desenho experimental que minimize o impacto da amostra heterogênea (Quadrado Latino). • Realizar treinamento para nivelar habilidades dos participantes. • Atribuir os participantes aleatoriamente aos grupos

		<ul style="list-style-type: none"> • Atribuir os tratamentos aleatoriamente aos participantes
23. Efeitos de fadiga podem afetar o desempenho do participante (Interna)	Se o experimento é muito longo, os participantes podem se sentir cansados ou entediados.	<ul style="list-style-type: none"> • Permitir intervalos durante a execução • Realizar as execuções em dias consecutivos
24. Aplicação de teste estatístico inadequado (Conclusão)	Existem fatores que devem ser atendidos para que um teste estatístico seja utilizado, como por exemplo, tamanho da amostra e normalidade dos dados.	<ul style="list-style-type: none"> • Utilizar teste estatístico de estudos anteriores • Utilizar ferramenta de teste estatístico • Citar as referências para os procedimentos estatísticos utilizados • Definir critérios que sustentam o rigor estatístico
25. Diferenças entre os participantes relacionadas ao nível de habilidade (Interna) 26. Diferenças entre os participantes relacionadas ao nível de experiência (Interna)	Ocorre quando há variações de habilidades entre os participantes. Geralmente, os participantes possuem diferentes níveis de experiência. Assim, o desequilíbrio entre os níveis de experiência dos participantes pode ser um risco para o experimento.	<ul style="list-style-type: none"> • Atribuir os participantes aleatoriamente aos grupos • Atribuir os tratamentos aleatoriamente aos participantes • Realizar treinamento para nivelar a habilidade dos participantes
27. Participante pode adquirir conhecimento durante o experimento (Interna)	Os participantes podem adquirir conhecimento utilizando o material experimental e aplicando os tratamentos.	<ul style="list-style-type: none"> • Estabelecer um design experimental que permita balancear o aprendizado • Aplicar o tratamento logo após o treinamento
28. Precisão dos dados coletados (Interna)	Os dados podem não ser precisos se o participante for responsável por coletá-los.	<ul style="list-style-type: none"> • Pedir para que os participantes sejam precisos na coleta dos dados
29. Estabelecer tempo para execução do experimento (Interna)	Com um tempo pré-estabelecido, o participante pode realizar as tarefas muito rápido para poder atender ao tempo fixado, sem a devida atenção que deveria ter.	<ul style="list-style-type: none"> • Estabelecer o mesmo tempo para todos os participantes
30. Participantes podem abandonar o experimento (Interna)	Essa ameaça ocorre quando os participantes abandonam o experimento.	<ul style="list-style-type: none"> • Formar os grupos experimentais no dia do experimento

		<ul style="list-style-type: none"> • Descartar dados incompletos ou incorretos
--	--	---

Relacionamentos entre Ameaças e Ações de Controle

Os possíveis relacionamentos entre as ameaças e ações identificadas são apresentados na tabela abaixo:

Ameaça à validade	Ações de controle	Pode causar
1. Coleta de dados incompleta ou incorreta. 2. Participantes podem abandonar o experimento (Interna)	<ul style="list-style-type: none"> • Descartar dados incompletos ou incorretos 	<ul style="list-style-type: none"> • Diminuição do tamanho da amostra, assim, a validade estatística dos resultados do experimento é baixa (Conclusão).
3. Treinamento Insuficiente	<ul style="list-style-type: none"> • Permitir que os participantes façam perguntas 	<ul style="list-style-type: none"> • Os participantes podem fazer suposições experimentais das hipóteses do estudo. Com isso, as informações dadas aos participantes podem afetar o desempenho (Constructo).
4. Familiaridade com o material experimental. 5. Diferenças entre os participantes relacionadas ao nível de habilidade	<ul style="list-style-type: none"> • Fornecer familiaridade aos participantes através do treinamento. • Realizar treinamento para nivelar a habilidade dos participantes 	<ul style="list-style-type: none"> • Treinamento desbalanceado (Interna) • Treinamento pode ser insuficiente (Interna) • Treinamento conduzido por única pessoa ou por pessoas diferentes (Interna)
6. Efeitos de fadiga podem afetar o desempenho do participante	<ul style="list-style-type: none"> • Permitir intervalos durante a execução 	<ul style="list-style-type: none"> • Fatos do cotidiano podem afetar o desempenho dos participantes durante os intervalos (Interna) • Comunicação entre os participantes durante os intervalos (Interna)

REFERÊNCIAS

JURISTO, N.; MORENO, A. **Basics of software engineering experimentation**. Springer Science & Business Media, 2013.

MALONEY, J. *et al.* The scratch programming language and environment. **ACM Transactions on Computing Education (TOCE)**, v. 10, n. 4, p. 16, 2010.

PIVA JR, D.; FREITAS, R. Estratégias para melhorar os processos de abstração na disciplina de Algoritmos. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**, 2010.

VAN SOLINGEN, R., *et al.* Goal question metric (gqm) approach. **Encyclopedia of software engineering**, 2002.

WOHLIN, C. *et al.* **Experimentation in software engineering**. Springer Science & Business Media, 2012.

ANEXO B – DESCRIÇÃO DA UNIDADE EXPERIMENTAL – UE2

Descrição da 2ª Unidade Experimental (UE2) apresentado aos alunos que participarão do experimento: Realizar o planejamento de um experimento controlado para verificar se a escolha de determinado método de inspeção de software (variável independente) afeta ou não a eficácia da inspeção (variável dependente). O experimento a ser planejado é avaliar se a quantidade de defeitos encontrados em um documento de requisitos de software, utilizando o processo tradicional de inspeção de software Ad-hoc é menor que utilizando o processo de inspeção por Checklist.

