ANÁLISE DA QUALIDADE DE EXPERIMENTOS CONTROLADOS NO CONTEXTO DA ENGENHARIA DE SOFTWARE EMPÍRICA

por

EUDIS OLIVEIRA TEIXEIRA

Dissertação de Mestrado

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CIN - CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2014

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

EUDIS OLIVEIRA TEIXEIRA

ANÁLISE DA QUALIDADE DE EXPERIMENTOS CONTROLADOS NO CONTEXTO DA ENGENHARIA DE SOFTWARE EMPÍRICA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação, área de concentração em Engenharia de Software, do Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco.

ORIENTADOR: *Sérgio Castelo Branco Soares.*

RECIFE
2014

Agradecimentos

A Deus por iluminar meu caminho e me dar forças para seguir sempre em frente. À Dayany Vieira Braga Teixeira, minha esposa, pelo carinho, paciência e suporte. À minha pequena Alice por simplesmente existir. Aos meus pais, Mª da Paz Oliveira e Emi Teixeira, mesmo distantes, porém, sempre apoiando minha qualificação profissional.

Ao meu orientador Sérgio Soares, obrigado pela oportunidade, confiança e pelo empenho durante o mestrado. Obrigado à Liliane Fonseca, Alex Borges, Waldemar Ferreira Nt, Adauto Almeida, Emanoel Barreiros e a todos os amigos que conviveram comigo durante o mestrado. Ao Centro de Informática da UFPE (CIn), IF SERTÃO-PE, Univasf, Facape e Facepe, 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."

Geraldo Eustáquio

Resumo estruturado

Contexto – Considerando o aumento do interesse em pesquisas que conduzem Estudos Empíricos (EE), assim como, do aumento no número de pesquisadores e instituições em todo o mundo que investigam processos experimentais em Engenharia de Software (ES), autores têm criticado a falta de qualidade e padronização dos experimentos quanto aos métodos, procedimentos e forma de divulgar os resultados de EE.

Objetivo – Realizar uma análise quantitativa da qualidade dos estudos categorizados como experimentos controlados no contexto da comunidade de Engenharia de Software Empírica (ESE) quanto aos mecanismos de suporte utilizados, replicabilidade e rigor estatístico.

Método – Em virtude de não ter encontrado na literatura revisada um padrão de perguntas para avaliar a qualidade de EE em ES, decidiu-se criar uma escala de qualidade baseada, principalmente, em listas de verificação amplamente utilizadas por pesquisadores da área de ES. O método de procedimento utilizado foi a abordagem GQM - *Goal Question Metric*, composta pelas fases: definição, planejamento, coleta e interpretação.

Resultados – Os estudos que mencionaram o uso de mecanismos de suporte tiveram um índice de qualidade igual a 58,54, numa escala que vai de zero a cem pontos, enquanto aos que não usaram, tiveram um índice igual a 51,32. A classificação da variável replicabilidade e rigor estatístico foi respectivamente 63,42 (Boa) e 70,79 (Muito Boa) e, no geral, os resultados mostraram que o índice de qualidade dos estudos foi 57,15, sendo que entre os locais avaliados, houve diferença estatisticamente significativa apenas quando comparamos o *Journal* (ESEJ) com os outros dois locais pesquisados (EASE e ESEM).

Conclusões – Houve evolução significativa na qualidade dos experimentos que relatou o uso de mecanismos de suporte, o que evidencia a importância da aplicação de metodologias de apoio que permitam planejar, executar e analisar resultados de EE em ES. Porém, o índice de qualidade dos estudos não apresentou diferenças estatísticas

no período avaliado, o que preocupa, pois não foram identificados avanços significativos na qualidade ao longo dos anos. Vale ressaltar que o instrumento de qualidade desenvolvido está estruturado de tal maneira que possa ser evoluído para avaliar a qualidade de outros tipos de estudos, uma vez que apresenta critérios gerais e outros específicos.

Além dos resultados encontrados, espera-se ter contribuído, também, no sentido de prover às outras pesquisas uma compreensão sobre um modelo, processo ou guia que possa dar suporte à avaliação da qualidade de EE e com isso outros pesquisadores conduzam estudos com maior qualidade.

Palavras-chave: Engenharia de Software Empírica, Avaliação da Qualidade, Experimentos Controlados.

Structured Abstract

Context - Considering the increasing interest in leading research of Empirical Studies (ES), as well as the increase in the number of researchers and institutions around the world investigating experimental procedures in Software Engineering (ES), authors have criticized the lack of quality and standardization of experiments on methods, procedures and forms to disclose the results of ES.

Objective - To carry out a quantitative analysis of the quality of the studies categorized as controlled experiments in the context of the community of Empirical Software Engineering (ESE) and the support mechanisms used, replication and statistical rigor.

Method - By virtue of having not found in the literature reviewed a pattern of questions to assess the quality of ES, it was decided to create a quality scale based mainly on checklists widely used by researchers in ES. The method of procedure used was the GQM approach - Goal Question Metric, comprised of the following steps: definition, planning, collection and interpretation.

Results - The studies that mentioned the use of support mechanisms had a quality index equal to 58.54, on a scale from zero to one hundred points, while those who did not use such mechanisms had a rate equal to 51.32. The classification variable replicability and statistical rigor was respectively 63.42 (Good) and 70.79 (Very Good) and overall the results showed that the index of quality of the studies was 57.15, whereas among local reviews, there was statistically a significant difference only when comparing Journal (ESEJ) with the other two sites surveyed (EASE and ESEM).

Conclusions - There was a significant improvement in the quality of the experiments which reported the use of support mechanisms, which highlights the importance of implementing support methodologies that allow to plan, execute and analyze results in ES. However, the quality score of the studies showed no statistical differences in the period studied, which raises concern because there were no significantly identified improvements in quality over the years. It is noteworthy that the quality instrument developed is structured in such a way that it can be evolved to assess the quality of other types of studies, since it presents general and other specific criteria.

Finally, it is hoped, also, that the work has contributed towards providing an understanding to other research on a model, process or guide that can support the quality assessment of ES and with that other researchers may conduct studies with higher quality.

Keywords: Empirical Software Engineering, Quality Assessment, Controlled Experiments.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

-

Sumário

Capítulo

1

1. Introdução

Este capítulo inicia a discussão que será abordada nesta dissertação e relata os motivos, questões, objetivos e métodos que levaram a realizar uma avaliação de qualidade de experimentos no contexto da comunidade de Engenharia de Software Empírica (ESE).

1.1 Motivação

Cada vez mais tem-se observado o aumento do interesse em pesquisas que conduzem estudos empíricos, bem como, na avaliação e uso de novas tecnologias, métodos, técnicas, ferramentas, linguagens, etc. [2].

Em 1986, Basili et al. [8] iniciaram uma ampla discussão sobre a necessidade de inclusão e aplicação dos processos experimentais em Engenharia de Software (ES).

Ao longo dos anos, Kitchenham et al. [36][42] e Wholin et al. [67] têm estudado a importância desses processos para a pesquisa em ES, o que possibilitou o desenvolvimento de metodologias de apoio que permitem planejar, executar e analisar resultados de estudos empíricos em ES.

Alguns métodos empíricos, principalmente os da área de Engenharia de Software Baseado em Evidências (ESBE), a exemplo da Revisão Sistemática da Literatura (RSL) e Estudo de Mapeamento Sistemático (EMS), constroem conhecimento através do agrupamento e avaliação de evidências de estudos primários, no entanto para que os resultados sejam válidos e confiáveis é necessário fazer uma correta avaliação de qualidade dos estudos que se quer incluir, pois, se não for realizada de forma organizada e sistemática, os resultados podem ser comprometidos [40].

Para compor essa pesquisa, realizou-se inicialmente uma revisão da literatura *ad-hoc*, porém, não foi encontrada pesquisas empíricas no contexto da comunidade de Engenharia de Software Empírica (ESE) relacionadas à evolução da qualidade dos experimentos quanto aos mecanismos de suporte utilizados, replicabilidade e rigor estatístico. Assim, acredita-se que seja a hora de investigar essa comunidade no sentido de saber se houve alguma melhora notável da qualidade dos estudos durante todo o seu período de divulgação.

1.2 Definição do Problema

Apesar do aumento no número de pesquisadores e instituições em todo o mundo que investigam processos experimentais em ES [59], autores têm criticado a falta de qualidade dos experimentos, bem como, a falta de padronização quanto aos métodos, procedimentos e forma de divulgar os resultados de estudos empíricos [21][27][29][38][59][66][68].

A literatura tem proposto o uso de mecanismos de suporte, que são metodologias, processos, guidelines, ferramentas, técnicas e boas práticas, que

permitam planejar, executar, analisar e empacotar estudos empíricos em ES [23][32][36][41][64][67]. No entanto, não se sabe se há alguma influencia na qualidade dos estudos que usam tais mecanismos.

1.3 Questões de Pesquisa

De acordo com o problema exposto, esse estudo objetiva responder às seguintes questões de pesquisa:

- **Q1**. Qual é a evolução da qualidade dos estudos em relação aos mecanismos de suporte utilizados?
- **Q2**. Qual é a evolução da qualidade dos estudos em relação à replicabilidade?
- **Q3**. Qual é a evolução da qualidade dos estudos em relação ao rigor estatístico?
- **Q4**. Qual é a evolução da qualidade dos estudos no período avaliado?
- **Q5**. Qual é a evolução da qualidade dos estudos por veículo de divulgação avaliado?

1.4 Objetivo Geral

A partir do contexto de um mapeamento sistemático, essa pesquisa propõe analisar a qualidade e reportagem dos estudos categorizados como experimentos controlados da comunidade de ESE quanto aos mecanismos de suporte utilizados, replicabilidade e rigor estatístico.

1.4.1 Objetivos Específicos

- Realizar uma revisão da literatura em busca de *checklists*, critérios, listas de verificação e escalas utilizadas atualmente para avaliação da qualidade de EE;
- Desenvolver um instrumento (mecanismo) de avaliação baseado em critérios de qualidade que sejam mais apropriados para o contexto desta investigação;
- Avaliar o instrumento de medição de qualidade desenvolvido e verificar a confiabilidade das respostas entre os avaliadores do estudo;
- Analisar a evolução da qualidade quanto aos mecanismos de suporte utilizados, replicabilidade e rigor estatístico dos estudos;
- Propor critérios de avaliação gerais e específicos no sentido de uma possível aplicação em mais de um tipo de estudo.

1.5 Resultados Esperados

Espera-se que os resultados deste trabalho possam, não apenas apresentar uma avaliação de qualidade de experimentos controlados em ES, como também auxiliar pesquisadores a realizar avaliações de qualidade em uma revisão sistemática e servir de apoio à compreensão de um modelo/processo/guia para avaliar a qualidade de estudos empíricos em ES, beneficiando assim a comunidade de ESE.

1.6 Metodologia

Esta pesquisa de mestrado usa um subconjunto dos resultados de um estudo de mapeamento sistemático realizado pelo grupo de pesquisa liderado pelo orientador deste trabalho. O autor desta dissertação foi coautor deste trabalho feito pelo grupo.

O mapeamento examinou 876 artigos publicados nos principais veículos de divulgação científica da comunidade de ESE. A saber: EASE – *International Conference on Evaluation and Assessment in Software Engineering*, ESEM – *International Symposium on Empirical Software Engineering and Measurement* e ESEJ – *Empirical Software Engineering Journal*.

A seguir, apresenta-se a abordagem metodológica em duas etapas: Classificação e Ciclo da Pesquisa.

1.6.1 Classificação da Pesquisa

A pesquisa apresentada nesta dissertação utiliza o método de abordagem dedutivo baseado em dados de natureza quantitativos. O processo de busca dos dados primários beneficiou-se de um Estudo de Mapeamento Sistemático (EMS), do inglês *Systematic Mapping Study*, que é um tipo de Revisão Sistemática da Literatura (RSL) [41]. O principal método de procedimento utilizado seguiu uma abordagem orientada à melhoria da qualidade através da abordagem GQM (*Goal, Question e Metric*) [60].

Quanto ao objeto, esta é uma pesquisa bibliográfica, pois é elaborada a partir de artigos já publicados, e quanto ao objetivo é descritiva [4]. A Tabela 1.1 apresenta o quadro metodológico do trabalho.

Tabela 1.1 - Classificação da Pesquisa

Quadro Metodológico	
Método de Abordagem	Dedutivo
Quanto ao Objeto	Pesquisa Bibliográfica
Quanto ao Objetivo	Pesquisa Descritiva
Métodos de Procedimento	• Estudo de Mapeamento Sistemático • Abordagem GQM – Definição, Planejamento, Coleta dos dados e Interpretação. • Estatístico
Natureza dos Dados	Quantitativa

De acordo com Andrade [4], o método de abordagem dedutivo segue os caminhos das consequências, numa cadeia de raciocínio em conexão descendente, do geral para o particular. Segundo Severino [56], *apud* Zanella [70], pode-se dizer que a dedução é um procedimento lógico, raciocínio pelo qual pode-se tirar de uma ou de várias proposições uma conclusão que delas decorre por força puramente lógica.

Quanto ao objetivo da pesquisa, esta é descritiva, pois os fatos foram sistematicamente observados, registrados, analisados, classificados e interpretados, sem interferência subjetiva do pesquisador [4].

O principal método de procedimento utilizado foi a abordagem GQM - *Goal Question Metric*, que é composta por quatro fases [60][65]:

- **Definição**: Quando os objetivos, questões, métricas e hipóteses são estabelecidos;
- **Planejamento**: Quando o projeto está selecionado, definido, caracterizado e planejado;
- **Coleta**: Quando a coleta dos dados experimentais é realizada, derivando em um conjunto de dados prontos para a análise;
- **Interpretação**: Quando os dados são processados e analisados a respeito dos objetivos anteriormente definidos.

Quanto à natureza dos dados, esta pesquisa é quantitativa, pois se caracteriza principalmente pelo emprego de instrumentos e técnicas estatísticas, tanto na coleta como no tratamento dos dados, traduzindo em números as informações para serem classificadas e analisadas, isto é, com a medição objetiva e quantificação dos resultados [53][70].

1.6.2 Ciclo da Pesquisa

Esta seção apresenta as principais fases que formaram o ciclo desta pesquisa, conforme a abordagem GQM, detalhada pelas etapas de definição, planejamento, coleta de dados e interpretação dos resultados. A Figura 1.1 apresenta o ciclo de pesquisa.

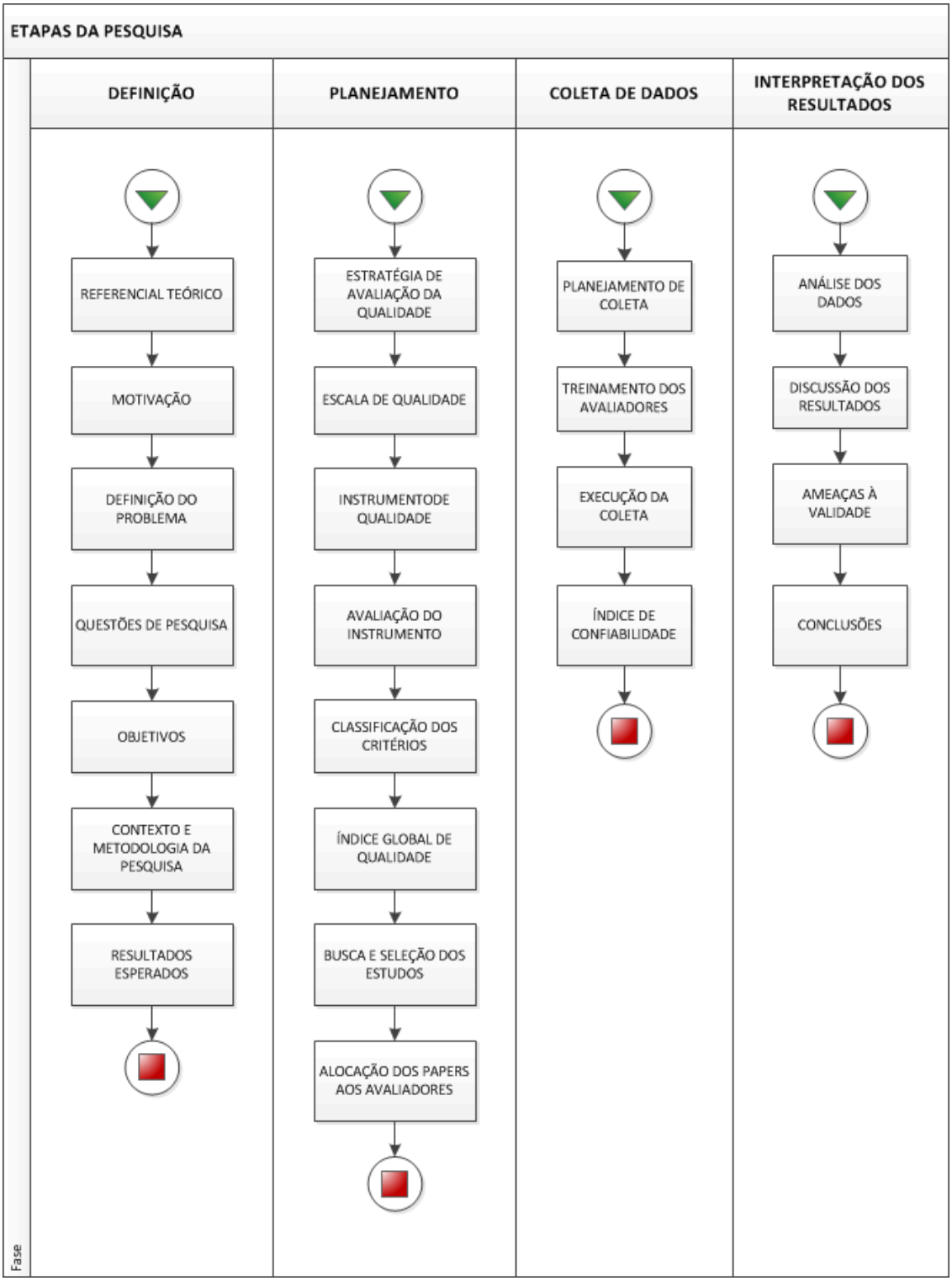


Figura 1.1 - Ciclo da Pesquisa

A etapa de Definição é composta pela construção do referencial teórico que contém a revisão dos principais conceitos que fundamentam a avaliação da qualidade de estudos empíricos, os motivos e as questões de pesquisa que norteiam o trabalho, os objetivos propostos, assim como a descrição dos resultados esperados.

A etapa de Planejamento começa com a discussão da estratégia utilizada para a avaliação da qualidade dos experimentos controlados, a construção do instrumento de avaliação, a definição do estudo piloto usado para verificar a concordância entre os avaliadores, a definição do indicador quantitativo utilizado para avaliar a qualidade dos experimentos e finalizou com o processo de busca e seleção dos estudos e a alocação dos artigos aos participantes da pesquisa.

A fase de Coleta de Dados iniciou com a definição dos procedimentos e instruções para a coleta dos dados, o treinamento dos avaliadores e findou com a execução do estudo e registro das principais informações relativas ao processo de coleta.

Por fim a etapa de Interepretação dos Resultados foi composta pela análise estatística e discussão dos resultados, a indicação das possíveis respostas às questões de pesquisa e encerrou com a apresentação das ameaças à validade e limitações do estudo, conclusões e trabalhos futuros.

1.7 Estrutura do Trabalho

Além deste capítulo introdutório, esta dissertação está organizada de acordo com a seguinte estrutura:

- **Capítulo 2 (Revisão da Literatura)**: apresenta o referencial teórico, que contém a revisão dos principais conceitos que fundamentam a avaliação da qualidade de estudos empíricos;
- **Capítulo 3 (Instrumento de Avaliação)**: descreve a estratégia utilizada para análise da qualidade dos experimentos, bem como, definição, planejamento e construção do instrumento de avaliação de qualidade;
- **Capítulo 4 (Metodologia de Avaliação da Qualidade)**: descreve a estratégia de análise da qualidade, o processo de busca e seleção dos estudos, a forma de alocação dos *papers* aos avaliadores, a estratégia de coleta dos dados, bem como, a execução do estudo;
- **Capítulo 5 (Análise dos Dados e Discussão dos Resultados)**: relata a análise dos dados e descreve os resultados;

- **Capítulo 6 (Conclusão e Trabalhos Futuros)**: neste capítulo final são apresentadas as limitações e ameaças à validade do trabalho, conclusões e propostas de trabalhos futuros.

Capítulo

2

2. Revisão da Literatura

Este capítulo descreve o contexto da pesquisa sobre avaliação da qualidade de Estudos Empíricos (EE) através de uma revisão da literatura conduzida de maneira *adhoc*. Os assuntos abordados oferecem uma reflexão dos conceitos que apoiam a pesquisa, ou seja, Engenharia de Software Empírica (ESE), principais métodos utilizados para estudos primários e secundários, importância e estratégias da avaliação de qualidade de EE.

2.1 Engenharia de Software Empírica

Um método empírico é um conjunto de princípios organizados em torno do qual os dados empíricos são coletados e analisados [23]. Easterbrook et al. [23] enfatizaram que em virtude dos métodos de pesquisa serem adaptados a partir de um número de áreas de pesquisa diferentes, não existe uma terminologia consistente para descrição dos métodos experimentais e não há um consenso sobre a forma de distinção entre eles.

Dessa forma, selecionar um método de estudo para pesquisa empírica não é algo trivial porque os benefícios e desafios na utilização de cada um, ainda, não estão bem descritos e aceitos na literatura. Também, não há método melhor ou pior do que outro, eles simplesmente são melhores em alguns aspectos e piores em outros [17].

A Engenharia de software (ES) é uma área da ciência que se preocupa com os procedimentos de produção de um software (concepção, especificação, desenvolvimento e manutenção) em escala industrial, no intuito de tornar o processo mais sistemático, científico e quantificável, de forma a se aproximar cada vez mais das engenharias tradicionais [61][28]. No entanto, Shull et al. [58] afirmam que pouco se sabe, ainda, sobre como os engenheiros de software executam o seu trabalho.

Nesse contexto, surgiu a Engenharia de Software Empírica (ESE), uma área da ES que busca investigar a realização de estudos experimentais, ou seja, verificar como os engenheiros de software resolvem problemas em seus ambientes reais. Ao longo de quase 30 anos autores tem evidenciado a importância de ESE em ES [7][8][50][63][67].

Segundo Wohlin et al. [67] existem basicamente quatro métodos relevantes para a condução de estudos empíricos na área de ES: científico, de engenharia, empírica e analítica. No entanto, Travassos et. al [65] sugerem que a abordagem mais apropriada para a experimentação na área de ES seja o método experimental.

Travassos et al. [65] dizem que o método experimental sugere um modelo e, em seguida desenvolve um método que pode ser qualitativo e/ou quantitativo, aplica um procedimento empírico (experimento, quase experimento, etc.), mede e analisa, avalia o modelo e repete novamente o processo, ou seja, segue uma abordagem orientada à melhoria evolucionária.

Em resumo, a ESE através de métodos experimentais, está preocupada em como as ferramentas e processos de desenvolvimento de software realmente funcionam, como compreender seus limites, inclusive sociais e cognitivos e como projetar melhorias. A literatura descreve quais os principais métodos empíricos para estudos

primários e secundários da área de ES: *survey research*, estudo de caso, experimento, etnografia, pesquisa-ação, *grounded theory e* revisão sistemática da literatura. Apesar de propormos a avaliação da qualidade de experimentos, os critérios gerais utilizados para essa avaliação poderão em tese avaliar outros métodos empíricos em ES como os apresentados abaixo.

2.1.1 Métodos Empíricos em Engenharia de Software

O objetivo desta seção é fornecer uma orientação para o entendimento dos métodos experimentais utilizados em ES para estudos primários e secundários.

2.1.1.1 Survey Research

Com origem nas áreas de economia e sociologia o *survey research* é um método de pesquisa empírica onde o pesquisador elabora um questionário para obter respostas de um conjunto de pessoas ou de uma população (censo), relacionadas às ações, experiências, comportamentos, opiniões ou perfis das pessoas com relação ao objeto de estudo pesquisado [26].

Segundo Travassos et al. [65], *survey research* é uma estratégia conduzida quando outras técnicas ou ferramentas já tenham sido usadas, cujo objetivo é descrever, explanar e explorar informações preliminares (quantitativas e qualitativas) através de um questionário, além de levantar as variáveis do estudo a serem avaliadas.