PLANO EXPERIMENTAL – UE2

Introdução

Há uma constante preocupação da Engenharia de Software com a satisfação das necessidades dos clientes quanto a assegurar e garantir a qualidade do produto e dos artefatos de software que está sendo desenvolvido. A partir desse contexto, pretende-se realizar uma avaliação experimental, para verificar se a escolha de técnicas de inspeção de software (conhecidas como técnicas de leitura) afeta ou não a eficácia da inspeção, empregadas na detecção de defeitos de software existentes em um documento de requisitos de software.

Existem diversas técnicas de inspeção de software específicas para a minimização das inconsistências de artefatos de software. Neste sentido, o presente planejamento consiste em avaliar se a quantidade de defeitos encontradas em um documento de requisitos de software, utilizando a técnica de leitura (inspeção de software) AD-HOC é menor que utilizando a técnica de leitura baseada (inspeção de software) em CHECKLIST.

Q1. Definição

O objetivo deste plano experimental foi estruturado segundo a abordagem GQM (Goal, Question, and Metric), descrito a seguir:

Analisar as técnicas de leitura e inspeção de software AD-HOC e CHECKLIST. **Com o propósito de** comparar as duas técnicas. **Com respeito à** quantidade de defeitos encontrados num documento de requisitos de software. **Sob o ponto de vista** dos alunos de ciência da computação. **No contexto** de uma disciplina de Engenharia de Software.

Q2. Questão de Pesquisa

O experimento visa responder à seguinte pergunta de pesquisa:

- O uso de uma determinada técnica de inspeção de software aplicada a um documento de requisitos, afeta ou não a eficácia da inspeção?

Para responder a essa questão, será medido a quantidade de defeitos de **omissão, ambiguidade e inconsistência** encontrados num documento de requisito de software. A classificação dos tipos de defeitos, seguiu a taxonomia proposta por Shull (1998).

Q3. Preocupações Éticas

É importante relatar que esta pesquisa terá o cuidado com as questões éticas, garantindo os direitos dos participantes. As informações estarão num Termo de Consentimento, que tratará sobre participação no estudo, objetivos do estudo, pesquisadores, procedimentos, coleta de dados, sigilo e privacidade dos registros, riscos e/ou benefícios, autonomia e ressarcimento e indenização. Em virtude de ser parte avaliativa da disciplina de Engenharia de Software, a participação no experimento é obrigatória, no entanto, os participantes terão acesso a todas as informações sobre o estudo e poderão solicitar que seus dados não sejam utilizados para análise.

Q4. Hipóteses

Na definição do nosso objetivo, expressamos que gostaríamos de comparar as técnicas de leitura e inspeção de software AD-HOC e CHECKLIST quanto aos defeitos de **omissão, ambiguidade e inconsistência** encontrados num documento de requisito de software. Para formular a hipótese formal (estatística), considere:

Quanto a Defeitos de Omissão

- **DO-Ad** como sendo a Quantidade de Defeitos encontrados de Omissão pela técnica *Adhoc*
- **DO-Ch** como sendo a Quantidade de Defeitos encontrados de Omissão pela técnica *Checklist*

Quanto a Defeitos de Ambiguidade

- **DA_Ad** como sendo a Quantidade de Defeitos encontrados de Ambiguidade pela técnica *Adhoc*
- **DA_Ch** como sendo a Quantidade de Defeitos encontrados de Ambiguidade pela técnica *Chcklist*

Quanto a Defeitos de Inconsistências

- **DI_Ad** como sendo a Quantidade de Defeitos encontrados de Inconsistências pela técnica *Adhoc*
- **DI_Ch** como sendo a Quantidade de Defeitos encontrados de Inconsistências pela técnica *Checklist*

Definição das Hipóteses Nulas:

H₀₁: DO_Ad = DO_Ch

H₀₂: DA_Ad = DA_Ch

H₀₃: DI_Ad = DI_Ch

Não existe diferenças estatisticamente significativas relacionadas à quantidade de defeitos de omissão, ambiguidade e inconsistência encontrados utilizando as técnicas Adhoc e Checklist.

Definição das Hipóteses Alternativas:

H₁: DO_Ad <> DO_Ch

H₁: DA_Ad <> DA_Ch

H₁: DI_Ad <> DI_Ch

Existe diferenças estatisticamente significativas relacionadas à quantidade de defeitos de omissão, ambiguidade e inconsistência encontrados utilizando as técnicas Adhoc e Checklist.

Q5. Variáveis

Para examinar as hipóteses apresentadas acima será utilizada a variável dependente corretude. Esta variável representa o percentual de respostas corretas em relação ao número total de respostas.

Q6. Participantes

Serão convidados a participar do experimento todos os alunos da disciplina de Engenharia de Software do programa de Pós-Graduação do Centro de Informática (CIn) da Universidade Federal do Pernambuco (UFPE). O experimento será realizado presencialmente no laboratório de informática do curso.

De forma geral, deve-se garantir que o público da disciplina possua conhecimentos básicos em engenharia de software e inspeção de artefatos de produto de software. O conhecimento dos alunos é, em geral, desenvolvido como parte da própria disciplina, onde eles aprendem os fundamentos da área. Dessa forma, o experimento deverá ser executado na segunda metade da disciplina. A participação no experimento será obrigatória, no entanto, os participantes terão a liberdade solicitar que seus dados não sejam utilizados para análise, sem nenhum prejuízo na disciplina. O perfil dos participantes será identificado através de um questionário demográfico antes da realização do experimento.