Easterbrook et al. [23] relatam que o método *survey research* é mais associado ao uso de questionários, porém, a coleta dos dados também pode ser realizada por meio de entrevistas estruturadas ou dados técnicos. Eles explicitam, também, que uma pré-condição para a realização de um *survey research* é uma questão de pesquisa clara sobre a natureza de uma população alvo (representativa da população), pois é inviável ou desnecessário realizar a consulta com todos os membros de uma população. Informação adicional sobre *survey research* pode ser encontrada em [23][26][65].

2.1.1.2 Estudo de Caso

Nascido das ciências sociais, ainda hoje, há divergências na literatura quanto ao conceito de estudo de caso. Frequentemente o termo é usado simplesmente para designar um exemplo de caso real, porém, como método empírico, autores têm conceituado estudo de caso de forma diferente [23].

Yin [69] explica o conceito de estudo de caso como “uma investigação empírica que investiga um fenômeno contemporâneo dentro de seu contexto de vida real, especialmente quando os limites entre fenômeno e contexto não são claramente evidentes”.

Fuks [26] conceitua-o como um método de estudo que investiga o fenômeno em seu ambiente (contexto) real sem nenhum tipo de controle sobre as variáveis envolvidas.

Flyvbjerg [25] revela que estudos de caso oferecem compreensão de como e porque certos fenômenos ocorrem e pode revelar os mecanismos pelos quais as relações causa-efeito são construídas.

Easterbrook et al. [23] relatam que o estudo de caso utiliza uma variedade de fontes de dados e que, diferente de outros métodos, este utiliza uma amostragem intencional em vez de usar amostragem aleatória, como é o caso de alguns experimentos. Segundo os autores, uma pré-condição para a realização de um estudo de caso é uma questão de pesquisa clara e preocupada como e/ou porque os fenômenos ocorrem. Informação adicional sobre o método estudo de caso pode ser encontrada em [23][25][26][69].

2.1.1.3 Experimento

Com origem nas ciências naturais, como biologia e física, o experimento normalmente é realizado em laboratório e oferece o maior nível de controle entre os diferentes tipos de classificação para métodos experimentais [26]. Travassos et al. [65] dizem que o tipo de experimento mais apropriado para uma pesquisa depende dos objetivos do estudo, propriedades do processo de software usado na experimentação ou dos resultados esperados. As principais características usadas para diferenciá-los estão relacionadas ao controle de medição e execução, custo e nível de replicabilidade do estudo.

Easterbrook et al. [23] relatam que as etapas de um experimento são orientadas através da definição clara de uma hipótese e sua teoria subjacente, onde o pesquisador tem controle sobre as variáveis: fixa algumas e varia outras. A Figura 2.1 exemplifica um processo experimental empírico [26].

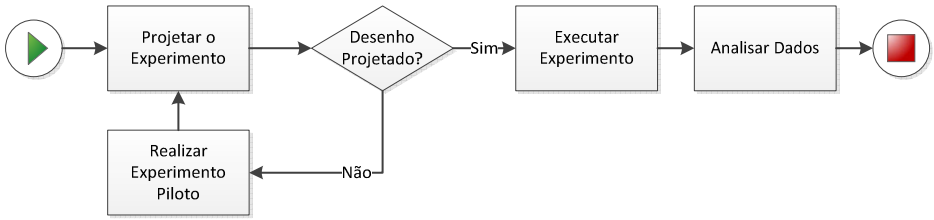


Figura 2.1 - Processo de um Experimento

Neste estudo propõe-se analisar a qualidade dos experimentos publicados nas principais fontes de divulgação da comunidade de Engenharia de Software Empírica (ESE).

2.1.1.4 Etnografia

A história mostra que a etnografia surgiu na antropologia com o objetivo de compreender culturas pouco conhecidas como, por exemplo, as tribos indígenas [26]. A aplicação do método surgiu da necessidade de obter um entendimento aprofundado das práticas de uma determinada população e que não foi possível conseguir a partir de outras técnicas como, por exemplo, entrevistas ou questionários.

Robinson et al. [54] asseguram que o objetivo da etnografia é estudar uma comunidade de pessoas e entender como os membros desta comunidade interagem socialmente, na busca pela resposta de um problema de pesquisa.

Para Easterbrook et al. [23], uma etnografia pode se concentrar tanto em pequenas (por exemplo, um pequeno grupo de desenvolvedores de uma empresa) quanto em grandes comunidades (por exemplo, programadores da linguagem Java em geral). Aborda, também, que uma pré-condição para o método inclui questões de pesquisa que envolve as práticas culturais da comunidade observada.

Em resumo, dado um problema cultural, o objetivo da etnografia é investigá-lo a partir das interações sociais do grupo pesquisado, ressaltando que a aplicação do método etnográfico evita impor qualquer teoria pré-existente relacionada ao problema e se concentra principalmente em investigar as questões de pesquisa a partir do contexto sócio cultural daquela comunidade. Outros detalhes sobre o método etnografia pode ser encontrado em [23][26][54].

2.1.1.5 Pesquisa-Ação

O método pesquisa-ação é mais uma estratégia de pesquisa científica para a investigação empírica em ES. A estratégia pesquisa-ação distingue-se das demais em virtude do pesquisador deixar de ser um observador neutro para atuar, modificar e aprender a partir da ação que realiza durante a execução do estudo [26].

Thiollent [62] define a pesquisa-ação como “um tipo de pesquisa social com base empírica que é concebida e realizada em estreita associação com uma ação ou com a resolução de um problema coletivo e no qual os pesquisadores e os participantes representativos da situação ou do problema estão envolvidos de modo cooperativo ou participativo”.

Davison et al. [16], e Easterbrook et al. [23], afirmam que na pesquisa-ação os pesquisadores tentam estudar (pesquisa) e resolver (ação) o problema ao mesmo tempo, ou seja, intervir (ação) nas situações estudadas (pesquisa) com o objetivo explícito de melhorá-las.

Em síntese, o duplo objetivo da pesquisa-ação reforça sua essência que é pesquisar (teoria) e agir (prática), onde o pesquisador e sujeitos da pesquisa colaboram, buscando sempre avançar na teoria e intervir no problema específico. Porém, a imaturidade é um dos principais desafios deste método empírico, pois é relativamente novo e há uma ampla discussão sobre a metodologia utilizada para aplicá-lo e até mesmo um debate sobre se o método é de fato empírico [23]. Informação adicional sobre o método pesquisa-ação pode ser encontrada em [16][23], [26][62].

2.1.1.6 Grounded Theory

Nascido da sociologia e também conhecido como teoria fundamentada nos dados, o método *grounded theory* preocupa-se com os procedimentos de coleta e análise, numa tentativa de extrair dos dados qualitativos, novos conceitos que possibilitem a criação ou aperfeiçoamento de teorias [14][26].

Segundo Corbin et al. [14], criar uma estrutura integrada que possa ser usada para explicar ou prever fenômenos (teoria), derivada dos dados primários é algo extremamente complexo, porém, este processo torna-se científico e quantificável a partir do método *grounded theory*, cuja principal característica é coletar, organizar e analisar dados de forma sistemática no intuito de elaborar ou estender uma teoria.

Fuks [26] explica que, diferente de outros métodos dos quais parte-se sempre de teorias existentes para o planejamento da pesquisa e coleta dos dados, com *grounded theory,* o processo é inverso. Partimos de um tema ou situação de pesquisa, coletamos e analisamos os dados para fazer surgir uma teoria fundamentada. Ainda segundo Fuks, os dados são coletados a partir de entrevistas e de observação direta, em seguida são comparados no intuito de identificar categorias (variáveis) e após a exaustão da coleta e análise, aí então o pesquisador inicia a comparação com as

teorias correntes na literatura científica. Informação adicional sobre o método *grounded theory* pode ser encontrada em [14][26].

2.1.1.7 Revisão sistemática da literatura

Alguns métodos empíricos preocupam-se em agrupar e avaliar evidências existentes sobre uma determinada tecnologia. Dyba et al. [22] definiram etapas através das quais a ESE agrupa e avalia evidências em ES, dentre elas destacamos duas: a) Pesquisar na literatura por melhores evidências disponíveis para responder às perguntas e b) Avaliar criticamente as evidências, quanto a sua validade, impacto e aplicabilidade. Essas etapas podem ser realizadas através de um método chamado Revisão Sistemática da Literatura (RSL) (do inglês *Systematic Literature Review*).

A RSL constitui uma forma de estudo secundário, pois dependem dos estudos primários para poder agregar evidências e construir conhecimento [38]. Pai et al. [49] conceitua RSL como uma busca abrangente e exaustiva por estudos primários relacionados com uma questão específica de pesquisa, tendo critérios de qualificação claros e reproduzíveis para a seleção dos estudos. A literatura diferencia alguns tipos de RSLs [51], entre eles:

- **RSLs convencionais** – Segundo Kitchenham et al. [41], são estudos secundários que usam uma metodologia bem definida para identificar, analisar e interpretar toda a evidência disponível relacionada com uma questão específica de pesquisa de maneira imparcial e de alguma forma repetível. Petticrew et al. [51] explicam que RSLs convencionais agregam resultados sobre a eficácia de um tratamento, de uma intervenção ou de uma tecnologia e normalmente estão relacionados a questões do tipo: Um determinado tratamento X aplicado a uma população Y é mais eficiente em obter o resultado W no contexto K em comparação com o tratamento Z?
- **Estudos de Mapeamento Sistemático (EMS)** – Segundo Kitchenham et al. [41], estudos de mapeamento sistemático, também categorizado como estudo exploratório, é uma ampla revisão de estudos primários em uma área de tópico específico que visa identificar quais provas estão disponíveis sobre o tema em questão. No mesmo sentido, Arksey et al.[5] explicam que EMS objetivam identificar todas as pesquisas relacionadas a um tópico específico, como por exemplo, responder questões mais amplas (exploratórias) sobre tendências em uma determinada área de

pesquisa. Normalmente estão relacionados a questões do tipo: O que sabemos sobre o Tema X?

Em resumo, RSLs e EMSs são métodos de pesquisa secundários e surgiram como duas importantes ferramentas para acrescentar e construir conhecimento em ES.

Os EMSs possibilitam uma visão mais ampla e exploratória dos estudos primários, tornando-o dependente da realização de mapeamentos para revelar as evidências da pesquisa, enquanto as RSLs convencionais visam identificar, avaliar e interpretar todas as pesquisas disponíveis relevantes para uma determinada questão de pesquisa, focada em um ponto específico da área de estudo [38]. Esta dissertação beneficiou-se de um EMS realizado pelo nosso grupo de pesquisa que examinou 876 artigos, dos quais analisamos 104, definidos como experimentos controlados.

2.1.2 Experimentos Controlados

Um experimento é uma investigação de uma hipótese testável, onde uma ou mais variáveis são manipuladas para medir o seu efeito (resultado) sobre outras variáveis [23].

Abaixo, descrevem-se alguns motivos apontados na literatura que levou esse estudo a avaliar a qualidade de experimentos. [65][23].

- São experiências realizadas normalmente em laboratórios e que oferecem maior nível de controle presumindo um relacionamento de causa (representado pelos tratamentos de variáveis) e efeito (representado pelos resultados);
- A força do experimento encontra-se no controle total sobre o processo e as variáveis, bem como, a possibilidade de replicá-lo;
- A literatura apresenta cinco fases (gerais) bem definidas para o processo de condução de um experimento: Definição de Objetivos (global, medição, questões), Planejamento (hipóteses, instrumentos, contexto, indivíduos, variáveis, analise e validade), Execução (perfil do participante, competências, coleta de dados, resultados), Análise e Apresentação (validação, estatística, verificação de hipóteses) e Empacotamento (pré-requisito para repetição, bibliotecas de experimentação);
- São apropriados para confirmar teorias, confirmar o conhecimento convencional, explorar os relacionamentos, avaliar a predição dos modelos, bem como, validar medidas;

- Apresentam características de aleatoriedade, agrupamento e balanceamento como princípios básicos;
- Normalmente os resultados de um experimento devem ser válidos segundo quatro tipos de validade (interna, externa, de constructo e de conclusão);
- Apresentam 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.

2.1.2.1 Mecanismos de suporte utilizados

Segundo Almeida et al. [2], estudos empíricos vêm ganhando espaço na área de ES e promovendo o desenvolvimento de ferramentas, metodologias e técnicas de suporte. Isto deve-se principalmente à capacidade de gerar evidências sobre a eficiência e eficácia das tecnologias através de experimentos sistemáticos.

Nos últimos anos, diversos estudos têm apresentado ambientes de apoio, *guidelines*, metodologias, *frameworks* e outros mecanismos que permitem planejar, executar, analisar e empacotar estudos empíricos em ES [23][32][36][41][64][67].

Neste sentido, essa pesquisa investiga o nível de qualidade dos estudos que declararam usar mecanismos de apoio ao seu planejamento e execução, pois poderemos verificar se há alguma diferença de qualidade entre esses estudos e os que não usaram mecanismos de suporte. Caso confirmado a influência na qualidade dos estudos que usam mecanismos de apoio, poderemos incentivar o uso e desenvolvimento de mecanismos que apoiem a realização de pesquisas empíricas e, com isso, pesquisadores poderão produzir resultados experimentais de maior qualidade. Os resultados e conclusões sobre a qualidade dos experimentos quanto aos mecanismos de suporte serão descritos em detalhes nos Capítulos 5 e 6.

2.1.2.2 Replicabilidade dos estudos

Um dos cuidados mais importantes que o pesquisador deve ter ao planejar e executar um experimento é descrever de forma detalhada as atividades que favoreçam a repetição do estudo. A repetição aumenta o conhecimento a respeito dos conceitos estudados, bem como, propicia maior confiabilidade às teorias verificadas.

Por isso, não faz nenhum sentido falar sobre o avanço de qualquer área da ciência a partir de experiências isoladas. Em ES não é diferente, os resultados de uma única experiência não pode ser visualizado como sendo representativo da realidade

global [31]. Segundo Do et al. [19], a descoberta científica não é confiável a menos que possa ser replicada de forma independente. Por isso, várias repetições dos experimentos em ES devem ser feitas para checar a confiança dos resultados obtidos [31].

Dessa forma, baseado em aspectos como o delineamento experimental, a instrumentação e a descrição adequada dos resultados [65], propõe-se a verificar, também, como está a qualidade dos experimentos quanto à replicabilidade. Os resultados e conclusões obtidas sobre a qualidade dos experimentos quanto à replicabilidade dos estudos serão descritos em detalhes nos Capítulos 5 e 6.

2.1.2.3 Rigor Estatístico dos estudos

O rigor estatístico de um estudo empírico é essencial para a interpretação dos resultados e para verificar a validade das conclusões. Um estudo sem um adequado tratamento estatístico não é capaz de fornecer informação suficiente para tirar conclusões sobre a aceitação ou rejeição das hipóteses da pesquisa [21].

Além disso, se o rigor estatístico é insuficiente, a probabilidade de encontrar resultados com significância estatística é pequena, e os resultados do estudo provavelmente serão insignificantes para a ciência. Assim, a incapacidade de fornecer um nível adequado de rigor estatístico tem implicações tanto para a execução como para os resultados do experimento.

No intuito de verificar como está a qualidade dos experimentos quanto ao tratamento estatístico, foi proposto verificar o nível de qualidade em relação ao rigor estatístico dos estudos. Os resultados e conclusões sobre a qualidade dos experimentos quanto ao rigor estatístico serão descritos em detalhes nos Capítulos 5 e 6.

2.1.3 Avaliação da qualidade

O trabalho de Kitchenham et al. [36], sobre Engenharia de Software Baseada em Evidências (EBSE) (do inglês *Evidence-Based Software Engineering*) em 2004, recomendou a avaliação da qualidade dos estudos, porém, naquele momento, este assunto ainda era bastante teórico e incipiente. Em seguida, várias RSLs foram publicadas sem dar muita ênfase à avaliação da qualidade e com isso o assunto foi tratado superficialmente sem que a literatura indicasse como fazer avaliação da qualidade dos estudos.

Juntamente com a publicação da segunda versão do Guia de Kitchenham et al. [38], em 2007, havia uma lista com 50 perguntas destinadas a avaliar a qualidade dos experimentos, organizados de acordo com a fase do estudo que iria ser avaliado. No entanto, não há, ainda, uma definição de “qualidade” no escopo da avaliação de estudos empíricos.

A literatura de EBSE sugere *checklists* de qualidade para diferentes tipos de estudos experimentais, no entanto, cada uma das fontes pesquisadas, identifica um conjunto ligeiramente diferente de questões e não há um padrão acordado para esse tipo de avalição.

Neste sentido, apresenta-se nas próximas seções o contexto da pesquisa sobre avaliação da qualidade de estudos empíricos através do conceito de qualidade, importância da avaliação de qualidade, dimensões de garantia de qualidade, estratégias de avaliação de qualidade e trabalhos relacionados.

2.1.3.1 Conceito de Qualidade

Como afirmado anteriormente, não existe um conceito absoluto de “qualidade” quanto à avaliação de estudos empíricos, porém, autores como Higgins et al. [27], Moher et. al [48] e Kitchenham et al. [36], têm apresentado algumas constatações:

- Moher et. al [48], relata em um estudo sobre tendência em ensaios sistemáticos da área de medicina que a validade do estudo sistemático quanto à sua concepção e execução, diminui a probabilidade de erros sistemáticos o que minimiza o viés e aumenta a confiabilidade da pesquisa.
- O *Handbook* “*The Cochrane Collaboration Open Learning Material for Reviewers*” sobre revisões sistemáticas [27] (pág. 49-58), dedica um dos capítulos à analise da avaliação da qualidade e recomenda que o conceito de qualidade diga respeito à validade dos estudos e como estes limitam a inclinação e tendências na condução das pesquisas e guiam à interpretação dos resultados.
- Kitchenham et al. [36], enfatizam que a qualidade de uma pesquisa sistemática deve ser identificada através da ampliação da validade interna (planejamento e execução) e externa (generalização e aplicação), esta identifica características que impactam na interpretação dos resultados.

Com isso, esse estudo seguirá as recomendações propostas pelos autores citados acima, onde o conceito de qualidade deve estar relacionado ao viés da

pesquisa, ao aumento da validade (interna e externa) e à correta interpretação dos resultados.

2.1.3.2 Importância da Avaliação da Qualidade

Avaliar a qualidade dos estudos primários é importante porque, se não for feita de forma organizada e válida, os resultados do experimento podem ser comprometidos [40].

Um estudo terciário [30] sobre a importância de utilização de pesquisas bibliográficas na avaliação da qualidade revelou que é fundamental avaliar a qualidade dos estudos em revisões sistemáticas e meta análises. O experimento mostrou que em virtude de não terem realizado uma análise da qualidade dos estudos, o efeito do tratamento foi modificado e com isso tornou-se menos benéfico.

Uma revisão sistemática sobre homeopatia [57], realizada em 2005, incluiu estudos de baixa qualidade e os resultados sugeriram que a homeopatia funcionava bem, no entanto experimentos que envolveram estudos de qualidade como experimentos classificados como controlados não mostraram efeito significativo dos remédios homeopáticos.

Em Engenharia de Software, autores têm relatado a falta de padronização quanto aos métodos e procedimentos na condução dos experimentos. Uma revisão sistemática, conduzida por Dyba et al. [21], constatou que não havia informação suficiente para uma análise completa dos estudos. Segundo Dyba et al. [21], em quatorze experimentos analisados não houve relato de qualquer análise estatística e em sete casos não conseguiu-se averiguar nem quais testes respondiam as hipóteses ou questões de pesquisa.

Dessa forma, é de crucial importância que haja avaliação da qualidade dos estudos, pois dessa forma pode-se diminuir a possibilidade de erros sistemáticos e aumentar a confiabilidade dos resultados da investigação.

2.1.3.3 Dimensões de Garantia de Qualidade

De forma semelhante ao *checklist* sugerido por Dieste et al. [18], desenvolveu-se nesta pesquisa, um instrumento de qualidade baseado numa escala de qualidade proposta por Dyba et al. [20], e organizado de acordo com as seis dimensões de garantia de qualidade propostas por Kitchenham et al. [36]: a) contexto experimental, b) delineamento experimental, c) condução do experimento e coleta dos dados, d) análise, e) apresentação e f) interpretação dos resultados.

- A finalidade das diretrizes de **contexto experimental** é garantir que os objetivos da pesquisa estejam devidamente definidos e se a descrição da pesquisa fornece detalhes suficientes para outros pesquisadores e profissionais. Aborda três elementos centrais: (i) Informações básicas sobre as circunstâncias industriais em que um estudo empírico ocorre ou em que uma nova técnica de Engenharia de Software é desenvolvida. (ii) Discussão das hipóteses de pesquisa e como elas foram obtidas e. (iii) Informações sobre pesquisas relacionadas;
- O **delineamento experimental** do estudo descreve os produtos, recursos e processos envolvidos no estudo, incluindo: (i) População estudada, (ii). Lógica e técnica de amostragem da população. (iii) Processo para a atribuição e gestão dos tratamentos, e. (iv) Métodos utilizados para reduzir o viés e determinar o tamanho da amostra. O objetivo é assegurar que o projeto é adequado para os objetivos do estudo;
- A **condução do experimento e coleta dos dados** envolve a coleta das medidas de resultados experimentais. Este é um problema particular para os experimentos de ES, porque as medidas não são padronizadas. Assim, o objetivo das orientações de coleta de dados é garantir que será bem definido o processo de coleta o suficiente para a experiência ser replicada. Além disso, é preciso monitorar e registrar quaisquer desvios dos planos experimentais, isto inclui tanto o acompanhamento das desistências como as questões que ficaram sem respostas;
- O objetivo das diretrizes de **análise** é garantir que os resultados experimentais sejam analisados corretamente. Basicamente, os dados devem ser analisados de acordo com o projeto de estudo. Assim, a vantagem de fazer um projeto cuidadoso, bem definido é que a análise posterior é geralmente simples e clara;
- O principal objetivo da **interpretação dos resultados** (conclusões) é que deve seguir diretamente os resultados. Assim, os pesquisadores não devem introduzir novos materiais na seção de conclusões, bem como, ter o cuidado para não deturparem as suas conclusões. Por exemplo, omitir o significado dos resultados que conflitam com uma pesquisa anterior. É também importante que os investigadores qualifiquem seus resultados de forma adequada;

- A **apresentação dos resultados** é tão importante como a análise propriamente dita. O leitor de um estudo deve ser capaz de compreender a razão e concepção para o estudo, a análise e o significado dos resultados.

2.1.3.4 Estratégias de avaliação da qualidade

Determinar a qualidade de um estudo não é algo simples, uma vez que o conhecimento sobre a validade do constructo “qualidade” ainda é incerto. Em geral, a "qualidade" de um estudo está ligada aos métodos de pesquisa utilizados e a validade dos resultados gerados [20]. Neste sentido, determinar a qualidade de um estudo implica em empregar métodos para diminuir os desvios da pesquisa e aumentar a validade interna e externa do experimento [36].

Dieste et al. [18] explicam que um bom experimento é aquele que usa a aleatorização para criar grupos experimentais homogêneos, utiliza sujeitos e pesquisadores de forma imparcial e acompanha todos os resultados com o objetivo de minimizar o viés da pesquisa.

Em tese, um procedimento para determinar a qualidade de uma experiência seria comparar o resultado desta com a média de todas as experiências de um conjunto de experimentos por meio de meta análise, porém, isso não é possível nem em áreas maduras como medicina, imagine em Engenharia de Software [18]. Dessa forma, no intuito de minimizar erros sistemáticos nos estudos, autores têm sugerido que sejam utilizadas outras abordagens para estimar a qualidade de um estudo, por exemplo, usar critérios simples de qualidade, listas de verificação ou escalas de qualidade [27][34].