Q7. Materiais Experimentais

Aqui serão descritos os principais objetos, orientações e instrumentos de medida que serão utilizados para a execução do experimento. A realização do experimento se dará em um laboratório de computação, com duração máxima de quatro horas e para cada tratamento o participante receberá um documento de requisitos (**objeto**) e uma lista de verificação (**objeto**) que deve ser seguida para os que farão a inspeção por *checklist* (**tratamento**). Como **orientação** cada participante utilizará um roteiro com as atividades e ordem em que deverão ser realizadas e os slides utilizados durante o treinamento.

Como instrumento de coleta de dados será disponibilizado um formulário a cada participante com o objetivo de organizar as respostas relacionadas à quantidade de defeitos de omissão, ambiguidade e inconsistência encontrados utilizando as técnicas Adhoc e Checklist. Cada participante terá um período de 30 minutos para ler as especificações antes de iniciar o experimento.

Serão definidas duas unidades experimentais que receberão a aplicação dos tratamentos. As unidades consistirão de dois fragmentos de documento de requisitos elaborados especificamente para este experimento. O documento deverá ser construído por especialistas da área de engenharia de software e engenharia de requisitos e deverão ser inseridos, seguindo a taxonomia definida por Shull (1998), os seguintes tipos de defeitos: omissão, ambiguidade, inconsistência, fato incorreto, informação estranha. No entanto, serão objetos de avaliação deste experimento apenas a identificação de defeitos de omissão, ambiguidade e inconsistências.

Com o objetivo de obter parâmetros de comparação das variáveis dependentes coletadas, será criado um gabarito das unidades experimentais que será o documento de requisitos que receberá a aplicação dos tratamentos. Esse gabarito será validado pelos especialistas responsáveis pela construção do documento de requisitos.

As seguintes tarefas serão realizadas pelos participantes:

- Identificar defeitos de omissão, ambiguidade e inconsistência em um documento de requisitos, usando as técnicas de leitura e inspeção de software AD-HOC e CHECKLIST.

- Preencher o formulário de coleta de dados de acordo com os tipos de defeitos encontrados em cada fragmento do documento de requisitos disponibilizado.

Todos os materiais necessários à execução do experimento serão disponibilizados aos participantes no momento da execução do experimento, conforme cronograma estabelecido. Após a execução do experimento serão aplicados questionários de opinião para avaliar a avaliação experimental, bem como, as percepções positivas e negativas dos participantes sobre a funcionalidade e facilidade de uso das técnicas de inspeção utilizadas.

Q8. Treinamento

O treinamento será realizado em três seções distintas, com uma hora cada parte do treinamento, ministrado pelo experimentador. O primeiro momento será a explanação básica sobre qualidade, defeitos e inspeção de software, necessários para contextualizar e balizar os participantes quanto ao experimento. No segundo momento será realizado o treinamento dos participantes em relação ao uso das técnicas de leitura *Ad-hoc* e baseada em *Checklist*, com a aplicação de exemplos de inspeção utilizando as técnicas propostas.

O terceiro e último momento do treinamento será prático, ou seja, uma execução a partir de um fragmento de documento de requisitos, criado especificamente para este momento, utilizando ambas técnicas. Com o objetivo de evitar influência do treinamento na execução do experimento, este será realizado em duas oportunidades, a primeira logo após o sorteio da primeira unidade experimental a ser utilizada e a segunda antes do início do trabalho com a segunda unidade experimental.

Nesta etapa do experimento, os pesquisadores poderão esclarecer possíveis dúvidas dos participantes, a fim de ser possível realizar o experimento de forma adequada. Servirá também para alinhar o entendimento dos conceitos envolvidos no experimento e permitir a realização de ajustes nos materiais que serão adotados no estudo. Além de fornecer informações mais específicas sobre a média de tempo necessário para a execução do experimento.

Q9. Design Experimental

Em virtude das diferentes experiências e habilidades dos participantes, assim como da influência das diferenças de complexidade das unidades experimentais que serão usadas, o design experimental proposto é quadrado latino 2x2, ilustrado pela figura a seguir:

	Und Exp (UE1)	Und Exp (UE2)
Grupo P1	Tratamento <i>Adhoc</i>	Tratamento <i>Checklist</i>
Grupo P2	Tratamento <i>Checklist</i>	Tratamento <i>Adhoc</i>

Os participantes serão divididos em dois grupos (P1 e P2), ambos sofrerão ambos os tratamentos (*Adhoc* e *Checklist*), utilizando as unidades experimentais (UE1 e UE2). De forma randômica, serão realizados sorteios para a aleatorização das posições dos participantes no quadrado latino, ordem de aplicação dos tratamentos e ordem de utilização das unidades experimentais.

Q10. Procedimentos

Os procedimentos para execução do experimento se darão a partir das seguintes atividades: a) Aprendizagem dos conceitos qualidade e inspeção de software pela dinâmica natural da disciplina e; b) treinamento dos participantes em relação ao uso das técnicas de leitura *Adhoc* e baseada em *Checklist*.

O experimento será presencial, realizado no laboratório de informática do curso. Será disponibilizado para cada participante os materiais necessários e um roteiro de execução do experimento, com a ordem de aplicação de cada material experimental, as especificações das unidades experimentais e o formulário de coleta de dados. A execução do experimento terá duração estimada de duas horas para cada unidade experimental, em dias alternados. Os dois grupos receberão os materiais ao mesmo tempo.

Q11. Coleta e Análise de Dados

O instrumento de coleta de dados será uma planilha eletrônica onde serão transcritos os dados apresentados pelo formulário disponibilizado aos participantes. Serão utilizados procedimentos estatísticos exploratórios, testes de normalidade e de hipóteses. Para isso, será utilizada a ferramenta “R” para execução das análises estatísticas, e uma significância estatística de $\alpha = 0,05$. A ferramenta “R” é um ambiente de desenvolvimento integrado, para cálculos estatísticos e geração de gráficos.

A análise exploratória será inicialmente realizada através da organização dos dados, de forma que facilite a interpretação e análise das informações, em seguida, a representação visual dos dados amostrais será feita através de técnicas gráficas como o histograma e o *box-plot*, recursos que permitem investigar a simetria e estatísticas dos dados como média, moda e mediana.

Após a verificação da normalidade e/ou homogeneidade das variâncias coletadas, através dos testes de *Shapiro-Wilk* e *Lilliefors*, serão definidos se usaremos testes estatísticos paramétricos a exemplo do *Paired t-test*, ou testes estatísticos não paramétricos a exemplo do teste de *Wilcoxon*. Ao reportar os resultados do experimento em questão, iremos indicar onde os dados brutos serão disponibilizados, o que facilitará a análise por outros pesquisadores e/ou revisores e conseqüentemente possíveis replicações do estudo.

Q12. Ameaças à Validade e Ações de Controle

Para este plano experimental, o conjunto de possíveis ameaças à validade estão organizadas por nome da ameaça, descrição, tipo de ameaça e suas respectivas ações de controle, conforme tabela abaixo:

Ameaça à validade / Tipo	Descrição da Ameaça	Ações de controle
31. Comunicação entre os participantes (Interna).	A disseminação de informação durante o experimento pode afetar o desempenho dos participantes.	<ul style="list-style-type: none"> • Pedir que os participantes não conversem durante o experimento <ul style="list-style-type: none"> • Não permitir o compartilhamento de material durante o experimento • Discutir posteriormente a possibilidade de descartar dados dos participantes que conversarem.

32. Treinamento conduzido por diferentes pessoas (Interna).	Ocorre quando os participantes são treinados por pessoas diferentes. Assim, o treinamento pode proporcionar informações diferentes para cada grupo do experimento.	<ul style="list-style-type: none"> • Estabelecer procedimentos consistentes de treinamento
33. Coleta de dados incompleta ou incorreta (Interna).	Os participantes poderão coletar dados em uma primeira sessão e não nas sessões seguintes. Também pode ocorrer quando o participante não preenche o formulário de coleta de dados da maneira correta.	<ul style="list-style-type: none"> • Descartar dados incompletos ou incorretos. • Fornecer um exemplo durante o treinamento de como preencher e coletar os dados.
34. Restrição de tempo para execução do experimento (Interna).	Quando um experimento possui tempo limitado para realização das tarefas. Com um tempo pré-estabelecido, o participante pode realizar as tarefas muito rápido para poder atender ao tempo fixado, sem a devida atenção que deveria ter. Portanto, as tarefas podem ser executadas de forma incorreta, causando prejuízo ao experimento.	<ul style="list-style-type: none"> • Conduzir o treinamento para verificar se o tempo estabelecido é suficiente para executar o experimento • Estabelecer o mesmo tempo para todos os participantes.
35. Ordem de apresentação do material experimental para os participantes (Interna).	Essa ameaça ocorre quando a ordem de apresentação do material experimental pode afetar o desempenho dos participantes.	<ul style="list-style-type: none"> • Para evitar possível influência, cada participante receberá um roteiro e ordem de aplicação de cada material experimental.
36. Os participantes se comportam de maneira diferente quando observados (Constructo).	Os participantes podem ter seu desempenho afetado quando observados por outras pessoas durante o experimento.	<ul style="list-style-type: none"> • Proporcionar adaptação dos participantes ao ambiente experimental através do treinamento.
37. Falta de motivação aos participantes (Interna).	Diz respeito à falta de motivação ao realizar o experimento por parte dos participantes.	<ul style="list-style-type: none"> • Informar aos participantes que os tratamentos utilizados oferecerão aprendizagem a eles. • O experimento é parte obrigatória da disciplina de lógica de programação.
38. Familiaridade com o material experimental (Interna).	Quando o participante possui pouca (ou nenhuma) familiaridade com o material experimental, a execução do experimento pode ser comprometida.	<ul style="list-style-type: none"> • Fornecer familiaridade aos participantes através do treinamento.
39. Amostra Heterogênea (Conclusão)	Quando por exemplo, podem existir diferenças entre os participantes relacionadas à experiência e habilidades. Tais	<ul style="list-style-type: none"> • Utilizar um desenho experimental que minimize

	diferenças podem causar impacto nos resultados do experimento.	o impacto (Quadrado Latino). <ul style="list-style-type: none"> Realizar treinamento para nivelar habilidades dos participantes.
40. Aplicação de teste estatístico inadequado (Conclusão).	Existem fatores que devem ser atendidos para que um teste estatístico seja utilizado, como por exemplo, tamanho da amostra e normalidade dos dados.	<ul style="list-style-type: none"> Verificar os requisitos de normalidade e tamanho da amostra dos testes estatísticos definidos. Utilizar ferramenta de teste estatístico. Citar as referências para os procedimentos estatísticos utilizados <ul style="list-style-type: none"> Definir critérios que sustentam o rigor estatístico
41. Efeitos de fadiga podem afetar o desempenho do participante (Interna)	Se o experimento é muito longo, os participantes podem se sentir cansados ou entediados.	<ul style="list-style-type: none"> Permitir intervalos durante a execução Realizar as execuções em dias consecutivos
42. Diferenças entre os participantes relacionadas ao nível de habilidade (Interna) 43. Diferenças entre os participantes relacionadas ao nível de experiência (Interna)	Ocorre quando há variações de habilidades entre os participantes. Geralmente, os participantes possuem diferentes níveis de experiência. Assim, o desequilíbrio entre os níveis de experiência dos participantes pode ser um risco para o experimento.	<ul style="list-style-type: none"> Atribuir os participantes aleatoriamente aos grupos Atribuir os tratamentos aleatoriamente aos participantes Realizar treinamento para nivelar a habilidade dos participantes
44. Replicação do estudo em contextos similares (Externa).	Acontece quando o experimentador não indica onde os dados brutos estão disponíveis, complicando a análise e/ou replicação por outros pesquisadores.	<ul style="list-style-type: none"> Apresenta ou indicar onde os dados brutos estão disponíveis para análise por outros pesquisadores, revisores ou auditores independentes.
45. Fatores técnicos podem causar dificuldade aos participantes (Interna).	Tal ameaça ocorre quando os participantes não possuem habilidades ou capacidades técnicas suficientes para realizar o experimento.	<ul style="list-style-type: none"> Realizar treinamento para nivelar a habilidade dos participantes. Selecionar participantes com experiência no contexto do estudo.
46. Participante pode adquirir conhecimento durante o experimento (Interna)	Os participantes podem adquirir conhecimento utilizando o material experimental e aplicando os tratamentos.	<ul style="list-style-type: none"> Estabelecer um design experimental que permita balancear o aprendizado Aplicar o tratamento logo após o treinamento