Conforme sugerido por Higgins et al. [27] e Khan et al. [34], podemos estimar a qualidade de um estudo utilizando um conjunto simples de critérios de qualidade e respondê-los de forma qualitativamente. Por exemplo: claramente atendido, não atendido ou parcialmente atendido. Outra abordagem, apontada por Rowe et al. [55], é utilizar listas de verificação, pois são compostas por um número considerável de questões relacionadas com a qualidade e podem aumentar a confiabilidade da análise. Para o desenvolvimento deste trabalho de mestrado escolheu-se utilizar uma escala de qualidade. Uma instância desta escala pode ser visualizada na Figura 2.2

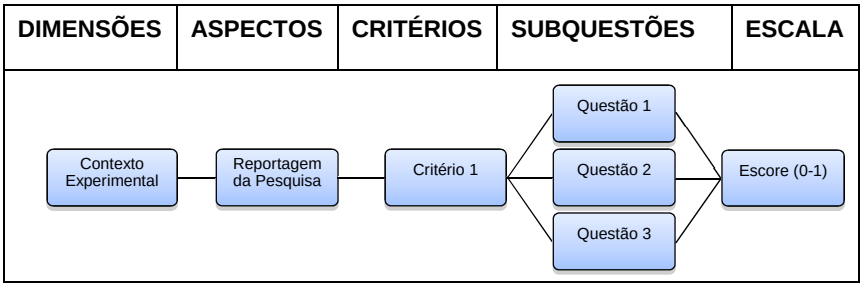


Figura 2.2 - Instância da escala de qualidade

A instância da escala de qualidade da Figura 2.2 representa o instrumento de avaliação da qualidade utilizado. Este foi organizado de acordo com as seis dimensões de garantia da qualidade de um experimento propostas por Kitchenham et al. [36] e contempla os quatro principais aspectos de qualidade que precisam ser considerados quando se avalia um estudo empírico. Em torno dos quatro aspectos de qualidade há um conjunto de critérios que podem ser satisfeitos à medida que as subquestões relacionadas a este são atendidas. O método de avaliação usado para mensurar o conjunto de questões foi um escore de três pontos numa escala que vai de zero a um.

2.1.3.5 Trabalhos Relacionados

Grande parte das pesquisas disponíveis, e estudadas, sobre avaliação da qualidade de estudos empíricos, está preocupada apenas com o processo de inclusão e exclusão dos estudos em revisões sistemáticas ou meta análises. Poucos trabalhos em ES relatam a avaliação da qualidade dos estudos com o objetivo de reportar a qualidade global das pesquisas.

Em buscas realizadas com o *Google Scholar* não foi encontrado pesquisas empíricas relacionadas à evolução da qualidade em relação aos mecanismos de suporte utilizados em EE, replicabilidade e rigor estatístico dos experimentos da Comunidade de ESE. A seguir, a descrição de alguns trabalhos que estão relacionados com a avaliação da qualidade:

Kampenes et al. [33] publicaram um estudo cujo objetivo foi avaliar a qualidade dos experimentos em Engenharia de Software no período entre 1993 e 2002, quanto à validade e qualidade dos relatórios de divulgação dos resultados. Concluíram que em geral os resultados mostraram que os experimentos de Engenharia de Software foram planejados de forma inadequada e pouca relevância foi dada ao tratamento estatístico dos dados.

Ainda, segundo Kampenes et al. [33], a importância do efeito (resultados) foi raramente relatada nos estudos, bem como, não foram interpretadas em relação à sua importância prática para o contexto da pesquisa. Além disso, a análise revelou a necessidade de informações completas e padronizadas, o que é crucial para a compreensão e replicabilidade de experimentos em ES.

Kitchenham et al. [39] investigaram a usabilidade de uma lista de verificação de avaliação de qualidade, com o intuito de determinar a importância em reportar os experimentos e especificar um processo adequado para avaliar a qualidade. Os seguintes resultados foram obtidos: (i) A confiabilidade entre avaliadores foi ruim para as avaliações individuais, mas foi melhor para avaliações conjuntas. (ii) A soma dos critérios foi considerada mais confiável do que questões individuais. Com isso, o estudo aconselhou o uso de pelo menos dois revisores, rodadas de discussão para resolução dos conflitos, bem como, verificar a qualidade dos estudos quanto à soma dos critérios utilizados.

Silva et al. [59], e Kitchenham et al. [40], investigaram se houve aumento na qualidade dos experimentos e quasi-experimentos que envolvessem a experiência humana, no período de 1993 a 2010. Os estudos relataram que o número de RSLs em ES está aumentando, a qualidade geral dos estudos de ESE está melhorando, e o número de pesquisadores e instituições de investigação que estão conduzindo RSLs em todo o mundo também está aumentando. No entanto, os pesquisadores constataram que a maioria das RSLs não avalia a qualidade dos estudos primários, diminuindo assim o impacto sobre a prática de ES.

Em 2011, Dieste et al. [18] tentaram identificar se existe uma relação entre a validade interna e viés em experiências de ES e quais os aspectos de validade interna (ou falta de validade interna) causam maior quantidade de viés. Uma escala de qualidade foi aplicada a um conjunto de 28 experiências que faziam parte de duas RSLs e os resultados indicaram que não há uma relação entre o índice de qualidade global (resultante da escala de qualidade) e o viés (inclinação da pesquisa).

No entanto, os resultados da pesquisa de Dieste et al. [18] identificaram correlações importantes entre o viés e alguns aspectos particulares da validade interna, bem como, identificaram tendências que mostram que os instrumentos de controle de qualidade (escalas) utilizados em RSLs não deve conter uma infinidade de itens e sugeriu que devemos usar critérios cuja relação com o viés da pesquisa é razoavelmente bem definida. Por fim, concluiu que é possível aplicar avaliação da

qualidade em RSLs, porém, tendo cuidado quanto aos limites de avaliação que diz respeito aos aspectos de validade interna.

Kitchenham et al. [37] investigaram três diferentes técnicas (avaliações individuais, conjuntas e consensuais entre avaliadores) que podem ser utilizadas para melhorar a confiabilidade da avaliação da qualidade, em particular, o número de colaboradores, utilizando um conjunto de critérios de consulta e de avaliação. Os resultados levaram às seguintes conclusões:

- Recomendar o uso de três revisores independentes, adotando uma avaliação média das pontuações para os critérios individuais;
- Usar uma lista (escala) de verificação de qualidade, porém, garantir que todos os avaliadores entendam como utilizá-las;
- Se houver forte divergência entre os três avaliadores para um artigo em específico, aplicar uma exceção ao processo, encontrar um quarto avaliador ou instituir uma nova rodada de discussão.

Em síntese, os trabalhos [18][33][37][40][59] abordaram preocupações e resultados relacionados à validade interna e externa, usabilidade de listas de verificação, viés das pesquisas, bem como, procedimentos e técnicas para avaliação de qualidade. É importante ressaltar que nenhum dos trabalhos fez algum tipo de relação da qualidade quanto ao uso de mecanismos de suporte no planejamento e/ou execução dos estudos.

2.2 Resumo

Esse capítulo descreveu o contexto da pesquisa sobre avaliação da qualidade de Estudos Empíricos (EE). Os próximos capítulos trarão à tona as estratégias e metodologia utilizada para a análise da qualidade de experimentos controlados no contexto da ESE, bem como, a preocupação em fornecer bases para uma futura definição de um modelo, processo ou guia de avaliação da qualidade de EE em ES.

Capítulo

3

3. Instrumento de Avaliação

Este capítulo discute a estratégia utilizada para análise da qualidade dos experimentos controlados, bem como, a construção do instrumento de qualidade criado, principalmente, com base em critérios amplamente utilizados por pesquisadores da área de ES.

3.1 Conceitos e Definições

Em razão de não haver um padrão para avaliar a qualidade de estudos empíricos, a literatura apresenta um conjunto ligeiramente diferente de questões e estratégias de avaliação da qualidade. Dessa forma, nesta seção serão apresentados alguns conceitos e definições a respeito das abordagens utilizadas para estimar a qualidade de estudos primários.

Diversos autores têm sugerido usar diferentes táticas de estimativa de validade dos estudos, dentre as quais iremos destacar: Abordagens simples, Listas de verificação e Escalas de qualidade [27][34][52].

- **Abordagens Simples –** Conjunto de critérios de validade, normalmente respondido de forma qualitativa, utilizando uma escala classificatória (por exemplo: o critério foi claramente atendido, não atendido ou parcialmente atendido), e o estabelecimento de um risco de viés (por exemplo: baixo, moderado, alto).
- **Listas de Verificação –** São instrumentos baseados em itens de qualidade, geralmente compostas por um número considerável de questões relacionadas com a qualidade, com respostas, normalmente, dicotômicas (sim ou não). Por exemplo, existe uma questão de pesquisa bem definida? A análise dos dados foi suficientemente rigorosa?
- **Escalas de Qualidade –** São instrumentos baseados em uma série de itens de qualidade, marcados numericamente para fornecer uma estimativa quantitativa da qualidade global do estudo. Normalmente, a classificação da pontuação tende a ser subjetiva e a qualidade global pode ser gerada através da soma do total de itens ou atribuindo-lhes pesos em relação à importância dos critérios avaliados.

Moher et al. [48] identificaram nove listas de verificação e 25 diferentes tipos de escalas de qualidade, as quais têm sido utilizadas para avaliar a validade e "qualidade" de estudos primários na área de medicina. Estas listas e escalas incluem de três a 57 questões de qualidade e são baseadas em critérios que são mencionados em livros da área médica.

Há dúvidas quanto à validade das listas e escalas de avaliação da qualidade usando uma ampla gama de critérios de qualidade. Segundo Higgins et al. [27], muitos desses instrumentos de avaliação como os sugeridos por Moher et al. [48] são passíveis de confundir o julgamento feito pelos avaliadores e consequentemente

podem colocar em risco a confiabilidade dos resultados dos estudos, pois não há um padrão em que possamos afirmar como válido. Dessa forma, nenhuma das estratégias disponíveis atualmente para medir a validade ou a "qualidade" dos estudos pode ser recomendada sem reservas.

É importante ressaltar que listas de verificação e escalas de qualidade com vários itens e sistemas de pontuação complexos levam mais tempo para concluir uma avaliação do que abordagens simples e que não fornecem, necessariamente, avaliações mais confiáveis [30]. Em se tratando de avaliação da qualidade em revisões sistemáticas, estudos de mapeamento sistemático ou meta análises, autores têm sugerido usar métodos simples para avaliar a validade ou “qualidade” dos estudos [27] [34][35].

Das alternativas descritas nesta seção (abordagens simples, listas de verificação e escalas de qualidade), escolheu-se utilizar a estratégia escala de qualidade. Abaixo descrevemos as razões para esta decisão:

- Este trabalho não se propõe realizar uma avaliação de qualidade com o objetivo de incluir ou excluir estudos primários, como é a maioria dos casos em se tratando de revisões sistemáticas, estudos de mapeamento sistemático ou meta análises;
- O objetivo da pesquisa é fornecer uma estimativa quantitativa da qualidade global dos estudos (experimentos controlados);
- Escalas de qualidade são instrumentos baseados em uma série de itens de qualidade, organizados numericamente e por categorias (escalas numéricas) para fornecer uma estimativa quantitativa da qualidade global do estudo;
- Escalas de qualidade retornam valores contínuos e por isso podem ser correlacionados com outras variáveis do estudo.

3.2 Instrumento de Avaliação de Qualidade

Para construir o instrumento de qualidade utilizado para analisar a qualidade dos experimentos controlados, foi realizada uma revisão da literatura *ad-hoc* em busca de *checklists* e critérios utilizados para avaliação da qualidade de EE. A revisão da literatura indicou que não existe um conjunto padrão de perguntas usadas em ES para avaliar a qualidade de EE.

Desta forma, decidiu-se usar uma lista de critérios proposta por Dyba e Dingsoyr [20], como base para desenvolvermos o instrumento. A principal razão para usar as questões propostas por Dyba e Dingsoyr justifica-se pelo fato dos itens terem sido baseados em uma lista de verificação que é amplamente utilizado por pesquisadores de outras áreas e, atualmente, está sendo adotada por pesquisadores de ES que realizam revisões sistemáticas [1][3][10].

Novos critérios foram adicionados à lista, pois a proposta por Dyba e Dingsoyr não considerou questões como aleatorização ou duplo cego (do inglês *double-blind*), no qual nem os pesquisadores nem os participantes sabem quais indivíduos pertencem ao grupo de teste ou ao grupo controlado.

A lista proposta por Dyba e Dingsoyr aborda quatro grandes questões de qualidade que precisam ser consideradas quando se avalia estudos primários:

- **Relatório –** Verifica se o contexto e resultados do estudo foram descritos de forma clara e adequada;
- **Rigor –** Verifica se o método de pesquisa principal adotou uma abordagem completa e adequada;
- **Credibilidade –** Verifica se a descoberta foi bem apresentada e significativa;
- **Relevância –** Verifica se a descoberta é útil para a indústria e comunidade científica de engenharia de software.

O instrumento de qualidade foi organizado de acordo com as seis dimensões de garantia de qualidade de um experimento, propostos por Kitchenham et al. [36], e sugerido por Dieste et al. [18], são elas: contexto experimental, delineamento experimental, condução do experimento e coleta de dados, análise, apresentação dos resultados e interpretação dos resultados.

A lista de critérios utilizada foi exaustivamente analisada e com isso, decidiu-se excluir questões que não interessam ao estudo, uma vez que iríamos avaliar experimentos formais, bem como ajustamos as subquestões para reduzir a subjetividade dos critérios.

Em virtude dos critérios propostos por Dyba e Dingsoyr serem genéricos e não específicos a experimentos formais, foram adicionados novos critérios, relacionados ao processo de condução de um experimento: Definição de Objetivos, Planejamento, Execução, Análise, Apresentação e Empacotamento [65].

É importante ressaltar que o instrumento de qualidade desenvolvido está estruturado de tal maneira que possa ser evoluído para avaliar a qualidade de outros tipos de estudos, uma vez que apresenta critérios gerais e outros específicos para experimentos. No entanto, essa classificação poderá mudar dependendo do tipo de estudo avaliado. Esse é apenas um primeiro passo no sentido de termos um conjunto comum de critérios e que poderá ser reutilizado para avaliar a qualidade de outros tipos de EE.

A categorização em critérios gerais e específicos foi baseada nas fontes de pesquisa utilizada para montagem do instrumento [18][20][36][65], por exemplo, questões usadas para avaliação de experimentos foram definidas como específicas e os demais critérios foram considerados gerais. Em tese, um critério definido como geral pode ser aplicado em mais de um tipo de estudo ou a todos os tipos de estudos e os específicos se aplicam aos experimentos. Ao todo foram definidas 39 subquestões gerais e 16 subquestões específicas.

Apesar do esforço, reconhece-se que o conjunto de questões selecionadas não garante que a lista está completa e que, em algumas situações de avaliação, algumas subquestões podem ser irrelevantes. A versão final do instrumento de avaliação utilizado é mostrada no **Quadro 3.1**.

Dimensão avaliada	Aspecto de Avaliação	Critérios de qualidade	Questões a considerar	Critério
Contexto Experimental	**Report**	**C1**. Os autores indicaram claramente os objetivos da pesquisa?	Existe uma razão por que o estudo foi realizado?	Geral
			Os autores indicaram de forma clara as questões de pesquisa do estudo?	Geral
			Os autores definiram as hipóteses do estudo e suas teorias subjacentes?	Específico
		C2. Existe uma descrição adequada do contexto em que a pesquisa foi realizada?	Há uma descrição do contexto em que o experimento foi realizado?	Geral
			Há uma descrição da natureza da organização (ES) em que o experimento foi realizado?	Geral
			Há uma descrição das habilidades e experiências dos participantes do experimento? (por exemplo, com uma linguagem, um método, uma ferramenta ou um domínio de aplicação)	Geral
			Há uma descrição do tipo de produto de software utilizado? (por exemplo, uma ferramenta de design, um compilador, etc.)	Geral
			Foi descrito o processo de software utilizado? (por exemplo, um processo padrão da empresa, os procedimentos de garantia de qualidade, o processo de gerenciamento de configuração).	Geral
		C3. Houve a descrição de pesquisas similares e como estas se relacionam com o estudo?	Houve a descrição de pesquisas similares e como estas se relacionam com o estudo?	Geral

Dimensão avaliada	Aspecto de Avaliação	Critérios de qualidade	Questões a considerar	Critério
Delineamento Experimental	**Rigor**	**C4**. Os autores descrevem o projeto experimental? O projeto de pesquisa foi adequado para resolver os objetivos da pesquisa?	Os autores descrevem claramente a metodologia escolhida?	Geral
			Os autores definem ou descrevem os tratamentos das variáveis do experimento?	Específico
			O pesquisador fez uma discussão em relação aos métodos utilizados na pesquisa?	Geral
		C5. A estratégia de recrutamento foi adequada aos objetivos da pesquisa?	O pesquisador explicou como os participantes ou os casos foram identificados e selecionados?	Geral
			Os participantes ou casos foram definidos e descritos com precisão?	Geral
			Há indícios de que os participantes ou casos foram representativos de uma população definida?	Geral
			Os pesquisadores explicaram por que os participantes ou casos selecionados foram adequados? Explicam como eles tiveram acesso ao conhecimento necessário ao estudo?	Geral
			O experimento realizou alguma pré-experiência ou cálculo para identificar ou estimar o tamanho mínimo necessário para a amostra?	Específico
		C6. Os autores indicam e descrevem de forma clara as variáveis da pesquisa?	Os autores indicaram as variáveis dependentes e independentes do experimento?	Específico
			As variáveis do experimento são descritas de forma precisa?	Específico
			As variáveis de resposta (dependentes) foram definidas e medidas relacionadas em termos de sua relevância para as metas listadas na seção objetivos da pesquisa?	Específico
		C7. O pesquisador define o processo aplicado ao tratamento de objetos e sujeitos?	Os sujeitos são alocados aos grupos de maneira imparcial, de modo a não comprometer o experimento?	Específico
			A distribuição dos tratamentos aos grupos foi realizada de forma aleatória?	Específico

Dimensão avaliada	Aspecto de Avaliação	Critérios de qualidade	Questões a considerar	Critério
Condução do Experimento e Coleta de Dados	**Rigor**	**C8**. Os autores descrevem os procedimentos de coleta de dados e definição de medidas?	Foram utilizados métodos de controle de qualidade para garantir a consistência, integridade e precisão dos dados coletados?	Específico
			Todas as medidas foram claramente definidas? (por exemplo, a escala, unidade ou regra de contagem)	Geral
			Foi utilizado método para coleta adequada dos dados?	Geral
			O pesquisador justifica os métodos que foram escolhidos?	Geral
			Há a definição clara de um instrumento utilizado para coleta de dados?	Geral
			O experimento informa o cronograma de execução, bem como, tempo ou período utilizado para realizar a coleta dos dados?	Específico
	Credibilidade	**C9**. Os autores descreveram ou apresentaram alguma medida de concordância entre os avaliadores?	Foi apresentada alguma medida de concordância entre examinadores?	Geral
		C10. A relação entre o pesquisador e os participantes foi considerada adequada?	O pesquisador examinou criticamente o seu próprio papel, preconceito e potencial influência na formulação de questões de investigação, no recrutamento da amostra, coleta de dados e análise e seleção de dados para a apresentação?	Geral
			O pesquisador informou ocorrências de eventos (desvios) durante o estudo (condução e coleta) e se considerou as implicações (como foram tratadas) das alterações no projeto de pesquisa?	Geral
		C11. As perdas e desistências de participantes ocorridas desde a seleção até o final do estudo foram descritas?	Os autores relataram os casos de abandono de participantes depois de terem sido distribuídos por grupos de tratamento?	Específico

Dimensão avaliada	Aspecto de Avaliação	Critérios de qualidade	Questões a considerar	Critério
Análise dos Dados	**Rigor**	**C12**. A análise de dados foi suficientemente rigorosa?	Os autores descreveram de forma detalhada os procedimentos para análise dos dados, bem como, justificaram suas escolhas?	Geral
			Os autores forneceram referências às descrições dos procedimentos de análise de dados?	Geral
			Os autores mencionam níveis de significância estatística e tamanhos de efeito?	Específico
			Os *outliers* são mencionados e casos encontrados são tratados durante a análise?	Geral
			Foi utilizada alguma técnica ou forma de análise dos dados? (Por ex: análise temática, grounded theory, etc.).	Geral
			Foram apresentados dados suficientes para apoiar os resultados?	Geral
			Os dados contraditórios foram levados em consideração?	Geral
			Foram utilizados métodos de controle de qualidade para verificar os resultados?	Geral
		C13. Os autores discutiram o potencial viés do experimento quanto a análise dos dados?	Os autores discutiram as implicações do tratamento utilizado no estudo?	Geral
			O treinamento foi equivalente para todos os grupos de tratamento?	Específico
			Houve ocultação da alocação de forma tal que os pesquisadores não sabem que tratamento cada sujeito foi submetido?	Específico
Interpretação dos Resultados	**Credibilidade**	**C14**. Os autores declararam de forma clara os resultados do estudo?	Os autores apresentam conclusões de forma clara?	Geral
			As conclusões são justificadas pelos resultados e as conexões entre os resultados e conclusões são apresentadas com clareza?	Geral
			Os autores discutem suas conclusões em relação às questões de investigação originais?	Geral
			O pesquisador discutiu a credibilidade dos seus resultados?	Geral
		C15. São mencionadas as ameaças à validade e como essas ameaças afetam os resultados e conclusões?	São mencionadas as ameaças à validade e também como essas ameaças afetam os resultados e conclusões?	Geral

Dimensão avaliada	Aspecto de Avaliação	Critérios de qualidade	Questões a considerar	Critério
Apresentação dos Resultados	**Credibilidade**	**C16**. O estudo fornece testes estatísticos apropriados e referenciados?	O estudo descreve e cita as referências para os procedimentos estatísticos utilizados?	Geral
			O experimento define critérios que sustentam o rigor estatístico?	Específico
		C17. As significâncias estatísticas são mencionadas com os resultados?	O experimento apresenta resultados baseados nos dados quantitativos, incluindo o tamanho do efeito e os limites de confiança?	Específico
	Relevância	**C18**. O estudo tem valor para a pesquisa e prática? O estudo poderá ser utilizado por outros pesquisadores ou profissionais?	O pesquisador discutiu a contribuição do estudo ou interpreta os resultados em relação ao conhecimento ou entendimento já existente?	Geral
			A investigação identifica novas áreas em que pesquisas são necessárias?	Geral
			O pesquisador discutiu se e como os resultados podem ser transferidos para outras populações, ou considerou outras maneiras em que a pesquisa pode ser utilizada?	Geral
	Report	**C19**. A pesquisa apresenta ou indica a disponibilidade dos dados brutos?	O estudo apresenta ou indica onde os dados brutos estão disponíveis para análise por outros revisores ou auditores independentes?	Geral

Quadro 3.1 – Instrumento de avaliação da qualidade

O instrumento de qualidade é composto por 19 critérios e 55 subquestões, entre elas algumas perguntas são subjetivas. A Tabela 3.1 mostra o número de perguntas que adotamos e os autores que as definiram.

Tabela 3.1 - Número de critérios adotados e autores que definiram

Seleção dos Critérios		
Autores	**Critérios**	**Subquestões**
Dyba e Dingsoyr [20]	09	32
Oscar Dieste [18]	03	8
Kitchenham [36]	01	6
Definidos pelo Autor da Dissertação	06	9
Total	**19**	**55**

3.3 Avaliação do Instrumento de Qualidade

Um estudo piloto foi realizado antes do estudo principal no intuito de avaliar o instrumento proposto. O estudo piloto foi realizado por quatro avaliadores, utilizando cinco *papers* (escolhidos de forma aleatória) para a avaliação. A principal métrica utilizada para analisar a confiabilidade do instrumento foi o *Cohen's Kappa*, um coeficiente para escalas nominal amplamente utilizado na literatura [13].

Para avaliar o nível de concordância entre os avaliadores realizamos o teste estatístico *Kappa* e para saber se os resultados seriam ou não satisfatórios nos baseamos na interpretação sugerida por Landis JR e Koch GG [43], conforme observamos a Tabela 3.2.

Tabela 3.2 - Interpretação sugerida para os valores de Kappa

Valores de *Kappa*	Interpretação
<0	Concordância nula
0 – 0.19	Concordância pobre
0.20 – 0.39	Concordância justa
0.40 – 0.59	Concordância moderada
0.60 – 0.79	Concordância substancial
0.80 – 1.00	Concordância quase perfeita

Apesar dos avaliadores terem experiência com a execução de experimentos e no intuito de minimizar os erros decorrentes do uso do instrumento, foi realizado um treinamento antes da realização do estudo piloto, com duração de aproximadamente duas horas, bem como, foram feitas rodadas de discussão para assegurar o entendimento dos critérios avaliados pelos avaliadores.