47. Participantes podem abandonar o experimento (Interna)	Essa ameaça ocorre quando os participantes abandonam o experimento.	<ul style="list-style-type: none"> • A atividade experimental é parte obrigatória da avaliação de uma disciplina • Formar os grupos experimentais no dia do experimento • Descartar dados incompletos ou incorretos
---	---	--

Relacionamentos entre Ameaças e Ações de Controle

Os possíveis relacionamentos entre as ameaças e as ações identificadas são apresentados na tabela abaixo:

Ameaça à validade	Ações de controle	Pode causar
7. Comunicação entre os participantes durante o experimento.	<ul style="list-style-type: none"> • Descartar dados dos participantes que conversarem 	<ul style="list-style-type: none"> • Diminuição do tamanho da amostra, assim, a validade estatística dos resultados do experimento é baixa (Conclusão).
8. Coleta de dados incompleta ou incorreta. 9. Participantes podem abandonar o experimento (Interna)	<ul style="list-style-type: none"> • Descartar dados incompletos ou incorretos 	<ul style="list-style-type: none"> • Diminuição do tamanho da amostra, assim, a validade estatística dos resultados do experimento é baixa (Conclusão).
10. Treinamento Insuficiente	<ul style="list-style-type: none"> • Permitir que os participantes façam perguntas 	<ul style="list-style-type: none"> • Os participantes podem fazer suposições experimentais das hipóteses do estudo. Com isso, as informações dadas aos participantes podem afetar o desempenho (Constructo).
11. Familiaridade com o material experimental. 12. Diferenças entre os participantes relacionadas ao nível de habilidade	<ul style="list-style-type: none"> • Fornecer familiaridade aos participantes através do treinamento. • Realizar treinamento para nivelar a habilidade dos participantes 	<ul style="list-style-type: none"> • Treinamento desbalanceado (Interna) • Treinamento pode ser insuficiente (Interna) • Treinamento conduzido por única pessoa ou por pessoas diferentes (Interna)
13. Efeitos de fadiga podem afetar o desempenho do participante	<ul style="list-style-type: none"> • Permitir intervalos durante a execução 	<ul style="list-style-type: none"> • Fatos do cotidiano podem afetar o desempenho dos participantes durante os intervalos (Interna) • Comunicação entre os participantes durante os intervalos (Interna)

REFERÊNCIAS

JURISTO, N.; MORENO, A. **Basics of software engineering experimentation**. Springer Science & Business Media, 2013.

MALONEY, J. *et al.* The scratch programming language and environment. **ACM Transactions on Computing Education (TOCE)**, v. 10, n. 4, p. 16, 2010.

PIVA JR, D.; FREITAS, R. Estratégias para melhorar os processos de abstração na disciplina de Algoritmos. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**, 2010.

SHULL, F.; BASILI, V. **Developing techniques for using software documents: a series of empirical studies**. 195 f. Tese de Doutorado. research directed by Dept. of Computer Science. University of Maryland, College Park, Md, 1998.

VAN SOLINGEN, R., *et al.* Goal question metric (gqm) approach. **Encyclopedia of software engineering**, 2002.

WOHLIN, C. *et al.* **Experimentation in software engineering**. Springer Science & Business Media, 2012.

ANEXO C – BANCO DE DADOS DE AMEAÇAS E AÇÕES DE CONTROLE

ID	Ameaça	Descrição da Ameaça	Ação de Controle	Descrição da Ação de Controle	Pode Gerar						
INT-A01	Diferenças entre os participantes relacionadas ao nível de experiência	Esta ameaça ocorre durante a seleção dos participantes. Geralmente, os participantes possuem diferentes níveis de experiência. Assim, o desequilíbrio entre os níveis de experiência dos participantes pode ser um risco para o experimento. Pereplechikov e Ryan (2011) preocuparam-se em selecionar participantes com nível de experiência similar.	INT-C01	Caracterizar a experiência do participante através de questionário	COT-A09	COT-A10					
			INT-C03	Atribuir os participantes aleatoriamente aos grupos							
			INT-C04	Caracterizar a experiência do participante através de atividade pré-experimento	INT-A18						
			INT-C16	Atribuir os tratamentos aleatoriamente aos participantes							
			INT-C25	Agrupar os participantes de acordo com o nível de experiência							
INT-A02	Familiaridade com o material experimental	Se um participante possui pouca (ou nenhuma) familiaridade com o material experimental, a execução do experimento pode ser comprometida. Os resultados podem ser influenciados não pelo tratamento, mas sim pelo nível de familiaridade do participante com o material experimental. Tonella e Ceccato (2005) mencionam que o nível de familiaridade dos participantes com Java e JDK poderia influenciar os resultados.	INT-C17	Realizar treinamento para fornecer familiaridade aos participantes	INT-A15	INT-A26	INT-A29	INT-A31			
			INT-C30	Utilizar material experimental familiar a todos os participantes							
			INT-C39	Aplicar atividade pré-experimento a fim de nivelar a familiaridade dos participantes	INT-A18						
			INT-C60	Explicar o material experimental							
			INT-C69	Estabelecer longo tempo de utilização do material experimental							
INT-A03	Participante pode adquirir conhecimento durante o experimento	Os participantes podem adquirir conhecimento utilizando o material experimental, aplicando os tratamentos, etc. Briand <i>et al.</i> (2005) afirmam que os participantes podem relembrar características do primeiro tratamento aplicado na primeira execução, no caso de	INT-C09	Utilizar diferentes materiais nas diferentes execuções							
			INT-C10	Estabelecer um design experimental que permita balancear o aprendizado							
			INT-C13	Estabelecer pequeno tempo de execução	INT-A18	INT-A15	INT-A26	INT-A29	INT-A31		
			INT-C19	Estabelecer ordem de uso do material							
			INT-C21	Definir critérios de interação entre o pesquisador e os participantes							

			INT-C67	Não permitir o compartilhamento de material durante o experimento							
			INT-C85	Descartar dados dos participantes que conversarem	INT-A01	INT-A06	INT-A30	INT-A35	EXT-A01	COS-A02	COS-A04
INT-A09	Falta de motivação dos participantes	Esta ameaça diz respeito à falta de motivação dos participantes. Thelin <i>et al.</i> , (2003) mencionam a preocupação em controlar tal ameaça.	INT-C06	Participar do experimento é parte obrigatória de um curso/disciplina							
			INT-C32	Selecionar atividades benéficas aos participantes							
			INT-C44	Recompensar o desempenho dos participantes							
			INT-C45	Selecionar participantes voluntários							
			INT-C46	Selecionar participantes do contexto real do estudo	COS-A05						
			INT-C55	Informar aos participantes que atividades semelhantes serão realizadas no futuro							
			INT-C79	Informar aos participantes que o tratamento é benéfico ao conhecimento							
INT-A10	Problemas de especificação do material experimental	Essa ameaça acontece quando o material experimental está inconsistente, ambíguo, incorreto ou incompleto. Arisholm <i>et al.</i> (2007) mencionam o erro de especificação do material como uma ameaça à validade.	INT-C15	Revisar o material experimental (pessoas diferentes, se possível)							
			INT-C24	Conduzir estudo piloto para avaliar o material experimental							
			INT-C43	Selecionar material experimental da literatura							
			INT-C65	Selecionar especialistas para elaborar o material experimental							
			INT-C71	Selecionar material experimental utilizado em estudos anteriores							
INT-A11	Diferenças no ambiente experimental	Tal ameaça pode ocorrer quando os participantes são atribuídos a diferentes ambientes experimentais, como feito por Knodel <i>et al.</i> (2008). Assim, os fatos que ocorrem nos ambientes podem influenciar de forma diferente os participantes.									
INT-A12	Efeitos de fadiga podem afetar o	Se o experimento é muito longo, os participantes podem se sentir cansados ou	INT-C13	Estabelecer pequeno tempo de execução	INT-A18						

	desempenho do participante	ficar entediados. Tal ameaça é mencionada por Mouchawrab <i>et al.</i> (2011). Assim, a execução do experimento é prejudicada.	INT-C22	Utilizar material experimental diferente em mais de uma execução									
			INT-C28	Permitir intervalos entre as execuções	INT-A13	INT-A14							
			INT-C31	Realizar as execuções em dias consecutivos	INT-A13	INT-A14							
			INT-C45	Selecionar participantes voluntários									
			INT-C52	Estabelecer interação entre o pesquisador e os participantes									
			INT-C64	Elaborar tarefas com pequeno tempo de duração									
			INT-C78	Permitir que os participantes escolham o dia de execução do experimento									
			INT-C82	Conduzir estudo piloto a fim de verificar se o tempo de execução causa fadiga									
INT-A13	Fatos do cotidiano podem afetar o desempenho dos participantes durante os intervalos	Essa ameaça ocorre quando o experimento possui intervalos. Durante as pausas, fatos cotidianos podem acontecer e afetar o desempenho do participante (Berling e Thelin, 2004).	INT-C13	Estabelecer pequeno tempo de execução	INT-A18								
			INT-C26	Conduzir apenas uma execução com um tratamento	COT-A08								
INT-A14	Comunicação entre os participantes durante os intervalos	Ocorre quando os participantes conversam durante os intervalos do experimento, não durante a execução. Mouchawrab <i>et al.</i> (2011) mencionam a difusão de informações e a imitação da aplicação dos tratamentos como ameaças à validade interna.	INT-C23	Permitir acesso ao material apenas durante a execução do experimento									
			INT-C29	Pedir aos participantes que não conversem sobre o experimento durante os intervalos									
INT-A15	Treinamento desbalanceado	O treinamento pode ser desbalanceado: diferenças entre o material de treinamento, tempo do treinamento e explicação podem afetar a compreensão dos participantes (Hochstein <i>et al.</i> , 2008).	INT-C57	Estabelecer procedimentos consistentes de treinamento									
INT-A16	Idioma dos participantes diferente do idioma do material experimental	Se o idioma oficial do participante for diferente do idioma do material experimental, existe um risco para o experimento (Genero <i>et al.</i> , 2007).	INT-C36	Utilizar uma ferramenta de tradução									
			INT-C51	Aplicar uma atividade pré-experimento a fim de verificar o nível de conhecimento do idioma do material experimental	INT-A18								