A amostra utilizada foi avaliada pelos quatro avaliadores utilizando o instrumento de qualidade proposto, após a aplicação do instrumento de avaliação, os *papers* foram classificados em Excelente (A), Muito Bom (B), Bom (C), Regular (D) e Ruim (E), conforme classificação de qualidade sugerida por Beecham et al. [9].

Tabela 3.3 - Resultados do Teste Piloto com o Coeficiente de *Kappa*

Papers	**Avaliadores**			
	1	**2**	**3**	**4**
PS089	B	C	B	B
PS100	D	D	D	C
PS242	B	C	B	B
PS256	C	C	D	C
PS836	D	D	D	D
***Kappa* Geral**			**0.623**	
P-valor geral			< 0.001	
Intervalo de 95% de Confiança do Kappa			**sup**: 0.803 **inf**: 0.442	

O teste estatístico *Kappa* foi calculado com base na fórmula básica $K = \frac{(Po-Pe)}{(1-Pe)}$, o valor de **Po** é a porcentagem de classificações que são iguais entre os quatro avaliadores e **Pe** é a probabilidade de concordância devido ao acaso. Conforme podemos verificar na Tabela 3.3 o valor de *Kappa* Geral foi **0.623**, indicando que a concordância entre os avaliadores foi classificada como **substancial**, a segunda melhor classificação possível. Com isso, o estudo piloto sugere que o instrumento avaliado apresenta uma boa confiabilidade, porém, os avaliadores sugeriram pequenas alterações na escrita das subquestões quanto ao entendimento dos critérios usados no instrumento, com o objetivo de melhorar a qualidade dos materiais antes da realização do estudo principal.

3.4 Resumo

Foram apresentados conceitos e definições em relação a estratégias para avaliação da qualidade, assim como, a tática usada por essa pesquisa para avaliar experimentos controlados. Descrevemos, também, um estudo piloto que foi usado para avaliar a confiabilidade do instrumento, bem como, para treinar e alinhar a definição da metodologia com os avaliadores. O próximo capítulo descreve a execução do estudo principal.

Capítulo

4

4. Metodologia de Avaliação da Qualidade dos Experimentos Controlados

Este capítulo discute a execução da estratégia de análise da qualidade, bem como, a realização do cálculo do índice de qualidade, o processo de busca e seleção dos estudos, a forma de alocação dos *papers* aos avaliadores, a estratégia de coleta dos dados e a execução deste estudo de mestrado.

4.1 Planejamento

O planejamento de um estudo experimental é muito importante, pois dele depende todo o processo de execução do experimento. Segundo Travassos et. al [65], negligenciar alguma etapa no processo de planejamento pode acarretar em resultados errôneos e com isso, causar modificações em todo os procedimentos que já haviam sido feitos, o que é muito difícil ou, às vezes, impossível de realizar.

Conforme descrito na seção de metodologia do Capítulo 1, o procedimento básico de avaliação da qualidade foi selecionar todo o conjunto de estudos publicados até 2012, dos três principais veículos internacionais da comunidade de Engenharia de Software Empírica. A saber: EASE – *International Conference on Evaluation and Assessment in Software Engineering*, ESEM – *International Symposium on Empirical Software Engineering and Measurement* e ESEJ – *Empirical Software Engineering Journal* e analisá-los com base em uma escala de qualidade criada, principalmente, com base em listas de verificação amplamente utilizadas por pesquisadores da área de ES.

Restringiu-se a investigação aos artigos publicados nos veículos mencionados acima devido ao objetivo do estudo que é realizar uma avaliação dos trabalhos da comunidade de ESE. Os materiais e estratégias de planejamento e execução da avaliação proposta são discutidos, em detalhes, nas seções seguintes.

4.2 Classificação dos itens de qualidade dos papers

Em várias áreas de conhecimento, e em ESE não é diferente, é comum o uso de mecanismos de avaliação e classificação para mensurar a realidade sobre um objeto em estudo.

Para realizar essas mensurações, os pesquisadores precisam desenvolver instrumentos adequados para que as medidas sejam válidas e correspondam efetivamente ao que se deseja medir. Precisam, também, reduzir ao mínimo possível o erro amostral (aumentar a confiabilidade) diante do escopo, custo e prazo disponível, e desta forma, os resultados das medidas apresentem um reflexo da realidade avaliada [15][47].

A estratégia utilizada para classificar os critérios do instrumento de qualidade proposto foi usar uma escala do tipo *Likert* [44], que é um instrumento que permite aos

pesquisadores atribuírem respostas gradativas sobre suas opiniões em relação aos itens avaliados.

Apesar do esforço para reduzir a subjetividade dos critérios, percebe-se que nem todas as respostas para o conjunto de subquestões selecionadas poderiam ser dicotômica, ou seja, sim ou não, dessa forma utilizou-se uma escala de três pontos, descrita com as seguintes classificações e valores: não atende (0), atende parcialmente (0,50) e atende totalmente (1), conforme mostra a Tabela 4.1.

Tabela 4.1 - Escala do Tipo Likert Utilizada

Escala	**Valores**		
Likert 3	Não Atende	Atende Parcialmente	Atende Totalmente
Score (0-1)	0	0,50	1

Churchill e Peter [11], bem como, Coelho et al. [12], sugerem que a confiabilidade de uma escala aumenta com a ampliação do número de categorias de resposta. No entanto, preferiu-se usar uma escala de tamanho mediano com três pontos. Dessa forma, manteve-se o equilíbrio na avaliação, pois uma escala menor poderia limitar as análises, bem como, gerar flutuações na normalidade dos dados [12] e uma escala com muitos itens aumentaria substancialmente o número de componentes a serem avaliados em virtude da complexidade e o elevado número de subquestões do nosso instrumento.

É importante ressaltar que a experiência, treinamento e habilidades dos pesquisadores envolvidos também tiveram influência na escolha da escala, pois avaliadores com maior capacidade e conhecimento permitem melhor mensurar a realidade sobre um objeto em estudo independente do número de pontos da escala [12].

De acordo com a escala utilizada (ver Tabela 4.1), os itens do instrumento de qualidade foram avaliados como **NÃO ATENDE**, quando não há nada no trabalho que atenda ao critério avaliado, **ATENDE PARCIALMENTE**, quando há evidências que atendam parcialmente ao requisito avaliado ou **ATENDE TOTALMENTE**, quando o trabalho apresenta atendimento total ao critério avaliado.

4.3 Índice global de qualidade

Na busca por encontrar mecanismos que ajudassem a definir um modelo ou guia que apontasse uma melhor forma de avaliar a qualidade, decidiu-se usar um índice global de qualidade, que é uma espécie de indicador quantitativo da qualidade de um estudo. Nesse caso, o estudo é resultante da soma dos pontos obtidos pela escala de qualidade utilizada.

O índice foi calculado pelo percentual dos valores obtidos da aplicação de uma escala tipo *Likert* (ver Tabela 4.1). O somatório da pontuação foi feito atribuindo a nota total do estudo em relação à qualidade e a classificação do índice foi realizada com base na categorização proposta por Beecham et al. [9], conforme mostra a Tabela 4.2.

Tabela 4.2 - Classificação da Qualidade segundo Beecham et al. [9]

Nota do Estudo (%)	Classificação da qualidade
N >= 86%	Excelente
66% =< N <= 85%	Muito Boa
46% =< N <= 65%	Boa
26% =< N <= 45%	Regular
N < 26%	Ruim

A nota total (NT) de classificação dos experimentos foi calculada conforme a fórmula do índice global de qualidade:

CLASSIFICAÇÃO = (NT / TOTAL POSSÍVEL) X 100 = N (%).

Conforme a Tabela 4.2, a classificação foi feita em cinco faixas (excelente, muito boa, boa, regular e ruim). Com isso, experimentos com uma pontuação global com valor de N menor que 26% de atendimento aos critérios, serão considerados de qualidade **RUIM** e experimentos classificados com valor de N maior ou igual a 86%, serão reconhecidos como estudos de **EXCELENTE** qualidade. Os escores individuais de todos os estudos avaliados estão disponíveis no Apêndice A.

4.4 Busca e seleção dos estudos

Segundo Kitchenham et al. [38], é necessário seguir uma estratégia rigorosa de busca para selecionar estudos primários que estejam relacionados com uma pergunta de pesquisa. Para isso, a estratégia pode ser realizar buscas automáticas e/ou manuais

nas principais bases de dados da área pesquisada. Neste trabalho, adotou-se um processo de busca manual, realizado a partir de todas as publicações dos três principais veículos internacionais (EASE, ESEM e ESEJ) da comunidade de engenharia de software empírica.

Esse processo de busca e seleção dos estudos beneficiou-se de um mapeamento sistemático que está em processo de submissão pelo grupo liderado pelo orientador deste trabalho, cujo objetivo foi identificar e analisar evidências sobre mecanismos utilizados no suporte à execução de estudos empíricos.

Foram considerados para a avaliação todos os artigos completos publicados nos três veículos pesquisados, desde 1997, primeiro ano de edição das fontes investigadas e vai até o ano de 2012, uma vez que a execução da avaliação de qualidade dos estudos foi realizada durante o ano de 2013 e, dessa forma, não tínhamos disponíveis todos os estudos das conferências do ano de 2013.

Dos 876 artigos das três fontes pesquisadas, encontrou-se 104 estudos que foram categorizados por seus autores como experimentos controlados, destes, 84 mencionaram explicitamente ter usado algum mecanismo de suporte ao planejamento ou execução do experimento, conforme observamos na Tabela 4.3.

Tabela 4.3 - Busca e seleção dos Experimentos Controlados

Fonte Avaliada	Número de Experimentos Controlados	Utilizaram Mecanismos de Suporte
EASE	14	09
ESEM	54	41
ESEJ	36	34
Total	**104**	**84**

Considerando as limitações de tempo e escopo da pesquisa, não foram incluídas estratégias de buscas automáticas. No entanto, os experimentos selecionados são representativos da Comunidade de ESE.

Durante a seleção dos estudos, foram descartados artigos resumidos, bem como, estudos que não foram categorizados por seus autores como experimentos controlados. Nos casos que houve estudos repetidos, foi considerado o mais completo. A lista com todos os estudos incluídos, bem como, todos os dados brutos coletados estão disponíveis nos Apêndices B e C.

4.5 Alocação dos papers aos avaliadores

Participaram diretamente desta avaliação de qualidade cinco pesquisadores. Além do autor participaram mais um estudante de mestrado e três estudantes de doutorado do Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE) e foram agrupados, por conveniência, em quatro duplas. Todas as duplas foram formadas pela junção do autor desse trabalho juntamente com um aluno de mestrado ou doutorado.

Os 104 estudos selecionados foram organizados numa planilha do *Microsoft Excel* de acordo com o ano de publicação das fontes pesquisadas. Em seguida, foi atribuído um número entre um e quatro para cada dupla e foi utilizada a função randômica do *Excel* para atribuir aleatoriamente cada dupla aos grupos de estudos selecionados. Embora limitada, a função que gera números aleatórios do *Microsoft Excel*, foi suficiente para garantir que a alocação dos avaliadores aos grupos de estudos não foi intencional.

É importante ressaltar que os avaliadores que participaram diretamente deste estudo não foram selecionados de forma aleatória, uma vez que são pesquisadores de um grupo de pesquisa com experiência em planejamento e execução de experimentos da área de engenharia de software empírica.

Cada dupla avaliou 26 *papers*, com isso, cada estudo foi revisado por no mínimo duas pessoas. A análise dos conflitos foi tratada em sessões de reunião de discussão e caso persistisse o desacordo, foi considerada a opinião de um terceiro pesquisador. Todo este processo foi realizado sob a supervisão do orientador deste trabalho de mestrado.

4.6 Coleta de dados e execução do estudo

O processo de realização de um EE envolve a coleta dos dados, a partir dos quais serão extraídos e interpretados os resultados experimentais. Segundo Kitchenham et al. [36], um dos objetivos das orientações de coleta de dados é garantir que esse processo seja suficiente para que a experiência possa ser replicada. Para isso, é preciso planejar e acompanhar a coleta, bem como, registrar quaisquer desvios no plano experimental.

Antes do início desta fase foi realizada uma reunião entre os avaliadores onde foram discutidos os procedimentos para coleta dos dados e execução do estudo. Em

seguida foi disponibilizado um guia (Apêndice D), de planejamento com um conjunto de instruções básicas para a realização do processo de extração dos dados.

É importante ressaltar que a planilha com o instrumento de avaliação de qualidade possuía uma coluna de observações para a indicação da localização exata da página, parágrafo ou seção do artigo, que motivou a escolha do avaliador por determinado valor da escala de qualidade em cada critério avaliado. Esse procedimento aumentou a confiabilidade nas respostas registradas no instrumento, bem como, facilitou as rodadas de discussões para consenso entre os avaliadores.

Após o treinamento e a realização do estudo piloto (descrito no Capítulo 3), todas as duplas receberam as planilhas com o instrumento de avaliação ao mesmo tempo. Inicialmente, o cronograma de execução do estudo previa que o trabalho fosse concluído em 15 dias, porém, devido à complexidade do instrumento e disponibilidade dos avaliadores, o preenchimento de todos os dados e discussão das divergências entre os avaliadores só foi concluído em aproximadamente 30 dias.

O processo de coleta dos dados foi acompanhado pelo autor desse trabalho de mestrado que, ao final, verificou se todas as planilhas estavam preenchidas de forma correta, bem como, observou se havia algum dado discrepante. Os dados brutos coletados estão disponíveis a partir dos Apêndices deste trabalho.

4.7 Resumo

Neste capítulo discute-se as etapas do planejamento e execução do estudo. Definiu-se o formato de cálculo do índice de qualidade utilizado, as estratégias utilizadas para busca e seleção dos estudos, bem como, a forma de alocação dos *papers* aos avaliadores de maneira que não comprometêssemos a imparcialidade do estudo. Define-se, também, o planejamento e acompanhamento da coleta dos dados e com isso, podemos debater no próximo capítulo a análise e discussão dos resultados.

Capítulo

5

5. Análise dos Dados e Discussão dos Resultados

Este capítulo descreve os procedimentos utilizados para aumentar a confiabilidade dos resultados, bem como, realizar a análise estatística e discussão. Após isso, serão apresentados os dados da estatística descritiva e da análise dos indicadores das variáveis observadas e será concluída, indicando as possíveis respostas às questões de pesquisa.

5.1 Índice de Confiabilidade entre os avaliadores

Conforme descrito no Capítulo 4, os 104 estudos selecionados foram divididos e distribuídos entre as quatro duplas de avaliação. Os critérios de avaliação foram aplicados de forma independente por cada avaliador.

Antes da discussão dos conflitos, foi medido o índice de concordância entre todos os avaliadores utilizando o coeficiente *Kappa*. O resultado do índice atingido por cada dupla é apresentado na **Tabela 5.1**. Para interpretar o coeficiente alcançado, consideramos a explicação sugerida por Landis JR e Koch GG [43], demonstrada pela **Tabela 3.2** do Capítulo 3, que apresenta uma classificação para os intervalos do coeficiente.

Tabela 5.1 - Índice de Concordância *Kappa*

Grupos	Índice *Kappa (K)*
Dupla K1	0,768
Dupla K2	0,444
Dupla K3	0,259
Dupla K4	0,722
Valor Médio	**0,548**

De acordo com os valores de concordância de cada dupla, um valor médio foi calculado, onde (K1 + K2 + K3 + K4) / 4 = **0,548**, tendo desse modo, de acordo com a classificação proposta por Landis e Koch [43], um nível de concordância **moderado**. Esse índice de concordância pode ter sido influenciado pela subjetividade dos critérios utilizados, bem como, pela complexidade do instrumento de avaliação e com isso, nos trouxe preocupação quanto ao nível de confiabilidade dos resultados desta etapa.

É importante ressaltar que antes e durante a realização do estudo principal, foram usadas algumas estratégias para melhorar a confiabilidade dos resultados. Dentre as quais, destacamos: *a)* Treinamento aos avaliadores, *b)* Estudo piloto com aplicação de índice *Kappa* e *c)* Indicação dos local específico na planilha de avaliação onde justificava a escolha da nota atribuída a cada critério por cada avaliador.

Após a verificação do índice de concordância entre os avaliadores, reuniu-se cada grupo para discutir e compor os resultados da dupla em uma única planilha, com isso, identificando e resolvendo as respostas divergentes.

5.2 Variáveis Compostas

No intuito de responder às questões de pesquisa relacionadas à replicação de um experimento, bem como, a importância estatística revelada pelos estudos, criaram-se as variáveis compostas, replicabilidade e rigor estatístico.

5.2.1 Replicabilidade dos estudos

A descoberta científica não é confiável a menos que possa ser replicada de forma independente [19]. Dessa forma, um dos cuidados mais importantes que o pesquisador deve ter ao planejar e executar um estudo experimental, é descrever de forma detalhada, as atividades que favoreçam a repetição do experimento.

A repetição dos estudos aumenta o conhecimento a respeito dos conceitos estudados, bem como, propicia maior confiabilidade às teorias verificadas. Entretanto, a repetição de um estudo experimental poderá acontecer caso os aspectos, como o delineamento experimental, a instrumentação e os resultados estejam adequadamente descritos [65].

Com isso, no intuito de verificar como está a qualidade dos experimentos controlados, quanto aos parâmetros que permitem a repetição dos estudos, criou-se uma variável composta chamada replicabilidade, a partir da soma das subquestões dos critérios do instrumento de qualidade que estão classificados e organizados nas seções de delineamento experimental, instrumentação, apresentação e interpretação dos resultados, conforme sugeridos por Travassos et al. [65]. Os critérios do instrumento que compõe a variável replicabilidade, foram: C1, C2, C4, C8, C16, C17, C18 e C19, conforme mostra a **Tabela 5.2**.

Tabela 5.2 - Critérios de qualidade da variável replicabilidade

N	Critérios
C1	**Os autores indicaram claramente os objetivos da pesquisa.** *Questões a considerar:* ▪ Existe uma razão por que o estudo foi realizado? ▪ Os autores indicaram de forma clara as questões de pesquisa do estudo? ▪ Os autores definiram as hipóteses do estudo e suas teorias subjacentes?
C2	**Existe uma descrição adequada do contexto em que a pesquisa foi**

	realizada. *Questões a considerar:* ▪ Há uma descrição do contexto em que o experimento foi realizado? ▪ Há uma descrição da natureza da organização (ES) em que o experimento foi realizado? ▪ Há uma descrição das habilidades e experiências dos participantes do experimento? (por exemplo, com uma linguagem, um método, uma ferramenta ou um domínio de aplicação). ▪ Há uma descrição do tipo de produto de software utilizado? (por exemplo, uma ferramenta de design, um compilador, etc.). ▪ Foi descrito o processo de software utilizado? (por exemplo, um processo padrão da empresa, os procedimentos de garantia de qualidade, o processo de gerenciamento de configuração).
C4	**Os autores descrevem o projeto experimental? O projeto de pesquisa foi adequado para resolver os objetivos da pesquisa.** *Questões a considerar:* ▪ Os autores descrevem claramente a metodologia escolhida? ▪ Os autores definem ou descrevem os tratamentos das variáveis? ▪ O pesquisador fez uma discussão em relação aos métodos utilizados na pesquisa?
C8	**Os autores descrevem os procedimentos de coleta de dados e definição de medidas.** *Questões a considerar:* ▪ Foram utilizados métodos de controle de qualidade para garantir a consistência, integridade e precisão dos dados coletados? ▪ Todas as medidas foram claramente definidas? (por exemplo, a escala, unidade ou regra de contagem). ▪ Foi utilizado método para coleta adequada dos dados? ▪ O pesquisador justifica os métodos que foram escolhidos? ▪ Há a definição clara de um instrumento utilizado para coleta de dados? ▪ O estudo informa o cronograma do experimento, bem como, tempo ou período utilizado para execução da coleta dos dados?

C16	**O estudo fornecer testes estatísticos apropriados e referenciados.** *Questões a considerar:* ▪ O estudo descreve e cita as referências para os procedimentos estatísticos utilizados? ▪ O estudo define critérios que sustentam o rigor estatístico do experimento?
C17	**As significâncias estatísticas são mencionadas com os resultados.** *Questões a considerar:* ▪ O estudo apresenta resultados baseados nos dados quantitativos, incluindo o tamanho do efeito e os limites de confiança?
C18	**O estudo tem valor para a pesquisa e prática? O estudo poderá ser utilizado por outros pesquisadores ou profissionais.** *Questões a considerar:* ▪ O pesquisador discutiu a contribuição do estudo ou interpreta os resultados em relação ao conhecimento ou entendimento já existente? ▪ A investigação identifica novas áreas em que pesquisas são necessárias? ▪ O pesquisador discutiu se e como os resultados podem ser transferidos para outras populações, ou considerou outras maneiras em que a pesquisa pode ser utilizada?
C19	**A pesquisa apresenta ou indica a disponibilidade dos dados brutos.** *Questões a considerar:* ▪ O estudo apresenta ou indica onde os dados brutos estão disponíveis para análise por outros revisores ou auditores independentes?

5.2.1.1 Índice de qualidade da variável replicabilidade

O índice de qualidade foi calculado pelo percentual dos valores obtidos a partir dos critérios que compõe a variável replicabilidade. Ao final foi feito o somatório da pontuação, atribuindo um índice, conforme a fórmula: Índice = (NT (Nota Total) / Total Possível) X 100 = N (%), e classificação conforme sugerido pela Tabela 4.2 de classificação, proposta por Beecham et al. [9], descrita na **Seção 4.3.**

Tabela 5.3 - Índice e classificação de qualidade da variável replicabilidade

Locais Avaliados	Índice	Classificação
EASE	56,57%	Boa
ESEM	60,33%	Boa
ESEJ	70,73%	Muito boa
Geral	63,42%	Boa

Conforme mostra a **Tabela 5.3** o índice de classificação geral de replicabilidade (63,42) entre os locais avaliados foi **BOM**, porém, esses valores são apenas descritivos e ainda serão discutidas as diferenças estatisticamente significativas entre os índices de qualidade dos locais avaliados.

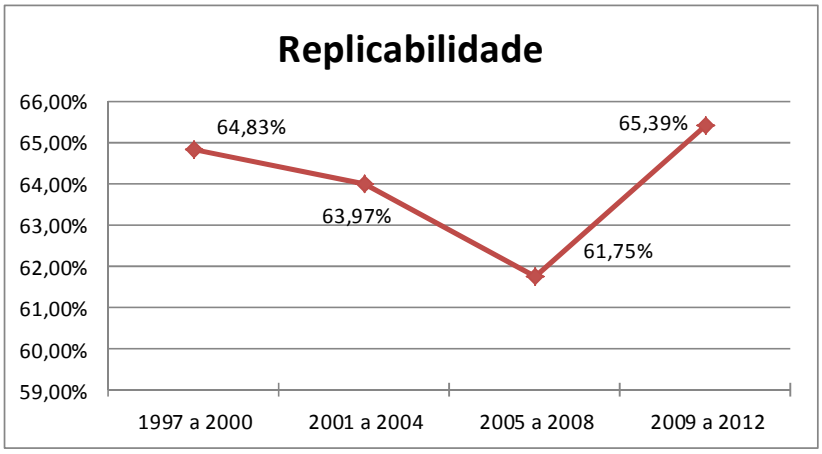


Figura 5.1 - Frequência da variável replicabilidade por período. Mostra a evolução do índice de replicabilidade no período avaliado (1997 a 2012).

O gráfico da **Figura 5.1** detalha a evolução do percentual dos valores obtidos quanto aos parâmetros que permitem a repetição dos estudos em quatro períodos: i) 1997 a 2000, ii) 2001 a 2004, iii) 2005 a 2008 e iv) 2009 a 2012. Observa-se que não houve alterações expressivas no índice ao longo dos tempos observados, que variaram entre 60% e 66%, porém, esses dados ainda não permitem afirmar que houve diferença estatisticamente significativa no período avaliado.

5.2.2 Rigor Estatístico dos Estudos

O rigor estatístico de um estudo empírico é essencial para a interpretação dos resultados e para a validade das conclusões. Um teste sem um adequado rigor estatístico não é capaz de fornecer informação suficiente para tirar conclusões sobre a aceitação ou rejeição das hipóteses da pesquisa [21].