			INT-C74	Selecionar participantes que estudam o idioma do material experimental								
INT-A17	Seleção dos participantes com base em dados históricos	Quando o pesquisador seleciona participantes com base em dados de atividades passadas (Mouchawrab <i>et al.</i> , 2011). Assim, o pesquisador pode causar viés se selecionar as pessoas com os melhores resultados.										
INT-A18	Aplicação de atividade pré-experimento pode influenciar o desempenho do participante	Quando uma atividade pré experimento influencia o desempenho do participante durante a execução do experimento (Jansen <i>et al.</i> , 2009). Nesse caso, o participante pode obter informações, a partir dos pré-testes, que podem influenciar no desempenho.										
INT-A19	No caso de replicação, alteração do pacote experimental do estudo original	Tal ameaça ocorre quando o pesquisador realiza a replicação de um experimento. Se o pacote experimental do estudo original sofrer alterações, os resultados podem não ser causados pelo tratamento, mas sim pelas mudanças feitas no material. Maldonado <i>et al.</i> (2006) realizaram pequenas mudanças no procedimento experimental e relataram que isso era uma ameaça à validade do estudo.	INT-C47	Conduzir estudo piloto para compreender os aspectos do pacote experimental								
INT-A20	Estabelecer tempo para execução do experimento	Quando um experimento possui tempo limitado para realização das tarefas, a ameaça INT-A20 pode ocorrer (Mouchawrab <i>et al.</i> , 2011). Com um tempo pré-estabelecido, o participante pode realizar as tarefas muito rápidas para poder atender ao tempo fixado, sem a devida atenção que deveria ter. Portanto, as tarefas podem ser executadas de forma incorreta, causando prejuízo ao experimento.	INT-C61	Conduzir estudo piloto para verificar se o tempo estabelecido é suficiente para executar o experimento								
			INT-C63	Estabelecer o mesmo tempo para todos os participantes								

INT-A21	Diferenças entre as tarefas experimentais	Essa ameaça acontece quando as tarefas experimentais que os participantes terão que realizar são diferentes, conforme mencionado por Deligiannis <i>et al.</i> (2004). Assim, as diferenças entre as tarefas podem influenciar o resultado do experimento, não sendo somente a aplicação do tratamento.	INT-C12	Elaborar tarefas as mais equivalentes possíveis											
			INT-C53	Utilizar as mesmas tarefas para todos os participantes											
			INT-C62	Conduzir estudo piloto para verificar a adequação das tarefas											
			INT-C77	Elaborar tarefas baseadas na literatura											
INT-A22	Rivalidade entre os grupos	Isso acontece quando os participantes de um grupo fazem o seu melhor utilizando o tratamento convencional (ou seja, o tratamento que é uma espécie de benchmark), para mostrar que tal tratamento pode ser competitivo (Thelin <i>et al.</i> , 2003). Assim, os resultados podem conter viés, visto que o tratamento convencional não foi aplicado da maneira que costuma ser.	INT-C34	Informar que todos os participantes utilizarão os mesmos tratamentos											
INT-A23	Os participantes podem ter realizado estudos semelhantes anteriormente	De acordo com Canfora <i>et al.</i> (2007), se os participantes realizaram estudos semelhantes anteriormente, existe um risco para o experimento. Assim, a experiência que os participantes adquiriram no estudo passado pode influenciar o desempenho.	INT-C18	Selecionar participantes que não tenham participado de experimentos anteriores											
INT-A24	Ordem de apresentação do material experimental para os participantes	Esta ameaça ocorre quando a ordem de apresentação do material experimental pode afetar o desempenho dos participantes, conforme mencionado por Ras e Rech (2009).	INT-C20	Estabelecer uma ordem de apresentação do material experimental											
INT-A25	Falta de conformidade com o processo a ser seguido	Se os participantes não seguirem o processo corretamente (conforme descrito no planejamento do experimento), os resultados podem ser prejudicados. Em seu experimento, Biffi e Halling (2003) identificaram essa ameaça à validade.	INT-C56	Inserir observadores no ambiente experimental	EXT-A08	COT-A04									
			INT-C59	Estabelecer diretrizes para guiar os participantes											
			INT-C68	Realizar treinamento para fornecer conformidade para a execução do experimento	INT-A15	INT-A26	INT-A29	INT-A31							

			INT-C86	Descartar dados dos participantes que não seguirem o processo	INT-A01	INT-A06	INT-A30	INT-A35	EXT-A01	COS-A02	COS-A04
INT-A26	Treinamento pode ser insuficiente	Tal ameaça acontece quando o treinamento não fornece todas as informações necessárias para que os participantes executem o experimento (Budgen <i>et al.</i> , 2011). Então, a realização do experimento é comprometida.	INT-C48	Permitir que os participantes façam perguntas	COT-A06						
INT-A27	Atribuição de recompensa pela a participação e não pelo desempenho	Estabelecendo recompensa pela participação, as pessoas podem não realizar o experimento de forma séria e correta. Assim, o resultado do experimento pode não representar o real desempenho do participante (Thelin <i>et al.</i> , 2004b).									
INT-A28	Fatores técnicos podem causar dificuldade aos participantes	Tal ameaça ocorre quando os participantes não possuem habilidades técnicas suficientes para realizar o experimento. Ahmed <i>et al.</i> (2005) afirmam que a maior ameaça no seu experimento foi a incapacidade dos participantes de conduzirem o experimento, ou seja, não possuíam habilidade técnica para isso.	INT-C40 INT-C42 INT-C48	Descartar dados dos participantes que apresentarem problemas técnicos Selecionar participantes com experiência no contexto real do estudo Permitir que os participantes façam perguntas	INT-A01 COT-A06	INT-A06	INT-A30	INT-A35	EXT-A01	COS-A02	COS-A04
INT-A29	Treinamento conduzido por pessoas diferentes	ocorre quando os participantes são treinados por pessoas diferentes. Assim, o treinamento pode proporcionar informações diferentes para cada grupo do experimento (Muller, 2004).	INT-C57	Estabelecer procedimentos consistentes de treinamento							
INT-A30	Diferenças entre os participantes relacionadas aos gêneros	Essa ameaça ocorre quando o gênero do participante pode causar viés ao experimento. (Salleh <i>et al.</i> , 2010)									
INT-A31	Treinamento conduzido por	Tal ameaça acontece quando consciente ou inconscientemente o instrutor influencia o	INT-C57	Estabelecer procedimentos consistentes de treinamento							

	apenas uma pessoa	treinamento do participante. Assim, o treinamento para um grupo pode ser melhor do que para outro (Muller, 2005). Por exemplo, o instrutor pode beneficiar a nova tecnologia que está sendo proposta ministrando um treinamento mediano para o grupo que for utilizar a tecnologia convencional.	INT-C84	Realizar treinamento com diferentes instrutores	INT-A15	INT-A26	INT-A29	INT-A31			
INT-A32	Coleta de dados incompleta ou incorreta	Essa ameaça ocorre quando os participantes não finalizam as tarefas experimentais. Por exemplo, os participantes podem coletar dados em uma primeira sessão e não fazê-lo nas seguintes. Pode surgir também quando o participante não preenche o formulário de coleta de dados da maneira correta.	INT-C14	Descartar dados incompletos ou incorretos	INT-A01	INT-A06	INT-A30	INT-A35	EXT-A01	COS-A02	COS-A04
			INT-C80	Fornecer um exemplo de como preencher o formulário de coleta de dados							
INT-A33	Os participantes podem procurar respostas em outras fontes	Acontece quando o participante, por exemplo, realiza o experimento na sua própria casa. Com isso, os participantes podem utilizar outras fontes de informação para executar as tarefas experimentais (Chatterji <i>et al.</i> , 2011).	INT-C76	Não permitir acesso à internet							
INT-A34	Parte dos participantes recebem recompensa e outra parte não	Acontece normalmente quando uma parte dos participantes selecionados é composta de profissionais e outra parte de estudantes. Com isso, foi dado recompensa para estimular a participação dos profissionais, enquanto para os estudantes a participação é obrigatória. (Müller & Höfer, 2007)									
INT-A35	Diferenças entre os participantes relacionadas ao nível de personalidade	Essa ameaça acontece quando a personalidade do participante pode causar viés à execução do experimento, como foi o caso no experimento conduzido por Salleh <i>et al.</i> (2009).	INT-C72	Agrupar os participantes de acordo com o nível de personalidade							
			INT-C73	Caracterizar a personalidade dos participantes através de questionários	COT-A09	COT-A10					
INT-A36	Adequação das tarefas a serem	Quando as tarefas a serem executadas no experimento não são adequadas ao contexto	INT-C62	Conduzir estudo piloto para verificar a adequação das tarefas							

	realizadas no experimento	estudado (Knodel <i>et al.</i> , 2008), a ameaça pode ocorrer. Assim, as tarefas influenciam o resultado do experimento.											
Tipo de Ameaça: Externa													
ID Ameaça	Ameaça	Descrição da Ameaça	ID Ação de Controle	Descrição da Ação de Controle	Pode gerar								
EXT-A01	Representatividade dos participantes	Tal ameaça ocorre quando a amostra dos participantes não representa a população para a qual o pesquisador quer generalizar os resultados. Lucia <i>et al.</i> (2011) identificaram essa ameaça em seu experimento. A amostra pode não representar a população alvo em termos de conhecimento, experiência, habilidade, tamanho, dentre outros.	EXT-C01	Selecionar participantes do contexto real do estudo.	COS-A05								
			EXT-C02	Selecionar participantes com conhecimento similar ao de pessoas do contexto real do estudo.									
			EXT-C03	Selecionar participantes próximas de iniciarem carreira no contexto real do estudo.									
			EXT-C04	Selecionar participantes que possuam experiência no contexto real do estudo.									
			EXT-C13	Realizar treinamento para os participantes.	INT-A15	INT-A26	INT-A29	INT-A31					
			EXT-C21	Selecionar participantes heterogêneos.	COS-A03								
EXT-A02	Representatividade dos artefatos.	Ocorre quando os artefatos experimentais não representam os artefatos do contexto real do estudo, em termos de complexidade e tamanho.	EXT-C06	Selecionar os artefatos do contexto real do estudo.									
			EXT-C08	Selecionar os artefatos mais comumente utilizados no contexto real do estudo, de acordo com pesquisa.									
			EXT-C09	Elaborar artefatos baseados no contexto real do estudo.									
			EXT-C14	Utilizar artefatos da literatura.									
			EXT-C23	Enviar os artefatos para análise de especialistas.									
EXT-A03	Representatividade do domínio dos artefatos	Essa ameaça ocorre quando o domínio dos artefatos não representa o domínio contexto real do experimento.	EXT-C15	Utilizar artefatos de diferentes domínios, no caso de replicação.									
EXT-A04	Representatividade do processo	Essa ameaça surge quando o processo utilizado no experimento não representa o	EXT-C07	Selecionar processo equivalente ao aplicado no contexto real do estudo.									