Além disso, se o rigor estatístico é insuficiente, a probabilidade de encontrar resultados com significância estatística é pequena, e os resultados do estudo provavelmente serão insignificantes para a ciência. Assim, a incapacidade de fornecer um nível adequado do rigor estatístico, têm implicações tanto para a execução, como para a interpretação e apresentação dos resultados do experimento.

No intuito de verificar como está a qualidade dos experimentos controlados quanto ao nível estatístico foi criado uma variável composta chamada Rigor Estatístico, a partir da soma das subquestões dos critérios do instrumento de qualidade que abordam questões estatísticas quanto à coleta, análise e significância estatística dos dados, bem como, quanto à apresentação e interpretação dos resultados. Os critérios do instrumento que compõe a variável rigor estatístico foram: C12, C14, C16 e C17, conforme mostra a **Tabela 5.4**.

Tabela 5.4 - Critérios de qualidade da variável rigor estatístico

N	Critérios
C12	**A análise de dados foi suficientemente rigorosa.** *Questões a considerar:* ▪ Os autores descreveram de forma detalhada os procedimentos para análise dos dados, bem como, justificaram suas escolhas? ▪ Os autores forneceram referências às descrições dos procedimentos de análise de dados? ▪ Os autores mencionam níveis de significância e tamanhos de efeito? ▪ Os outliers são mencionados e casos encontrados são tratados durante a análise? ▪ Foi utilizada alguma técnica ou forma para análise dos dados? (Por ex: análise temática, grounded theory, etc.). ▪ Foram apresentados dados suficientes para apoiar os resultados? ▪ Os dados contraditórios foram levados em consideração? ▪ Foram utilizados métodos de controle de qualidade para verificar os resultados?

C14	**Os autores declararam de forma clara os resultados do estudo.** *Questões a considerar:* ▪ Os autores apresentam conclusões de forma clara? ▪ As conclusões são justificadas pelos resultados e as conexões entre os resultados e conclusões são apresentadas com clareza? ▪ Os autores discutem suas conclusões em relação às questões de investigação originais?
C16	**O estudo forneceu testes estatísticos apropriados e referenciados.** *Questões a considerar:* ▪ O estudo descreve e cita as referências para os procedimentos estatísticos utilizados? ▪ O estudo define critérios que sustentam o rigor estatístico do experimento?
C17	**As significâncias estatísticas são mencionadas com os resultados.** *Questões a considerar:* ▪ O estudo apresenta resultados baseados nos dados quantitativos, incluindo o tamanho do efeito e os limites de confiança?

5.2.2.1 Índice de qualidade da variável rigor estatístico

O índice ou média de qualidade foi calculado pelo percentual dos valores obtidos a partir da soma dos critérios que compõe a variável rigor estatístico. Ao final foi feito o somatório da pontuação, atribuindo um índice, conforme a fórmula: Índice = (NT (Nota Total) / Total Possível)X100 = N (%), e classificação conforme sugerido pela Tabela 4.2 de classificação, proposta por Beecham et al. [9], descrita na **Seção 4.3**.

Tabela 5.5 - Índice e classificação de qualidade da variável rigor estatístico

Locais Avaliados	Índice	Classificação
EASE	60,93%	Boa
ESEM	67,70%	Muito Boa
ESEJ	79,25%	Muito Boa
Geral	70,79%	Muito Boa

Conforme mostra a **Tabela 5.5,** o índice geral da variável rigor estatístico (70,79%), foi classificado como **Muito Bom**, porém, serão apresentados na próxima seção os resultados quanto à diferença estatística entre os locais avaliados.

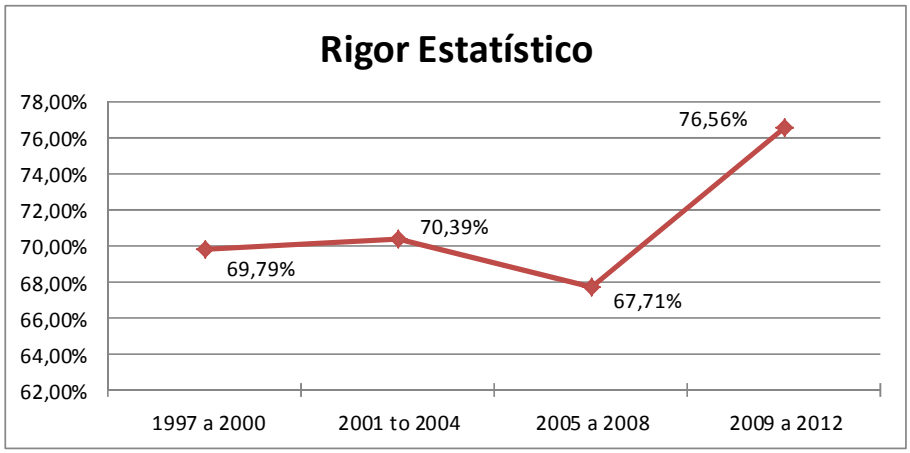


Figura 5.2 - Frequência da variável rigor estatístico por período. Mostra a evolução do índice de rigor estatístico dos estudos ao longo dos quatro períodos avaliados.

O gráfico da **Figura 5.2** mostra a evolução do percentual dos valores obtidos quanto à importância estatística revelada pelos estudos em quatro períodos: i) 1997 a 2000, ii) 2001 a 2004, iii) 2005 a 2008 e iv) 2009 a 2012.

Observa-se que nos três iniciais tempos avaliados não houve expressivas alterações no índice, porém, percebe-se que no período de 2009 a 2012 (76,56%), há uma tendência de aumento do rigor estatístico dos estudos em relação aos períodos anteriores. No entanto, esses dados são apenas descritivos e, por isso, não permitem afirmar, ainda, que houve diferença estatisticamente significativa entre os períodos.

5.3 Estatística Descritiva

A análise estatística foi realizada utilizando o pacote estatístico IBM SPSS (versão 2.0), e uma significância estatística de $\alpha = 0,05$. O SPSS foi desenvolvido para ser usado em diversas áreas da ciência e permite realizar uma variedade de análises (por exemplo: análises descritivas, análises inferenciais, multivariadas, gráficos, entre outros).

Uma das primeiras ações a serem feitas com dados estatísticos é sua descrição. Dessa forma esta seção apresenta as distribuições de frequência das variáveis: locais avaliados, mecanismos de suporte e período avaliado.

Tabela 5.6 - Frequências dos Experimentos Controlados

Período	Local Avaliado			Total
	EASE	ESEM	ESEJ	
1997 a 2000	02	00	07	**09** (*8,65%)*
2001 a 2004	03	10	07	**20** (*19,23%)*
2005 a 2008	05	28	14	**47** (*45,19%)*
2009 a 2012	04	16	08	**28** (*26,92%)*
Total	**14**	**54**	**36**	**104**
Por cento	*13,5%*	*51,9%*	*34,6%*	*100%*

A **Tabela 5.6** apresenta a descrição dos resultados da distribuição de frequência dos experimentos controlados no período de 1997 a 2012 nos locais avaliados (EASE, ESEM e ESEJ). Observa-se que dos 104 *papers* encontrados 13,5% foram publicados no EASE, 51,9% no ESEM e 34,6% no ESEJ. É importante verificar que 45,19% de todos os experimentos encontrados foram publicados no período entre 2005 a 2008.

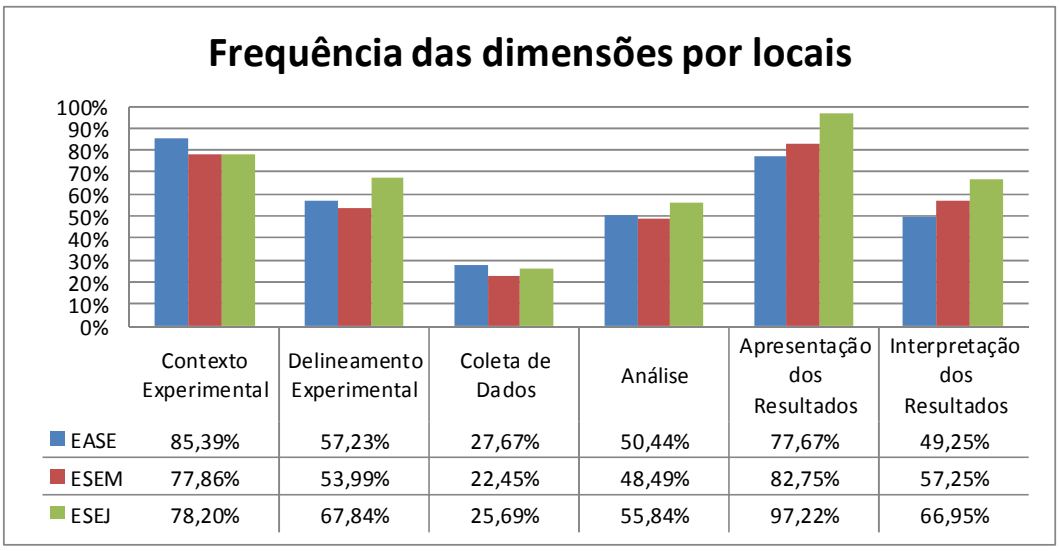


Figura 5.3 - Frequência das dimensões por locais avaliados. Mostra a diferença média entre os níveis de qualidade dos locais avaliados (conferências e Journal) por cada uma das dimensões do instrumento de avaliação de qualidade.

A **Figura 5.3** mostra a distribuição de frequência das dimensões em relação aos locais avaliados. Nota-se que em relação às dimensões Contexto Experimental (85,39%) e Coleta dos dados (27,67%), o evento EASE apresenta o maior nível de qualidade entre os eventos, no entanto, o *Journal* (ESEJ) mostra o melhor nível de qualidade em relação à média do conjunto de dimensões avaliadas. É importante

ressaltar neste momento que a descrição desses dados não indica existência de diferença estatisticamente significativa entre os grupos avaliados.

Tabela 5.7 - Índice de Qualidade por critério avaliado

Critérios	Médias
1. Os autores indicaram claramente os objetivos da pesquisa?	*79,48%*
2. Existe uma descrição adequada do contexto em que a pesquisa foi realizada?	*70,00%*
3. Houve a descrição de pesquisas similares e como estas se relacionam com o estudo?	*87,50%*
4. Os autores descrevem o projeto experimental? O projeto de pesquisa foi adequado para resolver os objetivos da pesquisa?	*62,82%*
5. A estratégia de recrutamento foi adequada aos objetivos da pesquisa?	*42,11%*
6. Os autores indicam e descrevem de forma clara as variáveis da pesquisa?	*69,23%*
7. O pesquisador define o processo aplicado ao tratamento de objetos e sujeitos?	*62,74%*
8. Os autores descrevem os procedimentos de coleta de dados e definição de medidas?	*56,97%*
9. Os autores descreveram ou apresentaram alguma medida de concordância entre os avaliadores?	*9,61%*
10. A relação entre o pesquisador e os participantes foi considerada adequada?	*16,58%*
11. As perdas e desistências de participantes ocorridas desde a seleção até o final do estudo foram descritas?	*13,94%*
12. A análise de dados foi suficientemente rigorosa?	*51,80%*
13. Os autores discutiram o potencial viés do experimento quanto à análise dos dados?	*50,80%*
14. Os autores declararam de forma clara os resultados do estudo?	*81,37%*
15. São mencionadas as ameaças à validade e como essas ameaças afetam os resultados e conclusões?	*92,78%*
16. O estudo fornecer testes estatísticos apropriados e referenciados?	*62,98%*
17. As significâncias estatísticas são mencionadas com os resultados?	*87,01%*

18. O estudo tem valor para a pesquisa e prática? O estudo poderá ser utilizado por outros pesquisadores ou profissionais?	*71,79%*
19. A pesquisa apresenta ou indica a disponibilidade dos dados brutos?	*16,34%*

A **Tabela 5.7** apresenta o índice de qualidade de todos os artigos por critério avaliado. O cálculo foi obtido pela soma das notas das subquestões de cada critério, divididas pelo total possível de notas e multiplicado por 100, conforme a fórmula a seguir: Média do Índice de qualidade = (NT / Total Possível)X100 = N (%).

É importante observar que os critérios melhores avaliados em termos de qualidade foram: C15 (92,78%), C3 (87,50%) e C17 (87,01%), que estão relacionados respectivamente com:

a) A descrição das ameaças à validade e como elas afetam os resultados e conclusões;
b) A definição de pesquisas similares e como se relacionam com o estudo e;
c) Apresentação das significâncias estatísticas do estudo.

No entanto, os critérios C9 (9,61%), C11 (13,94%) e C19 (16,34%) são os piores índices atribuídos pelos avaliadores. Estes estão relacionados respectivamente com:

a) Descrição de medidas de concordância entre os avaliadores;
b) Descrição de ocorrências e desistências ocorridas após a seleção dos participantes e;
c) Apresentação ou indicação dos dados brutos dos experimentos.

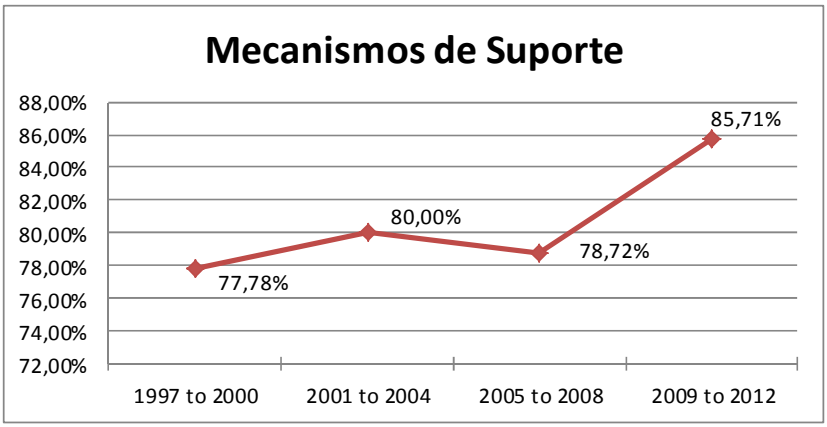


Figura 5.4 - Frequência da variável mecanismo de suporte por período avaliado. Mostra a evolução do uso dos mecanismos de suporte ao longo do período entre 1997 a 2012.

Em relação ao uso de mecanismos de suporte a **Figura 5.4** mostra que no período de 1997 a 2000 a porcentagem de estudos que indicaram usar algum mecanismo de suporte ao experimento era 77,78%, no período de 2001 a 2004 aumentou essa incidência para 80%, porém, percebe-se que de 2005 a 2008 (78,72%) houve uma diminuição e uma tendência de retomada é observada no último período avaliado que foi de 2009 a 2012 (85,71%).

5.4 Análise dos Indicadores

Para realizar os procedimentos de análises dos indicadores, foram criadas as seguintes variáveis compostas: replicabilidade e rigor estatístico dos estudos. O conjunto de critérios que formaram essas variáveis foi explicado em detalhes através dos itens 5.2.1 e 5.2.2 desse estudo.

Tabela 5.8 - Análise dos índices das variáveis

Variáveis	Locais Avaliados			Média Geral
	EASE	ESEM	ESEJ	
Índice Global de Qualidade	*55,21%*	*54,25%*	*62,25%*	*57,15%*
Replicabilidade	*56,57%*	*60,33%*	*70,73%*	*63,42%*
Rigor Estatístico	*60,93%*	*67,70%*	*79,25%*	*70,79%*
Mecanismo de Suporte	*64,30%*	*75,90%*	*94,40%*	*80,76%*

A **Tabela 5.8** mostra a média geral de avaliação da qualidade entre as conferências e o *Journal* pesquisado, bem como, os índices de avaliação das variáveis: replicabilidade, rigor estatístico e mecanismos de suporte.

Observa-se que o ESEJ (62,25%), apresenta o melhor índice de qualidade entre os locais pesquisados, porém, a média geral do índice global de qualidade entre todos os locais pesquisados foi 57,15%, o que segundo a tabela de classificação sugerida por Beecham et al. [9], foi **BOA**. A variável replicabilidade (63,42%), também, foi classificada como **BOA** e a variável rigor estatístico (70,79%), segundo a tabela de classificação, foi classificada como **MUITO BOA**.

Em relação ao uso de mecanismos de suporte utilizados para guiar o planejamento e/ou execução dos experimentos controlados avaliados, 80,76% dos estudos relataram explicitamente que usaram algum tipo de mecanismo de suporte.

5.4.1 Distribuição do índice global

Para compreender quais os testes que seriam necessários para realizar o cálculo das diferenças de média, verificamos se a distribuição do índice global e das variáveis compostas era normal. Para isso, realizou-se o teste K-S (*Kolmogorov-Smirnov*), um dos mais antigos e aceitos testes para verificar a gaussianidade dos dados [46].

Tabela 5.9 - Teste *One-Sample Kolmogorov-Smirnov*

Testes	Índice Global	Rigor Estatístico	Replicabilidade
Kolmogorov-Smirnov Z	0,776	1,180	0,563
Valor de *p.*	0,584	0,123	0,909

A **Tabela 5.9** exibe os resultados para o teste K-S, e mostra que a distribuição da amostra é normal, visto que o valor de *p*. para todas as variáveis analisadas é superior a 0,05.

Seguindo esses pressupostos de gaussianidade (valor de p= 0,05), decidiu-se usar os testes paramétricos, Teste de Hipótese (**Teste T**) e análise de variância (**ANOVA**), para verificar se houve diferenças estatisticamente significativas entre as variáveis analisadas.

As próximas seções descrevem os procedimentos realizados para calcular a diferença de média entre o índice global de qualidade, período avaliado e as variáveis compostas replicabilidade e rigor estatístico.

5.4.2 Diferença de Média entre o índice global de qualidade e o período avaliado

No intuito de calcular a diferença de média entre o índice e o período avaliado, foi recodificada a variável Ano em quatro grupos: i) 1997 a 2000, ii) 2001 a 2004, iii) 2005 a 2008 e iv) 2009 a 2012. Para isso foi usado uma análise de variância (ANOVA ONE WAY).

Tabela 5.10 - ANOVA – Diferença de média por período

Períodos Avaliados	N	Subconjunto para alfa = 0,05
		Índice
1997 a 2000	09	59,0107
2001 a 2004	20	58,1436
2005 a 2008	47	56,2593
2009 a 2012	28	57,3496
Valor de *p*.		0,819

A **Tabela 5.10** acima mostra que no período de 1997 a 2000, o índice de qualidade apresentado entre os locais avaliados foi 59,01 e com o passar dos anos esse índice foi diminuindo até o período de 2005 a 2008, quando atingiu 56,25. Observa-se que a partir do período 2009 a 2012, o índice de qualidade atinge 57,34, indicando uma provável tendência de aumento de qualidade.

No entanto, percebe-se que a análise de variância (ANOVA), indica que não houve diferença estatisticamente significativa (valor de *p*. 0,819) entre os períodos, pois a significância estatística para todos os grupos foi maior que 0,05, o que significa que ao longo dos anos, os experimentos controlados, têm mantido o mesmo padrão de qualidade, ou seja, um índice global de qualidade em torno de 57,15, classificado segundo a tabela proposta por Beecham et al. [9], como **BOA**.

5.4.3 Diferença de Média entre o índice global de qualidade e os locais avaliados

Para calcular a diferença de média entre o índice e os locais avaliados, utilizou-se uma análise de variância (ANOVA ONE WAY).

Ao observarmos os resultados do teste estatístico apresentado pela **Tabela 5.11** abaixo, verifica-se que há diferenças estatisticamente significativas, entre, ao menos, dois grupos, pois o valor de *p*. foi menor que 0,05 (significância estatística). Entretanto, o procedimento estatístico ANOVA, não indica quais grupos possuem diferenças em suas médias. Por isso, será preciso examinar a matriz gerada pelo Teste de *Tukey* (Post Hoc). O valor de *p*. igual a 0,000 é em decorrência do software estatístico utilizado não mostrar todas as casas decimais após a vírgula.

Tabela 5.11 - ANOVA – Diferença de Média por locais avaliados

	Soma dos quadrados	DF	Quadrado Médio	Valor de *p*.
Between Groups	1440,556	2	720,278	0,000
Within Groups	8270,445	101	81,886	
Total	9711,002	103		

O teste *Post Hoc* é utilizado para identificar quais os grupos ou variáveis cujas médias diferem ou não entre si, sendo um complemento da análise de variância (ANOVA).

De acordo com a **Tabela 5.12** abaixo, percebe-se que há diferença estatisticamente significativa apenas quando comparamos o *Journal* (ESEJ) com os outros dois locais (EASE e ESEM). Ao relacionarmos os eventos EASE (55,21) e ESEM (54,25), percebe-se que não há nenhuma diferença significativa de qualidade. O

valor de *p*. igual a 0,000 é em decorrência do software estatístico utilizado não mostrar todas as casas decimais após a vírgula.

Tabela 5.12 - *Tukey Test* – Diferença de Média por locais avaliados

Comparações Múltiplas			
(I) Média dos Locais Avaliados	(J) Locais Avaliados	Diferença de Média (I-J)	Valor de *p*.
EASE (*55,21%*)	ESEM	0,95853	0,934
	ESEJ	-7,03408	0,040
ESEM (*54,25%*)	EASE	-,95853	0,934
	ESEJ	-7,99261	0,000
ESEJ (*57,15%*)	EASE	7,03408	0,040
	ESEM	7,99261	0,000

5.4.4 Diferença de Média entre o índice de qualidade e mecanismos de suporte

A **Tabela 5.13** abaixo mostra que os experimentos que mencionaram explicitamente o uso de algum tipo de mecanismo de suporte, tiveram um índice de qualidade igual a 58,54%, e os experimentos que não declararam usar mecanismos de suporte, tiveram um índice de qualidade igual a 51,32%.

Tabela 5.13 - Teste T – Índice de qualidade por mecanismos de suporte

Mecanismos de Suporte		Média
NÃO		51,3289
SIM		58,5401
Teste de Levene para igualdade de variâncias		Teste-T para igualdade de médias
F	**Valor de *p*.**	**Valor de *p*.**
0,363	0,548	0,002

Para calcular a diferença da média entre o índice e a variável mecanismo de suporte, utilizou-se o Teste T com amostras independentes. O teste indicou que o valor de *p*. (significância estatística) foi de 0,002, portanto, menor que 0,05. Dessa forma, houve uma diferença estatisticamente significativa entre os grupos, o que significa que os estudos que mencionaram explicitamente o uso de mecanismos de suporte apresentaram uma melhora significativa na qualidade do experimento.

5.4.5 Diferença de Média entre o índice de Replicabilidade e os locais avaliados

Para calcular a diferença de média entre o índice de qualidade dos estudos e a variável replicabilidade, utilizou-se a análise de variância Anova, conforme mostra a **Tabela 5.14** abaixo.

Tabela 5.14 - ANOVA – Diferença de média por replicabilidade

	Soma dos quadrdos	***DF***	***Quadrado Médio***	Valor de *p*.
Between Groups	3096,802	2	1548,401	0,000
Within Groups	13828,187	101	136,913	
Total	16924,990	103		

De acordo com os resultados do teste estatístico, há diferenças estatisticamente significativas, entre, ao menos, dois grupos, pois o valor de *p*. foi menor que 0,05 (significância estatística). Entretanto, este procedimento não indicou quais grupos possuem diferenças em suas médias, logo será preciso examinar a matriz gerada pelo teste de *Tukey* (*Post Hoc*). O valor de *p*. igual a 0,000 é em decorrência do software estatístico utilizado não mostrar todas as casas decimais após a vírgula.

Tabela 5.15 - *Tukey Test* – Diferença de média por replicabilidade

Comparações Múltiplas			
(I) Locais por comparação	(J) Locais Avaliados	Diferença de Média (I-J)	Sig.
EASE	ESEM	-3,75441	0,535
	ESEJ	-14,15757	0,001
ESEM	EASE	3,75441	0,535
	ESEJ	-10,40316	0,000
ESEJ	EASE	14,15757	0,001
	ESEM	10,40316	0,000

De acordo com a **Tabela 5.15** acima, percebe-se que há diferença estatisticamente significativa apenas quando comparamos o *Journal* (ESEJ) com os outros dois eventos (EASE e ESEM). Ao relacionarmos os eventos EASE (56,57%) e ESEM (60,33%) nota-se que não há nenhuma diferença significativa no índice de qualidade. Portanto, pode-se afirmar que entre todos os locais avaliados, o *Journal* (ESEJ) possui melhores características que favorecem a replicabilidade dos experimentos controlados.

5.4.6 Diferença de Média entre a variável rigor estatístico e os locais avaliados

Para calcular a diferença da média entre o índice de qualidade dos estudos e a variável rigor estatístico, utilizou-se a análise de variância Anova, conforme mostra a **Tabela 5.16** abaixo.

Tabela 5.16 - ANOVA – Diferença de média por rigor estatístico

	Soma dos quadrados	***DF***	***Quadrado Médio***	Valor de *p*.
Between Groups	4450,517	2	2225,258	0,000
Within Groups	25466,851	101	252,147	
Total	29917,368	103		

A **Tabela 5.16** acima mostra que o teste estatístico realizado comprova diferenças estatisticamente significativas, entre, ao menos, dois grupos, pois o valor de p (Sig.) foi menor que 0,05 (significância estatística). Entretanto para verificar quais grupos possuem diferenças em suas médias, examinar-se-á o Teste de *Tukey* (*Post Hoc*). O valor de *p*. igual a 0,000 é em decorrência do software estatístico utilizado não mostrar todas as casas decimais após a vírgula.

Tabela 5.17 - *Tukey Test* – Diferença de média por rigor estatístico

Comparações Múltiplas			
(I) Locais por Comparação	(J) Locais Avaliados	Diferença de Média (I-J)	Sig.
EASE	ESEM	-6,77083	0,334
	ESEJ	-18,31597	0,001
ESEM	EASE	6,77083	0,334
	ESEJ	-11,54514	0,003
ESEJ	EASE	18,31597	0,001
	ESEM	11,54514	0,003

De acordo com a **Tabela 5.17** acima, percebe-se que há diferença estatisticamente significativa quando comparamos o *Journal* (ESEJ) com os outros dois locais (EASE e ESEM). Ao relacionarmos o EASE (60,93%) e o ESEM (67,70%) percebe-se que não houve diferença estatisticamente significativa no índice de qualidade.

5.5 Resposta às questões de pesquisa

Esta seção apresenta as respostas às questões de investigação e menciona algumas conclusões com base em tendências observadas:

5.5.1 RQ1. Qual é a evolução da qualidade dos estudos em relação aos mecanismos de suporte utilizados?

Com base nos resultados, pode-se dizer que 81% dos experimentos controlados avaliados mencionaram explicitamente ter utilizado algum tipo de mecanismo de suporte ao planejamento ou execução dos estudos. Avaliando os períodos, observamos que de 1997 a 2008, 78% dos estudos usaram mecanismos de suporte, já no período de 2009 a 2012, há uma tendência de aumento com 85,71%. No entanto, não há diferença de média estatisticamente significativa entre os períodos avaliados.

É importante ressaltar que os experimentos que mencionaram ter usado algum mecanismo de suporte tiveram um índice de qualidade igual a 58,54, enquanto os experimentos que não usaram nenhum tipo de mecanismo tiveram um índice de qualidade igual a 51,32. Essa diferença foi estatisticamente significativa e, portanto, quando há o uso de mecanismos de suporte houve uma evolução significativa na classificação de qualidade do experimento.

5.5.2 RQ2. Qual é a evolução da qualidade dos estudos em relação à replicabilidade?

Com base nos resultados, a classificação geral desta variável entre os locais avaliados foi **BOA** e o índice global de qualidade obteve 63,42 pontos. O bom resultado da classificação geral da variável replicabilidade deve ser analisado com cautela, uma vez que estudos experimentais raramente são replicados e geralmente não relatam seus resultados de tal forma que permita a comparação dos resultados [6]. Por isso, não podemos afirmar que, de fato, os critérios selecionados para esta variável estão avaliando o que se propuseram avaliar e novos estudos devem avaliar a qualidade de critérios com uma maior severidade ao conceito de replicabilidade.

Apesar de a variável replicabilidade obter uma boa classificação, os procedimentos realizados para calcular a diferença de média entre o índice global de qualidade e os períodos avaliados identificaram que não houve evolução estatisticamente significativa na qualidade ao longo do período avaliado (1997 a 2012).

5.5.3 RQ3. Qual é a evolução da qualidade dos estudos em relação ao rigor estatístico?

Os resultados mostraram que a variável rigor estatístico, obteve um índice de qualidade de 70,79 pontos. Avaliando os períodos, observou-se que de 1997 a 2008 não houve uma diferença significativa na média do índice de qualidade entre os locais

pesquisados, porém, percebe-se que no período de 2009 a 2012 (76,56%), há uma tendência de aumento no rigor estatístico dos estudos em relação aos períodos anteriores. Novos estudos devem ser realizados no intuito de confirmar ou não essa tendência de aumento no rigor estatístico dos estudos, bem como, novas metodologias de avaliação devem abordar uma maior severidade a este conceito.

Apesar da classificação geral de qualidade dessa variável ser **MUITO BOA** (70,79), bem como, apresentar uma tendência de aumento da qualidade no período de 2009 a 2012, os procedimentos realizados para calcular a diferença de média entre o índice global de qualidade, os locais e períodos avaliados identificaram que: i) quanto à diferença de média entre o índice da variável rigor estatístico e os locais avaliados, houve evolução estatisticamente significativa apenas quando comparamos o *Journal* (ESEJ) com os outros dois locais (EASE e ESEM), ao relacionarmos os eventos EASE (55,21) e ESEM (54,25), não há nenhum progresso significativo na qualidade. ii) quanto à diferença de média entre o índice e os períodos avaliados não houve evolução significativa da qualidade desta variável entre os quatro períodos avaliados, o que significa que, ao longo dos anos, o índice global do rigor estatístico têm mantido o mesmo padrão.

5.5.4 RQ4. Qual é a evolução da qualidade dos estudos no período avaliado?

Os resultados mostraram que o índice global de qualidade dos estudos foi 57,15 e com isso, segundo a tabela de classificação sugerida por Beecham et al. [9], foi **BOA**.

Avaliando os períodos, observou-se que de 1997 a 2000, o índice de qualidade apresentado foi 59,01 e com o passar dos anos o índice foi diminuindo até o período de 2005 a 2008 quando atingiu 56,25. Observa-se, ainda, que a partir do período 2009 a 2012 o índice de qualidade atinge 57,34%, indicando uma provável tendência de aumento de qualidade. Porém, não houve diferença estatisticamente significativa que justifique essa tendência, o que significa que não há um aumento de qualidade ao longo do período avaliado.

5.5.5 RQ5. Qual é a evolução da qualidade dos estudos por veículo de divulgação avaliado?

Em relação à análise de qualidade dos locais avaliados, os resultados revelaram que houve evolução estatisticamente significativa apenas quando comparamos o *Journal* (ESEJ) com os outros dois locais pesquisados (EASE e ESEM). Ao relacionar o EASE (55,21%) e o ESEM (54,25%), verifica-se que não houve nenhuma evolução

estatisticamente significativa no índice de qualidade dos estudos. Portanto, conforme previámos o *Journal* (ESEJ), apresenta maior nível de qualidade dos experimentos categorizados como controlados entre os locais avaliados.

5.6 Resumo

Neste capítulo discutiu-se as etapas de análise e discussão dos resultados. Para isso, iniciamos descrevendo os procedimentos utilizados para aumentar a confiabilidade dos resultados e em seguida apresentou-se a descrição dos dados e análise dos indicadores das variáveis observadas. O próximo capítulo irá descrever as possíveis ameaças à validade, a indicação de trabalhos futuros, bem como, conclusões do estudo.

Capítulo

6

6. Considerações Finais

Neste capítulo serão apresentadas as considerações finais. Entre elas são discutidas as ameaças à validade do estudo, recomendações para trabalhos futuros e as conclusões obtidas com a pesquisa.

6.1 Ameaças à validade

Nesta seção, apresentamos as ameaças à validade do estudo apresentado nesta dissertação. De acordo com a taxonomia proposta por Wohlin et al. [67], as limitações desse estudo, foram categorizadas em interna, externa, de conclusão e de constructo.

6.1.1 Ameaças à validade interna

A validade interna, está relacionada à capacidade de repetir o comportamento atual de um estudo (planejamento e execução) com os mesmos participantes e objetos para o qual foi executado. Este aspecto de validade é relevante para a avaliação da qualidade dos estudos primários [20]. Os seguintes aspectos da validade interna foram identificados:

a) **Confiabilidade do Instrumento de Qualidade** – Uma das principais ameaças aos resultados desse estudo é a confiabilidade do instrumento de avaliação da qualidade entre os avaliadores. Na tentativa de minimizar essa limitação, algumas estratégias foram implementadas como: i) Treinamento aos avaliadores e ii) Estudo piloto com uma amostra aleatória dos artigos no intuito de verificar o grau de confiabilidade do instrumento entre os avaliadores. Os resultados do estudo piloto indicaram que o instrumento avaliado apresentou uma confiabilidade substancial (seção 3.3, página 50);

b) **Subjetividade do Instrumento de Qualidade** – A lista de critérios utilizada foi revista e as questões não relevantes para avaliação de experimentos formais foram removidos. Algumas subquestões foram reescritas, pois verificavam se os estudos descreviam em detalhes os procedimentos, com isso, aumentou a objetividade dos critérios. No entanto, introduziu-se o risco de que um autor pode ter descrito em detalhes um método ou procedimento que pode ser completamente inadequado para os propósitos do estudo, mas que foi bem avaliado pelo instrumento de qualidade;

c) **Experimentos que realizaram mais de um tipo de estudo** – Um provável viés pode ter sido introduzido quanto aos artigos que realizaram mais de um tipo de estudo (por exemplo: experimento e estudo de caso ou experimento e survey, etc.) e por falta de espaço no documento, eles não foram capazes de descrever em detalhes os procedimentos do experimento, foco da pesquisa, etc. Contudo, apenas quatro artigos realizaram mais de um tipo de estudo e com isso, não houve alteração significativa nos resultados;

d) **Complexidade do Instrumento de Qualidade** – Apesar dos avaliadores envolvidos nesta pesquisa terem experiência com planejamento e execução de estudos empíricos, foi realizado um treinamento sobre avaliação de qualidade e 02 horas de rodadas de discussão para assegurar o entendimento dos critérios avaliados. Porém, em razão da complexidade do instrumento o treinamento e o tempo gastos com as discussões podem não ter sido suficientes.

6.1.2 Ameaças à validade externa

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. Uma ameaça à validade externa identificada e que pode limitar a capacidade de aplicar qualquer generalização está relacionada aos materiais usados.

O principal material usado neste estudo foi um instrumento de avaliação de qualidade desenvolvido a partir de listas de verificação genéricas e específicas, bem como, por intervenções do autor desta pesquisa. Como os critérios da lista sugerida por Dyba et al. [20], são genéricos e não são especificamente destinados a experimentos, foi utilizado, também, critérios definidos por Dieste et al. [18] e Kitchenham et al. [37], uma vez que essa lista apresenta recomendações de garantia de qualidade experimentais.

Mais estudos empíricos devem ser realizados no sentido de verificar a validade do instrumento utilizado em amostras maiores, assim como a relação entre os critérios utilizados e recomendações de garantia de qualidade experimentais mencionadas por Kitchenham et al. [36].

6.1.3 Ameaças à validade de conclusão

A validade de conclusão está relacionada à relação entre os tratamentos usados no estudo e os resultados obtidos. No intuito de aumentar a validade de conclusão dos resultados foram utilizados os seguintes procedimentos e decisões:

a) Uso de um pacote estatístico (IBM SPSS), que permitiu realizar de forma eficiente uma variedade de análises estatísticas descritivas e inferenciais;
b) Uso de um valor-p com nível de significância estatística = 0,05, o que aumentou a probabilidade dos nossos resultados não terem sido obtidos de um “acaso”;

c) Uso do *Cohen's Kappa*, um coeficiente estatístico para escalas nominal utilizado para verificar a confiabilidade entre os avaliadores do instrumento de avaliação de qualidade;
d) Uso do teste K-S (*Kolmogorov-Smirnov*), um dos mais antigos e aceitos testes para verificar a gaussianidade (normalidade) dos dados;
e) Uso do *Teste T, Post Hoc* e Análise de variância *Anova One Way* para calcular a probabilidade da diferença entre duas ou mais médias não ter sido obtida de um "acaso".

6.1.4 Ameaças à validade de constructo

A validade do constructo está relacionada ao grau em que as variáveis usadas no estudo conseguem medir com precisão os conceitos estudados. É importante ressaltar que não existe um conceito absoluto de "qualidade" quanto à avaliação de estudos empíricos e na prática não existe um conjunto padrão de critérios para avaliação da qualidade.

Para limitar a validade do constructo este estudo seguiu as recomendações de autores [27][41][48] que sugerem que o conceito de qualidade deve estar relacionado ao viés de pesquisa, ao aumento da validade interna e externa e a interpretação dos resultados.

No entanto, não se pode afirmar que os critérios selecionados estão avaliando o que se propuseram a avaliar e estudos adicionais são necessários para investigar outras dimensões da avaliação de qualidade no âmbito de estudos experimentais em ES, bem como, as correlações entre os critérios, aspectos e dimensões utilizadas.

6.2 Recomendações para Trabalhos Futuros

A partir da condução desta pesquisa, propõem-se algumas oportunidades de trabalhos futuros, bem como, direcionamentos para novas pesquisas, que poderão contribuir para a melhoria da qualidade dos estudos em ESE:

- Investigar a validade do constructo do instrumento de qualidade apresentado, no sentido de verificar se o conjunto de perguntas utilizado é realmente relacionado com avaliação da qualidade dos artigos;
- Expandir e atualizar esse trabalho a partir da avaliação de qualidade de outros tipos de estudos da área de ESE como: *survey*, estudos de caso, etnografias, pesquisa ação, revisões sistemáticas da literatura, etc.;

- Em virtude de termos realizado uma avaliação quantitativa, não foi possível revelar implicações para a prática na realização de experimentos controlados. Por isso, seria interessante realizar uma pesquisa qualitativa, por exemplo, fazendo entrevistas destinadas a extrair razões que ajudem a entender a falta de qualidade quanto às dimensões de garantia da qualidade de um estudo.
- Realizar uma análise detalhada de como o contexto e o delineamento experimental são descritos nos artigos;
- Avaliar os resultados encontrados nesta dissertação através da replicação desse estudo;
- Desenvolver um *framework,* ou modelo, ou guia, ou processo que apoie a avaliação de qualidade de Estudos Empíricos em Engenharia de Software.

6.3 Conclusões

Este estudo realizou uma avaliação quantitativa da qualidade dos experimentos publicados nas principais fontes de divulgação da Comunidade de Engenharia de Software Empírica, no período de 1997 a 2012. O estudo beneficiou-se do processo de busca e seleção de um mapeamento sistemático que examinou 876 artigos, dos quais analisamos 104, categorizados por seus autores como experimentos controlados.

Cinco perguntas de pesquisa foram usadas para guiar o estudo, que foi conduzido seguindo as quatro fases (definição, planejamento, coleta e interpretação) da abordagem GQM (*Goal Question Metric*). O procedimento de avaliação da qualidade foi selecionar um conjunto de estudos e analisá-los com base em uma escala de qualidade criada, principalmente, a partir de listas de verificação amplamente utilizadas por pesquisadores da área de ES. É importante ressaltar que todo o processo de planejamento e execução deste estudo foi apresentado nesta dissertação, permitindo posterior avaliação e replicação por terceiros.

De acordo com a análise dos indicadores, percebe-se que houve uma melhora estatisticamente significativa na qualidade dos experimentos que relataram explicitamente o uso de algum tipo de mecanismo de suporte, o que evidencia a importância da inclusão e da aplicação de metodologias de apoio que permitam planejar, executar e analisar resultados de estudos empíricos em engenharia de software.

Em relação a evolução da qualidade quanto a replicabilidade dos estudos os procedimentos estatísticos realizados identificaram que não houve evolução

significativa na qualidade desta variável ao longo dos quatro períodos avaliados (1997 a 2000; 2001 a 2004, 2005 a 2008 e 2009 a 2012), no entanto, a classificação geral da qualidade entre os três locais avaliados (EASE, ESEM e ESEJ) foi BOA e o índice global de qualidade obteve 63,42 pontos. Com respeito a evolução da qualidade em relação ao rigor estatístico, a classificação geral dessa variável foi MUITO BOA e apresentou uma provável tendência de aumento da qualidade no período de 2009 a 2012, porém, não houve evolução significativa entre o período avaliado, o que significa, também, que ao longo dos anos o índice global do rigor estatístico tem mantido o mesmo padrão de qualidade.

Em relação aos locais avaliados percebeu-se uma diferença estatisticamente significativa no índice de qualidade global quando comparamos o *Journal* (ESEJ) com os outros dois locais (EASE e ESEM). Portanto, podemos afirmar que entre os três locais analisados, o *Journal* (ESEJ) possui maior nível de qualidade. Este resultado, provavelmente, pode ser atribuído a uma maior rigidez exigida pelo *Journal*, bem como, um maior amadurecimento das pesquisas submetidas a este (que em geral já foram submetidas antes a outras conferências), o que tende a elevar a qualidade dos trabalhos.

Em geral, o índice global de qualidade dos estudos foi classificado como BOM e não apresentou diferenças significativas no período, o que pode-se entender que esses estudos mantiveram o mesmo padrão de qualidade. No entanto, isso preocupa, pois não foram identificados avanços estatisticamente significativos na qualidade ao longo dos anos.

Conclui-se esta pesquisa, explicitando algumas contribuições, dentre elas, destaca-se primeiro o fato de apresentar uma visão de qualidade dos trabalhos da comunidade de ESE e com isso, agregar conhecimentos sobre como foi e como estão os estudos empíricos em ES. Em seguida, a possibilidade de evolução do instrumento de avaliação utilizado para que possa analisar a qualidade de outros tipos de estudos, uma vez que possui critérios gerais e específicos. Por fim, espera-se, também, ter contribuído no sentido de fornecer às outras pesquisas uma compreensão sobre os conceitos que norteiam a construção de um modelo, processo ou guia que possa dar suporte à avaliação da qualidade de EE e com isso outros pesquisadores conduzam estudos com maior qualidade.

Referências

[1] AFZAL, Wasif; TORKAR, Richard; FELDT, Robert. **A systematic review of search-based testing for non-functional system properties**. Information and Software Technology, v. 51, n. 6, p. 957-976, 2009.

[2] ALMEIDA, A.; BARREIROS, E.; SARAIVA, J.; SOARES, S. **Mecanismos para Guiar Estudos Empíricos em Engenharia de Software: Um Mapeamento Sistemático**. Proceedings of 8 th Experimental Software Engineering Latin American Workshop, 2011. p.37.

[3] ALVES, V., NIU, N., ALVES, C., & VALENÇA, G. **Requirements engineering for software product lines: A systematic literature review**. Information and Software Technology, v. 52, n. 8, p. 806-820, 2010.

[4] ANDRADE, Maria Margarida de. **Introdução à Metodologia do Trabalho Científico**. 5. ed. São Paulo: Atlas. 2001.

[5] ARKSEY, H.; O'MALLEY, L. **Scoping studies: towards a methodological framework**. International *Journal* of social research methodology, 2005. v. 8, n. 1, p. 19–32.

[6] AZZEH, M. **A replicated assessment and comparison of adaptation techniques for analogy-based effort estimation**. Empirical Software Engineering, 1 fev. 2012. v. 17, n. 1-2, p. 90–127. Acesso em: 29 dez. 2013.

[7] BASILI, V. R. **The role of experimentation in software engineering: past, current, and future**.

[8] BASILI, V. R.; SELBY, R. W.; HUTCHENS, D. H. **Experimentation in software engineering**. IEEE Transactions on Software Engineering, jul. 1986. v. SE-12, n. 7, p. 733–743.

[9] BEECHAM, S., BADDOO, N., HALL, T., ROBINSON, H., & SHARP, H. **Motivation in Software Engineering: A systematic literature review**. Information and Software Technology, 2008. v. 50, n. 9, p. 860–878.

[10] CHEN, Lianping; BABAR, M. Ali; CAWLEY, Ciaran. **A status report on the evaluation of variability management approaches**. In: Proceedings of the 13th International Conference on Evaluation and Assessment in Software Engineering. British Computer Society. 2009.

[11] CHURCHILL JR, G. A.; PETER, J. P. **Research design effects on the reliability of rating scales: a meta-analysis**. *Journal* of marketing research, 1984. p. 360–375.

[12] COELHO, P. S.; ESTEVES, S. P. **The choice between a five-point and a ten-point scale in the framework of customer satisfaction measurement**. International *Journal* of Market Research, 2007. v. 49, n. 3, p. 313–339.

[13] COHEN, J.; OTHERS. **A coefficient of agreement for nominal scales. Educational and psychological measurement**, 1960. v. 20, n. 1, p. 37–46.

[14] CORBIN, J.; STRAUSS, A. **Basics of qualitative research: Techniques and procedures for developing grounded theory**. [S.l.]: Sage, 2008.

[15] COTE, J. A.; BUCKLEY, M. R. **Measurement error and theory testing in consumer research: An illustration of the importance of construct validation**. *Journal* of Consumer Research, 1988. p. 579–582.

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

[17] DENNIS, A.; VALACICH, J. **Conducting research in information systems**. Communications of the AIS, 2001. v. 7, n. 5, p. 1–41.

[18] DIESTE, O., GRIMÁN, A., JURISTO, N., & SAXENA, H. **Quantitative determination of the relationship between internal validity and bias in software engineering experiments: consequences for systematic literature reviews**.

[19] DO, H.; ELBAUM, S.; ROTHERMEL, G. **Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact**. Empirical Software Engineering, 2005. v. 10, n. 4, p. 405–435.

[20] DYBA, T.; DINGSØYR, T. **Strength of evidence in systematic reviews in software engineering**. Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement. 2008, p. 178-187.

[21] DYBA, T.; KAMPENES, V. B.; SJOBERG, D. I. **A systematic review of statistical power in software engineering experiments**. Information and Software Technology, 2006. v. 48, n. 8, p. 745–755.

[22] DYBA, T.; KITCHENHAM, B. A.; JORGENSEN, M. **Evidence-based software engineering for practitioners**. IEEE Software, 22:58–65, January 2005.

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

[24] FINK, A. **Conducting research literature reviews: from the Internet to paper**. [S.l.]: Sage, 2010.

[25] FLYVBJERG, B. **Five misunderstandings about case-study research**. Qualitative inquiry, 2006. v. 12, n. 2, p. 219–245.

[26] FUKS, H. **SISTEMAS COLABORATIVOS**. [S.l.]: Elsevier Brasil, [s.d.]. v. 1.

[27] HIGGINS, J. P.; GREEN, S.; COLLABORATION, C. **Cochrane handbook for systematic reviews of interventions**. [S.l.]: Wiley Online Library, 2008. v. 5.

[28] IEEE. **IEEE glossary of software engineering terminology**, IEEE standard 610.12. Technical report, IEEE, 1990.

[29] JEDLITSCHKA, A.; PFAHL, D. **Reporting guidelines for controlled experiments in software engineering**. IEEE.

[30] JUNI, P., BARTLETT, C., HOLENSTEIN, F., & STERNE, J. (2003). **How important are comprehensive literature searches and the assessment of trial quality in systematic reviews?: Empirical study**.

[31] JURISTO, N.; VEGAS, S. **The role of non-exact replications in software engineering experiments**. Empirical Software Engineering, 2011. v. 16, n. 3, p. 295–324.

[32] JURISTO, Natalia; MORENO, Ana M. **Basics of software engineering experimentation**. Springer Publishing Company, Incorporated, 2010.

[33] KAMPENES, V. **Quality of design, analysis and reporting of software engineering experiments**: A systematic review. 2007.

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

[35] KITCHENHAM, B. A., BRERETON, O. P., BUDGEN, D., & LI, Z. **An Evaluation of Quality Checklist Proposals-A participant-observer case study**. Proceeding of EASE. 2009.

[36] KITCHENHAM, B. A., PFLEEGER, S. L., PICKARD, L. M., JONES, P. W., HOAGLIN, D. C., El EMAM, K., & ROSENBERG, J. **Preliminary guidelines for empirical research in software engineering**. Software Engineering, IEEE Transactions on, 2002. v. 28, n. 8, p. 721–734.

[37] KITCHENHAM, B. A., SJØBERG, D. I., DYBÅ, T., PFAHL, D., BRERETON, P., BUDGEN, D., & RUNESON, P. **Three empirical studies on the agreement of reviewers about the quality of software engineering experiments**. Information and Software Technology, 2012. v. 54, n. 8, p. 804–819.

[38] KITCHENHAM, B. A.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. 2007.

[39] KITCHENHAM, B., SJØBERG, D. I., BRERETON, O. P., BUDGEN, D., DYBÅ, T., HÖST, M., & RUNESON, P. (2010, September). **Can we evaluate the quality of software engineering experiments?** In Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2010.

[40] KITCHENHAM, B., SJØBERG, D. I., BRERETON, O. P., BUDGEN, D., DYBÅ, T., HÖST, M., & RUNESON, P. **Trends in the Quality of Human-Intensive Software Engineering Experiments: A Quasi-Experiment**. IEEE, 2013.

[41] KITCHENHAM, B. **Procedures for performing systematic reviews**. Keele, UK, Keele University, 2004. v. 33, p. 2004.

[42] KITCHENHAM, B.; PICKARD, L.; PFLEEGER, S. L. **Case studies for method and tool evaluation. Software**, IEEE, 1995. v. 12, n. 4, p. 52–62.

[43] LANDIS, J. R.; KOCH, G. G. **The measurement of observer agreement for categorical data biometrics**, 1977. p. 159–174.

[44] LIKERT, R. **A technique for the measurement of attitudes**. Archives of psychology, 1932.

[45] MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de Metodologia Científica**.6. ed. São Paulo: Atlas, 2006.

[46] MASSEY JR, F. J. **The Kolmogorov-Smirnov test for goodness of fit**. *Journal* of the American statistical Association, 1951. v. 46, n. 253, p. 68–78.

[47] MATTAR, F. N. **Pesquisa de marketing**. 5ª. Edi\ccão. São Paulo. Atlas, 1999.

[48] MOHER, D., JADAD, A. R., NICHOL, G., PENMAN, M., TUGWELL, P., & WALSH, S. **Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists**. Controlled clinical trials, 1995. v. 16, n. 1, p. 62–73.

[49] PAI, M., MCCULLOCH, M., GORMAN, J. D., PAI, N., ENANORIA, W., KENNEDY, G., & COLFORD JR, J. M. **Systematic reviews and meta-analyses: an illustrated, step-by-step guide**. The National medical *Journal* of India, 2004. v. 17, n. 2, p. 86.

[50] PERRY, D. E.; PORTER, A. A.; VOTTA, L. G. **Empirical studies of software engineering: a roadmap**.

[51] PETTICREW, M.; ROBERTS, H. **Systematic Reviews in the Social Sciences: A Practical Guide**. Wiley-Blackwell, 2005.

[52] REVIEWS, U. OF Y. C. FOR; DISSEMINATION; AKERS, J. **Systematic reviews: CRD's guidance for undertaking reviews in health care**. [S.l.]: Centre for Reviews and Dissemination, 2009.

[53] RICHARDSON R. J. **Pesquisa social: métodos e técnicas**. 3. ed. rev. ampl. São Paulo: Atlas, 2007.

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

[55] ROWE, B. H., STROME, T. L., SPOONER, C., BLITZ, S., GRAFSTEIN, E., & WORSTER, A. **Reviewer agreement trends from four years of electronic submissions of conference abstract**. BMC medical research methodology, 2006. v. 6, n. 1, p. 14.

[56] SEVERINO, Antônio Joaquim. **Metodologia do trabalho científico**. 23. ed. rev. e atualizada. São Paulo: Cortez, 2007.

[57] SHANG, A., HUWILER-MÜNTENER, K., NARTEY, L., JÜNI, P., DÖRIG, S., STERNE, J. A., & EGGER, M. **Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy**. The Lancet, 2005. v. 366, n. 9487, p. 726–732.

[58] SHULL, F.; SINGER, J.; SJOBERG, D. I. **Guide to advanced empirical software engineering**. [S.l.]: Springer, 2008.

[59] SILVA, F. Q., SANTOS, A. L., SOARES, S., FRANÇA, A. C. C., MONTEIRO, C. V., & MACIEL, F. F. **Six years of systematic literature reviews in software engineering: An updated tertiary study**. Information and Software Technology, 2011. v. 53, n. 9, p. 899–913.

[60] SOLINGEN, R. VAN; BERGHOUT, E. **The Goal/Question/Metric Method: a practical guide for quality improvement of software development**. [S.l.]: McGraw-Hill London, 1999. v. 2

[61] SOMMERVILLE Ian. **Software Engineering**. Addison-Wesley Publishing Company, USA, 2007.

[62] THIOLLENT, M. **Metodologia da pesquisa-ação**. 2. ed. São Paulo: Editora Autores Associados, 1986.

[63] TICHY, W. F. **Should computer scientists experiment more?** Computer, 1998. v. 31, n. 5, p. 32–40.

[64] TORII, K., MATSUMOTO, K. I., NAKAKOJI, K., TAKADA, Y., TAKADA, S., & SHIMA, K. Ginger2: **An environment for computer-aided empirical software engineering**. Software Engineering, IEEE Transactions on, v. 25, n. 4, p. 474-492, 1999.

[65] TRAVASSOS, G. H.; GUROV, D.; AMARAL, E. **Introduão à engenharia de software experimental**. [S.l.]: UFRJ, 2002.

[66] WELCH II, G. E.; GABBE, S. G. **Review of statistics usage in the American *Journal* of Obstetrics and Gynecology**. American *Journal* of obstetrics and gynecology, 1996. v. 175, n. 5, p. 1138–1141.

[67] WOHLIN, C., RUNESON, P., HÖST, M., OHLSSON, M. C., REGNELL, B., & WESSLÉN, A. **Experimentation in software engineering**. [S.l.]: Springer Publishing Company, Incorporated, 2012.

[68] YANCEY, J. M. **Ten rules for reading clinical research reports**. The American *Journal* of Surgery, 1990. v. 159, n. 6, p. 533–539.

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

[70] ZANELLA, Liane Carly Hermes. **Metodologia de estudo e de pesquisa em administração**. 1. ED. Florianópolis: Departamento de Ciências da Administração / UFSC; [Brasília]: CAPES: UAB, 2009.

Apêndice A – Escores Individuais dos estudos avaliados

ID	ANO DE PUBLICAÇÃO	LOCAIS DE PUBLICAÇÃO	ÍNDICE DE AVALIAÇÃO
PS022	1999	EASE	61,56
PS024	2000	EASE	50,64
PS050	2002	EASE	63,27
PS055	2002	EASE	57,00
PS060	2004	EASE	51,01
PS082	2005	EASE	52,17
PS089	2006	EASE	53,14
PS100	2006	EASE	57,08
PS116	2008	EASE	54,82
PS120	2008	EASE	57,50
PS140	2010	EASE	53,18
PS165	2011	EASE	51,58
PS171	2012	EASE	58,53
PS187	2012	EASE	51,56
PS192	2002	ESEM	56,64
PS196	2002	ESEM	49,74
PS214	2003	ESEM	61,45
PS218	2003	ESEM	56,05
PS226	2003	ESEM	63,86
PS240	2004	ESEM	66,86
PS242	2004	ESEM	62,00
PS243	2004	ESEM	49,78
PS251	2004	ESEM	54,58
PS256	2004	ESEM	32,24
PS267	2005	ESEM	61,95
PS268	2005	ESEM	46,80
PS281	2005	ESEM	54,82
PS288	2005	ESEM	71,40
PS295	2005	ESEM	45,20
PS303	2005	ESEM	34,45
PS309	2006	ESEM	49,01
PS316	2006	ESEM	42,79
PS317	2006	ESEM	53,86
PS325	2006	ESEM	64,69
PS326	2006	ESEM	55,31
PS328	2006	ESEM	54,52
PS330	2006	ESEM	50,88

ID	ANO DE PUBLICAÇÃO	LOCAIS DE PUBLICAÇÃO	ÍNDICE DE AVALIAÇÃO
PS336	2006	ESEM	54,21
PS341	2006	ESEM	39,85
PS343	2006	ESEM	50,13
PS349	2007	ESEM	61,71
PS352	2007	ESEM	26,45
PS353	2007	ESEM	58,07
PS355	2007	ESEM	62,50
PS363	2007	ESEM	58,09
PS368	2007	ESEM	48,66
PS380	2007	ESEM	59,21
PS382	2007	ESEM	65,24
PS383	2007	ESEM	39,63
PS385	2007	ESEM	58,60
PS405	2008	ESEM	23,51
PS415	2008	ESEM	68,38
PS417	2009	ESEM	62,68
PS423	2009	ESEM	63,22
PS430	2009	ESEM	52,83
PS446	2009	ESEM	58,46
PS470	2010	ESEM	67,39
PS477	2010	ESEM	61,54
PS483	2010	ESEM	59,67
PS487	2010	ESEM	65,64
PS495	2011	ESEM	54,45
PS498	2011	ESEM	53,53
PS499	2011	ESEM	47,61
PS508	2011	ESEM	48,42
PS511	2011	ESEM	61,43
PS518	2011	ESEM	52,32
PS522	2012	ESEM	60,00
PS531	2012	ESEM	47,61
PS550	1998	ESEJ	45,07
PS551	1997	ESEJ	60,48
PS556	1997	ESEJ	62,13
PS558	1998	ESEJ	57,50
PS561	1998	ESEJ	59,04
PS563	1998	ESEJ	73,66
PS601	1999	ESEJ	61,03
PS623	2001	ESEJ	64,56
PS644	2002	ESEJ	43,97
PS648	2002	ESEJ	60,72
PS652	2003	ESEJ	70,11
PS664	2004	ESEJ	75,44
PS668	2004	ESEJ	64,96

ID	ANO DE PUBLICAÇÃO	LOCAIS DE PUBLICAÇÃO	ÍNDICE DE AVALIAÇÃO
PS671	2004	ESEJ	58,64
PS683	2005	ESEJ	62,35
PS684	2005	ESEJ	58,11
PS685	2005	ESEJ	75,02
PS686	2005	ESEJ	63,55
PS692	2005	ESEJ	59,80
PS700	2006	ESEJ	71,16
PS709	2006	ESEJ	49,85
PS714	2006	ESEJ	68,03
PS721	2007	ESEJ	65,94
PS728	2007	ESEJ	66,05
PS729	2007	ESEJ	68,75
PS741	2008	ESEJ	70,44
PS745	2008	ESEJ	79,87
PS752	2008	ESEJ	50,64
PS764	2009	ESEJ	62,74
PS766	2009	ESEJ	61,64
PS769	2009	ESEJ	58,60
PS770	2009	ESEJ	60,31
PS774	2009	ESEJ	70,88
PS792	2010	ESEJ	58,60
PS836	2011	ESEJ	45,42
PS848	2012	ESEJ	55,96

Apêndice B – Estudos Primários Incluídos

ID	TITTLE	YEAR	SOURCE
PS22	Experimental assessment of the e ect of inheritance on the maintainability of object-oriented systems	1999	EASE
PS24	Early lifecycle work: influence of individual characteristics, methodological constraints, and interface constraints.	2000	EASE
PS50	Investigating the influence of software inspection process parameters on inspection meeting performance	2002	EASE
PS55	Validating metrics for data warehouses	2002	EASE
PS60	Are Reviews an Alternative to Pair Programming?	2004	EASE
PS82	Assessing multiview framework (MF) comprehensibility and efficiency: A replicated experiment	2005	EASE
PS89	Assessing the value of Architectural Information Extracted from Patterns for Architecting	2006	EASE
PS100	An Experiment Measuring the Effects of Maintenance T asks on Program Knowledge	2006	EASE
PS116	Comparing Inspection Methods using Controlled Experiments	2008	EASE
PS120	Impact of Experience and Team Size on the Quality of Scenarios for Architecture Evaluation	2008	EASE
PS140	A Controlled Experiment on Team Meeting Style in Software Architecture Evaluation	2010	EASE
PS165	Using Background Colors to Support Program Comprehension in Software Product Lines	2011	EASE
PS171	Evaluating Methods and Technologies in Software Engineering with Respect to Developers' Skill Level	2012	EASE
PS187	Using the ISOIIEC 9 126 product quality model to classify defects a Controlled Experiment	2012	EASE
PS192	An Experimental Comparison of Checklist-Based Reading and Perspective-Based Reading for UML Design Document Inspection	2002	ESEM
PS196	Empirical Validation of Class Diagram Metrics	2002	ESEM
PS214	An Experimental Evaluation of Inspection and Testing for Detection of Design Faults	2003	ESEM
PS218	Applying Use Cases to Design versus Validate Class Diagrams – A Controlled Experiment Using a Professional Modelling Tool	2003	ESEM
PS226	Investigating the Accuracy of Defect Estimation Models for Individuals and Teams Based on Inspection Data	2003	ESEM
PS240	Assessing the Reproducibility and Accuracy of Functional Size Measurement Methods through Experimentation	2004	ESEM
PS242	Comparing Code Reading Techniques Applied to Object-oriented Software Frameworks with regard to Effectiveness and Defect Detection Rate	2004	ESEM
PS243	Comparing the Fault Detection Effectiveness of N-way and Random Test Suites	2004	ESEM
PS251	The Influence of the Level of Abstraction on the Evolvability of Conceptual Models of Information Systems	2004	ESEM
PS256	Using Students as Subjects in Requirements Prioritization	2004	ESEM

ID	TITTLE	YEAR	SOURCE
PS267	An Empirical investigation on the Visualization of Temporal Uncertainties in Software Engineering Project Planning	2005	ESEM
PS268	An Empirical Study of Factors that Affect User Performance when Using UML Interaction Diagrams	2005	ESEM
PS281	Evaluating a Rapid Simulation Modelling Process (RSMP) through Controlled Experiments	2005	ESEM
PS288	Investigating Training Effects on Software Reviews: A Controlled Experiment	2005	ESEM
PS295	Quality vs. Quantity: Comparing Evaluation Methods in a Usability-Focused Software Architecture Modification Task	2005	ESEM
PS303	Tool Assisted Identifier Naming for Improved Software Readability: An Empirical Study	2005	ESEM
PS309	A Family of Empirical Studies to Compare Informal and Optimization-based Planning of Software Releases	2006	ESEM
PS316	An Empirical Comparison Between Pair Development and Software Inspection in Thailand	2006	ESEM
PS317	An Empirical Evaluation of a Testing and Debugging Methodology for Excel	2006	ESEM
PS325	Distributed Versus Face-to-Face Meetings for Architecture Evaluation: A Controlled Experiment	2006	ESEM
PS326	Documenting Design Decision Rationale to Improve Individual and Team Design Decision Making: An Experimental Evaluation	2006	ESEM
PS328	Eliciting Better Quality Architecture Evaluation Scenarios: A Controlled Experiment on Top-Down vs. Bottom-Up	2006	ESEM
PS330	Evaluating Advantages of Test Driven Development: a Controlled Experiment with Professionals	2006	ESEM
PS336	Maximising the Information Gained From an Experimental Analysis of Code Inspection and Static Analysis for Concurrent Java Components	2006	ESEM
PS341	Requirement Error Abstraction and Classification: An Empirical Study	2006	ESEM
PS343	Testing and Inspecting Reusable Product Line Components: First Empirical Results	2006	ESEM
PS349	A Replicate Empirical Comparison between Pair Development and Software Development with Inspection	2007	ESEM
PS352	An Estimation Model for Test Execution Effort	2007	ESEM
PS353	An Experimental Evaluation of the Effectiveness and Efficiency of the Test Driven Development	2007	ESEM
PS355	Assessing, Comparing, and Combining Statechart- based testing and Structural testing: An Experiment	2007	ESEM
PS363	Defect Detection Efficiency: Test Case Based vs. Exploratory Testing	2007	ESEM
PS368	Evaluating the Usefulness and Ease of Use of a Groupware Tool for the Software Architecture Evaluation Process	2007	ESEM
PS380	Test Inspected Unit or Inspect Unit Tested Code?	2007	ESEM
PS382	The Impact of Group Size on Software Architecture Evaluation: A Controlled Experiment	2007	ESEM
PS383	Toward Reducing Fault Fix Time: Understanding Developer Behavior for the Design of Automated Fault Detection Tools	2007	ESEM
PS385	Usability Evaluation Based on Web Design Perspectives	2007	ESEM
PS405	Model-based Functional Size Measurement	2008	ESEM
PS415	The Impact of Time Controlled Reading on Software Inspection Effectiveness and Efficiency	2008	ESEM

ID	TITTLE	YEAR	SOURCE
PS417	Impact of the Visitor Pattern on Program Comprehension and Maintenance	2009	ESEM
PS423	An Empirical Study of the Effects of Personality in Pair Programming using the Five-Factor Model	2009	ESEM
PS430	Does Explanation Improve the Acceptance of Decision Support for Product Release Planning?	2009	ESEM
PS446	Test Case Prioritization Based on Data Reuse An Experimental Study	2009	ESEM
PS470	Evaluating a Model of Software Managers' Information Needs – An Experiment	2010	ESEM
PS477	On the Effectiveness of Screen Mockups in Requirements Engineering: Results from an Internal Replication	2010	ESEM
PS483	The Effects of Neuroticism on Pair Programming: An Empirical Study in the Higher Education Context	2010	ESEM
PS487	Usability Evaluation of Multi-Device/Platform User Interfaces Generated by Model-Driven Engineering	2010	ESEM
PS495	An Experimental Evaluation of the Impact of System Sequence Diagrams and System Operation Contracts on the Quality of the Domain Model	2011	ESEM
PS498	Experimental Analysis of Textual and Graphical Representations for Software Architecture Design	2011	ESEM
PS499	Exploring Software Measures to Assess Program Comprehension	2011	ESEM
PS508	Preserving Aspects via Automation: a Maintainability Study	2011	ESEM
PS511	Supporting Online Updates of Software Product Lines: A Controlled Experiment	2011	ESEM
PS518	Assessing the Impact of Real-Time Machine Translation on Requirements Meetings: A Replicated Experiment	2011	ESEM
PS522	Does the Prioritization Technique Affect Stakeholders' Selection of Essential Software Product Features?	2012	ESEM
PS531	Recommender Systems for Manual Testing	2012	ESEM
PS550	A Comparison of Tool-Based and Paper-Based Software Inspection	1998	ESEJ
PS551	A Controlled Experiment to Evaluate On-Line Process Guidance	1997	ESEJ
PS556	An Experimental Comparison of the Maintainability of Object-Oriented and Structured Design Documents	1997	ESEJ
PS558	An Extended Replication of an Experiment for Assessing Methods for Software Requirements Inspections	1998	ESEJ
PS561	Comparing Detection Methods For Software Requirements Inspections: A Replication Using Professional Subjects	1998	ESEJ
PS563	Does Every InspectionReallyNeed a Meeting?	1998	ESEJ
PS601	Perspective-based Usability Inspection: An Empirical Validation of Efficacy	1999	ESEJ
PS623	Assessing the Changeability of two Object-Oriented Design Alternatives: a Controlled Experiment	2001	ESEJ
PS644	Experimental Evaluation of Program Slicing for Fault Localization	2002	ESEJ
PS648	Using a Reliability Growth Model to Control Software Inspection	2002	ESEJ
PS652	An Externally Replicated Experiment for Evaluating the Learning Effectiveness of Using Simulations in Software Project Management Education	2003	ESEJ

ID	TITTLE	YEAR	SOURCE
PS664	A Controlled Experiment Comparing the Maintainability of Programs Designed with and without Design Patterns—A Replication in a Real Programming Environment	2004	ESEJ
PS668	Are Reviews an Alternative to Pair Programming?	2004	ESEJ
PS671	Group Processes in Software Effort Estimation	2004	ESEJ
PS683	Collecting Feedback During Software Engineering Experiments	2005	ESEJ
PS684	Expert Estimation of Web-Development Projects: Are Software Professionals in Technical Roles More Optimistic Than Those in Non-Technical Roles?	2005	ESEJ
PS685	Investigating the Role of Use Cases in the Construction of Class Diagrams	2005	ESEJ
PS686	Methodology Support in CASE Tools and Its Impact on Individual Acceptance and Use: A Controlled Experiment	2005	ESEJ
PS692	The Influence of the Level of Abstraction on the Evolvability of Conceptual Models of Information Systems	2005	ESEJ
PS700	An experimental evaluation of a higher-ordered- typed-functional specification-based test-generation technique	2006	ESEJ
PS709	Prioritizing JUnit Test Cases: An Empirical Assessment and Cost-Benefits Analysis	2006	ESEJ
PS714	Using patterns for the refinement and translation of UML models: A controlled experiment	2006	ESEJ
PS721	Building measure-based prediction models for UML class diagram maintainability	2007	ESEJ
PS728	Maximising the information gained from a study of static analysis technologies for concurrent software	2007	ESEJ
PS729	Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques	2007	ESEJ
PS741	An experiment on the role of graphical elements in architecture visualization	2008	ESEJ
PS745	Comparing distributed and face-to-face meetings for software architecture evaluation: A controlled experiment	2008	ESEJ
PS752	Presenting software engineering results using structured abstracts: a randomised experiment	2008	ESEJ
PS764	A subject-based empirical evaluation of SSUCD's performance in reducing inconsistencies in use case models	2009	ESEJ
PS766	An experimental investigation of personality types impact on pair effectiveness in pair programming	2009	ESEJ
PS769	Assessing IR-based traceability recovery tools through controlled experiments	2009	ESEJ
PS770	Assessing the understandability of UML statechart diagrams with composite states—A family of empirical studies	2009	ESEJ
PS774	Experimental evaluation of a tool for the verification and transformation of source code in event-driven systems	2009	ESEJ
PS792	An experimental comparison of ER and UML class diagrams for data modelling	2010	ESEJ
PS836	A replicated assessment and comparison of adaptation techniques for analogy-based effort estimation	2011	ESEJ
PS848	Program comprehension of domain-specific and general-purpose languages: comparison using a family of experiments	2012	ESEJ

Apêndice C – Dados Brutos da Pesquisa

Paper	PS22	PS24	PS50	PS55	PS60	PS82	PS89	PS100	PS116	PS120	PS140	PS165	PS171	PS187	PS192
Year	1999	2000	2002	2002	2004	2005	2006	2006	2008	2008	2010	2011	2012	2012	2002
Source	EASE	EASE	EASE	EASE	EASE	EASE	EASE	EASE	EASE	EASE	EASE	EASE	EASE	EASE	ESEM
Mecan	0	0	1	1	0	1	1	0	1	1	0	1	1	1	1
SQ1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ2	1	0,5	1	1	1	1	0	1	1	1	1	1	1	1	1
SQ3	0,5	0	1	0,5	1	0,5	0,5	0,5	0,5	0,5	1	1	0,5	0,5	1
SQ4	1	1	1	1	1	1	1	0,5	1	1	1	1	1	1	1
SQ5	1	1	1	0,5	1	0,5	0,5	0,5	1	1	0,5	1	1	1	1
SQ6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ7	1	1	1	0,5	1	0	1	0,5	1	1	1	1	1	1	0
SQ8	0	0,5	0	0,5	0,5	0	1	0,5	0,5	0,5	0,5	0	0,5	0	0
SQ9	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
SQ10	0,5	0,5	1	1	0,5	1	0,5	0,5	0,5	0,5	0,5	0	0	0	0,5
SQ11	0,5	0,5	0,5	1	1	1	1	0,5	1	1	0,5	0	1	0	1
SQ12	0	0	0	0,5	1	1	0	0,5	0,5	0	0	0	0	0	0
SQ13	1	1	1	1	1	1	1	1	1	0,5	1	1	1	1	1
SQ14	1	1	1	0,5	0,5	0,5	1	0,5	0,5	1	1	1	1	1	1
SQ15	0,5	0,5	0,5	0,5	0,5	1	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
SQ16	0	0,5	0,5	0	0,5	0,5	1	0,5	0,5	0	0	0	1	0,5	0,5
SQ17	0	0,5	0,5	0	0	0	0,5	0,5	0,5	0	1	0	0	0	0
SQ18	1	0	0	1	0,5	1	1	0	1	1	1	0	0	0	1
SQ19	1	0	0	1	0	0,5	1	0	1	1	0,5	0	0	0	1
SQ20	1	0,5	0	0,5	0	1	1	0	1	1	1	0	0	0	0,5
SQ21	0,5	1	1	0,5	0,5	1	1	1	0,5	1	1	0	0,5	0	1
SQ22	1	1	1	0,5	1	1	1	0,5	0	1	1	0	1	1	1
SQ23	0,5	0,5	0	0	0	0,5	0,5	0	0,5	0,5	0	0	0	0	0
SQ24	0	1	0	1	1	0,5	0,5	0,5	0,5	0,5	1	1	0,5	0,5	1
SQ25	0	0,5	0	0	0,5	0,5	1	0,5	1	0,5	1	0	0,5	0,5	0
SQ26	0	0	0	0	0	0	0,5	0,5	0	0	0	0	0	0	0
SQ27	1	1	0	0	1	0	1	0,5	1	1	1	1	1	1	1
SQ28	0	0,5	0	0,5	1	1	0,5	0,5	1	1	0,5	0,5	1	0	1
SQ29	0	0,5	0	0,5	0	0	1	0,5	0	0,5	0	0	0	1	0
SQ30	0	0,5	0,5	0	0,5	0	0	0	0	0	0	0	0	0	0
SQ31	0,5	1	1	0,5	0	0	0	1	0	0	0	0	0	0	0
SQ32	0,5	1	1	0	0	0	0	0	0	0	0	0	0	0	0
SQ33	0,5	0,5	0,5	1	1	0	0,5	1	1	0,5	0,5	0,5	1	1	1
SQ34	0	0,5	0,5	1	0,5	1	0	1	0	0	1	1	1	1	1
SQ35	1	0	1	1	1	1	1	0,5	1	1	0,5	1	1	1	1
SQ36	0	0	0,5	0	0	0	1	0	0	0	0	0	0	0	0
SQ37	1	0	1	1	1	0	0	0	1	0,5	1	1	1	1	1
SQ38	0,5	0,5	0,5	0,5	0,5	0,5	1	1	1	0,5	0,5	0,5	0,5	0,5	0,5
SQ39	0	0,5	0,5	0	0	0	0	0	0	0	0	0	0	0	0
SQ40	0,5	0,5	0	0	0	1	0	0	0	0,5	0	0	0	0	0
SQ41	0	0,5	0,5	1	0	0	0,5	0	0	0	0	0	1	0	1
SQ42	0,5	1	1	1	0,5	1	1	1	1	1	1	1	0,5	1	1
SQ43	0,5	0,5	0,5	0,5	0	0,5	1	1	0	1	0	0	1	0	0
SQ44	1	1	1	0,5	1	1	0,5	0,5	1	1	1	1	1	1	1
SQ45	1	0,5	1	0,5	1	1	0,5	0,5	1	0,5	1	1	1	1	1
SQ46	1	0,5	1	0	1	0,5	1	0,5	1	0,5	1	0,5	1	1	1
SQ47	0,5	0,5	0,5	0	0,5	0	0,5	0	1	0	0	0,5	0,5	0	0
SQ48	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1
SQ49	0,5	0,5	0,5	0,5	0	1	0	1	0,5	0	0	1	1	0,5	0,5
SQ50	0,5	0	0,5	1	0	0	0	1	0,5	0	0	0	1	0	0
SQ51	0,5	0	1	1	1	1	0,5	1	1	1	1	1	1	1	1
SQ52	1	0,5	1	0,5	0	1	0,5	0	0	0	0	1	1	1	1
SQ53	1	0,5	1	0	0,5	1	0	1	0,5	1	1	1	1	1	0,5
SQ54	0,5	0	0	0,5	0	0	0,5	1	0	1	0	1	0,5	0,5	0
SQ55	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0

Paper	PS196	PS214	PS218	PS226	PS240	PS242	PS243	PS251	PS256	PS267	PS268	PS281	PS288	PS295	PS303
Year	2002	2003	2003	2003	2004	2004	2004	2004	2004	2005	2005	2005	2005	2005	2005
Source	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM
Mecan	1	1	1	1	1	1	1	0	0	1	1	0	0	0	1
SQ1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ2	1	1	1	1	1	1	0,5	1	1	1	0,5	1	1	0,5	0
SQ3	1	1	0,5	0	1	0,5	0,5	1	0,5	1	1	0	1	0	1
SQ4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ5	1	1	1	1	1	1	0,5	1	1	1	1	1	1	1	1
SQ6	1	1	1	1	1	1	0	1	0	1	1	0	0,5	1	0
SQ7	1	1	1	1	1	1	1	1	0,5	0,5	1	1	0	0,5	1
SQ8	0	0	0	0,5	0,5	0	0	0	0	0	0	0	0	0	0
SQ9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,5
SQ10	1	0,5	0	1	1	1	0,5	1	0,5	1	1	1	1	0,5	0,5
SQ11	1	1	0,5	1	1	1	1	0,5	0	1	1	0	1	0,5	1
SQ12	0	0	0	0	0,5	1	0	0	0	0,5	0	1	0,5	0	0
SQ13	1	1	1	0,5	1	1	0	0,5	1	0,5	0,5	1	0,5	0,5	0,5
SQ14	0,5	1	1	1	1	0,5	0	0,5	1	1	1	1	1	0,5	0,5
SQ15	0,5	0,5	1	0,5	0	0,5	0	0,5	0,5	0	0	1	0	0	0
SQ16	0	0	0	0	1	0,5	0	0	0	0	0	1	0,5	0	0,5
SQ17	0	0	0	0	0,5	0	0	0	0	0	0	0,5	0	0	0
SQ18	1	1	1	0,5	1	1	1	0,5	0	1	1	0	1	0	1
SQ19	1	1	0,5	0	1	1	1	0	0	1	1	0	1	0	1
SQ20	1	0,5	0,5	0	1	1	0,5	0,5	0	1	1	0	1	0	0,5
SQ21	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0
SQ22	0	0,5	1	1	1	1	1	0	0	0	1	0	0	1	0
SQ23	0	0	0	0,5	0,5	0,5	0	0	0	0	0	0	1	0,5	0
SQ24	0,5	1	1	1	1	0,5	0	0,5	0	1	1	0,5	1	0	0,5
SQ25	0	0,5	0,5	1	1	0,5	0	0,5	0,5	1	0	0	1	0,5	0
SQ26	0	0	0	0	1	0	0	0	0	0,5	0	0	0	0	0
SQ27	1	1	0,5	1	1	1	0,5	0	0	1	1	1	1	0,5	0
SQ28	0,5	1	0,5	0	1	1	0	0,5	0,5	1	0,5	0,5	1	0,5	0,5
SQ29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ30	0	0	0	0,5	0	0	0	0	0	0	0	0	1	0	0
SQ31	0	0,5	1	0	0	0	0	1	0	1	0	1	1	0	0
SQ32	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0
SQ33	0,5	0,5	0,5	0,5	1	0	1	0,5	0	0,5	1	1	0,5	0,5	0,5
SQ34	1	1	0	1	1	1	1	0	1	0	0	0	1	0	0
SQ35	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
SQ36	0	0	0	0,5	0	0	1	0	0,5	0	0	0	0	0	0
SQ37	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
SQ38	0,5	0,5	0,5	1	1	1	1	1	1	1	1	1	1	1	1
SQ39	0	0	0	1	0,5	0	1	1	0,5	1	1	1	0,5	0	0
SQ40	0	0	0	0	0	0,5	0	0	0	0	0	0	0	0	0
SQ41	0	0,5	0	0,5	1	0	0	0	0	1	0	0	0,5	0	1
SQ42	1	1	0,5	1	1	1	0	1	1	1	0	1	1	1	0,5
SQ43	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SQ44	1	1	1	1	1	1	1	1	1	1	1	1	1	0,5	1
SQ45	1	1	1	1	1	1	1	1	1	1	1	1	1	0,5	1
SQ46	1	1	1	1	0,5	1	0,5	1	1	1	0,5	1	1	1	1
SQ47	0	0,5	0,5	0,5	1	0,5	1	1	0	1	0	1	1	0,5	0
SQ48	1	1	1	1	1	1	1	1	0	1	0	1	1	0,5	0
SQ49	1	0,5	0,5	1	1	1	0,5	1	0	0,5	0,5	0,5	0,5	0,5	0,5
SQ50	0	0	0	1	0,5	0,5	0	0,5	0	0	0	0	0,5	1	0
SQ51	1	1	1	1	1	1	1	1	0	1	1	1	1	0,5	1
SQ52	0	0,5	0	1	0	0,5	1	1	1	0,5	1	1	1	1	0,5
SQ53	0,5	1	0	1	0,5	0,5	1	1	1	1	0,5	1	1	0,5	0
SQ54	0	1	0	1	1	0	1	0	1	1	0	0	1	0	0
SQ55	0	0,5	0	0,5	0	0	0	0	0	0	0	0	0	0	0

Paper	PS309	PS316	PS317	PS325	PS326	PS328	PS330	PS336	PS341	PS343	PS349	PS352	PS353	PS355	PS363
Year	2006	2006	2006	2006	2006	2006	2006	2006	2006	2006	2007	2007	2007	2007	2007
Source	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM
Mecan	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1
SQ1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ2	1	0,5	0	1	0,5	0	1	0,5	0,5	0,5	0,5	1	0,5	1	1
SQ3	1	1	0,5	1	0,5	1	1	0,5	0,5	1	0,5	0	1	0	0,5
SQ4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ5	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
SQ6	0,5	0	0,5	1	1	1	1	0	0	1	1	1	1	1	1
SQ7	1	0	1	1	0	0,5	1	1	0	0,5	1	0	1	1	1
SQ8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ9	0,5	0,5	1	0	1	0,5	1	1	0,5	0,5	1	1	1	1	1
SQ10	1	1	0,5	1	1	1	0,5	1	1	1	1	0	1	1	1
SQ11	0,5	0	1	1	1	1	1	1	0,5	1	1	0	0,5	1	1
SQ12	1	0	0,5	0,5	0	0	0	0	0	0	0,5	0	0	0	0
SQ13	0,5	0,5	0,5	0,5	0,5	0,5	1	1	1	0	0,5	0	0,5	1	0,5
SQ14	0,5	0,5	0,5	0,5	0,5	0,5	1	1	1	1	1	0	1	1	1
SQ15	0	0	0	0	0	0	0,5	0,5	0	0	0	0	0	1	0
SQ16	0,5	0,5	0	0	0	0	1	1	0	0	0	0	0,5	0,5	0
SQ17	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ18	0	0	1	1	1	1	0,5	0	0,5	1	0,5	0	0,5	1	1
SQ19	0	0	1	1	1	1	1	0	0,5	1	0,5	0	1	1	1
SQ20	0	0	1	1	0,5	1	0,5	0	0,5	1	1	0	1	1	1
SQ21	0	0	0	1	1	1	0	1	0	1	1	0	1	1	1
SQ22	0	1	0	1	1	1	0	0	1	0	1	1	0	0	0
SQ23	0,5	1	0	0	0	0,5	0,5	0,5	0	0	0	1	1	0,5	0
SQ24	0	0	1	1	1	0,5	0,5	1	0,5	1	1	1	1	1	1
SQ25	1	0,5	0	1	1	1	0	0	0	0	1	1	0	1	0,5
SQ26	1	0,5	0	0,5	0	0	0	0	0	0	1	0	0	1	0
SQ27	1	1	1	1	1	1	1	1	0	0,5	1	1	1	1	1
SQ28	0,5	1	0,5	0,5	0	1	0,5	0,5	0	1	1	0,5	1	0,5	0,5
SQ29	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
SQ30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ31	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0
SQ32	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
SQ33	1	0,5	1	1	0,5	1	1	1	0,5	0	1	0	1	1	0,5
SQ34	1	1	0	1	0	0	0	1	0	0	1	0	1	0	0
SQ35	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1
SQ36	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1
SQ37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ38	0,5	0,5	0,5	1	0,5	1	1	1	1	1	1	1	1	1	1
SQ39	1	0,5	0,5	1	1	0,5	0	0	0	0	0	0	0	0	0
SQ40	0	0	0	0	0	0,5	0	1	0	0	0	0	0	0	0
SQ41	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1
SQ42	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1
SQ43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ44	1	0,5	1	1	0,5	1	1	0,5	0,5	1	1	1	0,5	0,5	1
SQ45	1	0,5	1	1	0,5	0,5	0,5	0	0	1	1	0,5	1	0,5	1
SQ46	1	0,5	0,5	1	0,5	0,5	1	0	1	0	0,5	0	1	1	1
SQ47	1	1	1	1	1	1	0,5	1	0	1	0	0,5	0	0	0
SQ48	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
SQ49	0,5	0,5	0,5	0,5	0,5	0	0	1	0,5	0	1	0,5	1	1	0,5
SQ50	1	0,5	0,5	0,5	0	0,5	1	1	0,5	0	1	0	0,5	1	1
SQ51	0,5	0,5	1	1	1	1	1	1	1	0,5	1	0	0,5	1	1
SQ52	1	0,5	1	1	1	0,5	0	1	0	0,5	1	0,5	1	0	1
SQ53	0,5	1	1	1	1	0,5	1	1	1	1	1	1	1	1	1
SQ54	0,5	0	0,5	1	1	1	0,5	1	0,5	0,5	0,5	0	0,5	1	0,5
SQ55	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5	0

Paper	PS368	PS380	PS382	PS383	PS385	PS405	PS415	PS417	PS423	PS430	PS446	PS470	PS477	PS483	PS487
Year	2007	2007	2007	2007	2007	2008	2008	2009	2009	2009	2009	2010	2010	2010	2010
Source	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM
Mecan	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
SQ1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ2	0,5	0,5	1	0,5	0,5	0	1	0,5	1	0,5	1	0,5	0,5	0,5	0,5
SQ3	0	0,5	0	0	0,5	0	0,5	0,5	1	1	0,5	1	1	1	1
SQ4	1	1	1	1	1	0,5	1	1	1	1	1	1	1	1	1
SQ5	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
SQ6	1	1	1	1	1	1	1	0,5	0	1	0	1	0,5	1	1
SQ7	1	1	0	1	1	0	0	0	0	1	1	0	1	0	1
SQ8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ9	1	1	1	1	1	1	1	1	1	0,5	1	1	1	1	1
SQ10	0,5	0,5	1	1	1	0,5	1	1	1	0,5	1	1	1	0,5	1
SQ11	0	0,5	1	0	1	0	1	1	0,5	1	0,5	1	1	0,5	1
SQ12	0	0	1	0	0	0	1	0	1	0	0	0,5	0	0	1
SQ13	0,5	0,5	0,5	1	1	0	1	0,5	0,5	0,5	0,5	0,5	0,5	0,5	1
SQ14	1	0,5	1	1	1	0	1	0,5	0,5	1	0,5	1	0,5	1	1
SQ15	0	0	0	0	0	0	0,5	0	0	0	0	0,5	0	0	0
SQ16	0	0	0,5	0,5	0,5	0	0,5	0	0,5	0	0,5	1	0	0	0
SQ17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ18	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1
SQ19	0,5	0,5	1	0	1	0	1	1	1	0,5	0,5	1	1	1	1
SQ20	0	1	1	0	1	0	1	0,5	1	1	1	1	1	1	1
SQ21	0	1	1	0	0	0	1	1	0	0,5	1	1	1	1	1
SQ22	0	1	1	0	0	0	0	1	0	1	1	1	0	1	1
SQ23	0	0	1	0	0	0	1	0	0	0	0,5	0,5	1	0	0
SQ24	0,5	1	1	0	0,5	1	1	1	0,5	1	1	1	1	1	1
SQ25	0	1	0,5	0	1	0	0,5	1	1	0	1	1	1	0	0
SQ26	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
SQ27	1	0,5	1	1	1	0	1	1	1	1	1	1	1	0	1
SQ28	0	1	1	1	0,5	0	0,5	0,5	0,5	0,5	0	0	0,5	0,5	0,5
SQ29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ31	0	0	0	1	1	1	1	0	0	0	1	0	0	1	0
SQ32	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0
SQ33	0,5	1	1	0,5	1	0	1	1	1	0,5	1	1	1	1	1
SQ34	0	1	0	1	1	0	0	0	0,5	0	1	1	1	1	1
SQ35	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1
SQ36	0	1	1	0	1	0	1	1	1	0	0	1	0	1	1
SQ37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ38	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
SQ39	0	0	1	0	0	0	0	0	0	0	0	0,5	0	0,5	0,5
SQ40	1	1	0,5	0	0	0	0	0	0	0	0	0	0	0	0
SQ41	1	1	1	0,5	1	0	1	1	1	1	1	1	0,5	1	1
SQ42	1	1	1	1	1	0,5	0,5	1	1	1	1	1	0	1	0
SQ43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ44	1	1	1	1	0,5	0	1	0,5	1	1	1	1	1	1	1
SQ45	1	0,5	1	1	0,5	0	1	1	1	1	1	1	1	1	1
SQ46	1	0,5	0,5	0	0	0	0,5	1	1	0	1	1	0,5	1	1
SQ47	0,5	1	1	1	0,5	0	1	0	0	0,5	0,5	1	1	0	1
SQ48	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
SQ49	0,5	1	0,5	0	1	0	0,5	1	1	0,5	1	1	1	1	1
SQ50	1	1	1	0	1	0	1	1	1	0,5	0,5	1	0	0,5	1
SQ51	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1
SQ52	0,5	1	1	1	0,5	0	0	0	1	1	0	1	1	1	1
SQ53	1	0,5	1	1	1	1	1	1	0,5	1	0,5	0	1	1	1
SQ54	1	0	1	1	1	0	1	1	1	1	0	0,5	1	0	1
SQ55	0,5	0	0	0	0	0,5	0	1	0	0	0	0,5	1	0	0

Paper	PS495	PS498	PS499	PS508	PS511	PS518	PS522	PS531	PS550	PS551	PS556	PS558	PS561	PS563	PS601
Year	2011	2011	2011	2011	2011	2011	2012	2012	1998	1997	1997	1998	1998	1998	1999
Source	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEM	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ
Mecan	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1
SQ1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ2	0,5	1	1	1	1	1	1	0,5	0,5	1	1	0,5	0,5	1	0,5
SQ3	0,5	0,5	0	0,5	0,5	0	0,5	0,5	1	1	1	1	1	0,5	1
SQ4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ5	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
SQ6	0,5	1	1	1	0,5	0	0,5	0,5	0,5	1	1	0	1	1	1
SQ7	0	0	1	1	1	1	0	0	1	1	0	0	1	1	1
SQ8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ9	0,5	1	0,5	1	1	1	1	1	0	0,5	0,5	0,5	1	1	0,5
SQ10	1	1	0,5	0,5	1	0,5	0,5	1	1	1	1	1	1	1	1
SQ11	1	1	0	0,5	1	0	1	0	0	1	1	1	1	1	1
SQ12	0	0,5	0	0	0	0,5	0,5	0	0,5	0	1	1	0	0	1
SQ13	0,5	1	0,5	1	0,5	0	1	0,5	1	1	1	0,5	0,5	0,5	0,5
SQ14	0	0,5	0	0,5	0,5	0	0,5	0,5	0,5	0,5	1	1	1	0,5	0,5
SQ15	0,5	0	0	0	0	0	0	0	0	0	0,5	0,5	0,5	0	0
SQ16	0	0	0	0,5	0	0	0,5	1	0	0,5	0,5	0	0,5	0,5	1
SQ17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ18	1	1	0	0,5	1	0,5	1	0	0	1	1	1	1	1	1
SQ19	1	1	0	0	1	0,5	0,5	0	0	1	1	1	1	1	1
SQ20	1	1	0	0,5	1	0,5	0,5	0	0	1	1	1	1	1	1
SQ21	0	0	0	0	0	0	1	0	0,5	1	1	1	1	0,5	0,5
SQ22	1	0	0	1	1	0	0	1	0,5	1	0	0	1	1	0,5
SQ23	0	1	0	0,5	1	0	0	0	1	1	0	0	0,5	0	0
SQ24	1	1	0,5	0,5	1	1	1	0,5	1	0,5	0,5	1	0,5	1	1
SQ25	0	1	0	0,5	1	1	0,5	0,5	1	1	0	1	1	1	1
SQ26	0	1	0	0	0	1	1	0	0	0	0,5	1	1	1	0,5
SQ27	1	1	1	1	1	1	1	0,5	1	0,5	1	1	1	1	1
SQ28	0,5	0	0	0	1	0,5	0,5	0,5	1	0	0,5	1	1	1	1
SQ29	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0
SQ30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5
SQ31	0	0	0	0	0	0	1	1	0	0,5	0	0	0,5	0	1
SQ32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ33	1	1	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	1	1	1	1	1
SQ34	1	0	1	0	1	1	0	0	0	1	1	1	1	1	0
SQ35	1	1	1	1	1	1	1	1	1	1	1	0,5	1	1	1
SQ36	0	0	0	1	0	0	0	0	1	0	0,5	0	0	0	0
SQ37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ38	1	1	1	1	1	0,5	0,5	1	0,5	1	1	1	1	1	1
SQ39	0,5	0,5	0	0,5	0	1	0	0	1	1	1	0,5	1	0,5	1
SQ40	0	0	0	1	1	0	0	0	0,5	0,5	0	0	0	0	0,5
SQ41	1	1	0,5	0	0	0	1	0	0	0,5	0,5	0,5	0	1	0,5
SQ42	1	0	1	1	1	1	0,5	1	1	1	1	1	1	1	1
SQ43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ44	1	1	0,5	1	1	0,5	1	1	1	1	1	1	1	1	1
SQ45	1	1	0,5	0,5	0,5	0,5	0,5	1	1	0,5	1	0,5	1	1	1
SQ46	1	1	0,5	0	1	1	1	0,5	1	0,5	1	1	1	1	1
SQ47	1	1	1	1	1	0,5	0,5	0,5	1	1	1	1	0,5	1	1
SQ48	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ49	1	0,5	1	0,5	1	1	0,5	0,5	1	1	1	1	0,5	1	0,5
SQ50	1	0,5	1	0,5	1	1	0,5	1	1	1	1	0,5	0,5	0,5	0,5
SQ51	1	1	1	0,5	1	1	1	1	0,5	1	1	1	0,5	1	0,5
SQ52	0	1	1	1	0,5	0,5	0,5	0,5	0,5	0,5	1	1	0,5	1	1
SQ53	1	0	1	1	0,5	1	0,5	1	0	0	1	1	1	1	1
SQ54	1	0,5	1	1	1	1	1	0,5	1	1	1	0,5	0	0,5	1
SQ55	0	0	1	0	0,5	0	0	0	0	0	0	0	0	1	0

Paper	PS623	PS644	PS648	PS652	PS664	PS668	PS671	PS683	PS684	PS685	PS686	PS692	PS700	PS709	PS714
Year	2001	2002	2002	2003	2004	2004	2004	2005	2005	2005	2005	2005	2006	2006	2006
Source	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ
Mecan	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1
SQ1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ2	0,5	0,5	0,5	0,5	1	1	0,5	1	0,5	0,5	0,5	1	1	1	0,5
SQ3	1	0,5	0,5	1	1	1	1	0	1	0,5	1	1	1	0	1
SQ4	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
SQ5	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
SQ6	0,5	1	1	1	1	1	1	1	1	1	0	1	1	0	1
SQ7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ8	0	0	0	0	0	1	1	1	0,5	0	0	0	0	0	0
SQ9	0,5	0,5	1	0,5	1	1	0,5	1	1	1	1	1	1	1	1
SQ10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ11	1	1	1	1	1	1	0	0,5	0,5	1	0,5	1	1	1	1
SQ12	0,5	0	0	0,5	1	0,5	0	0,5	0,5	1	0,5	0,5	1	0	0,5
SQ13	1	0,5	0,5	1	1	0,5	1	0,5	1	0,5	1	0,5	0,5	0	1
SQ14	1	0,5	1	1	1	0,5	1	0,5	1	1	1	1	1	0	1
SQ15	0	0	1	0	1	0	0,5	0	0,5	1	0,5	0	1	0	0
SQ16	0	0	0,5	0,5	0,5	0,5	0,5	0	0,5	0,5	0,5	1	1	0	0,5
SQ17	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SQ18	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1
SQ19	1	1	0	1	1	1	0	0,5	0	1	1	1	1	1	1
SQ20	1	1	0	1	1	0	0	0,5	0	1	0,5	0,5	1	1	1
SQ21	1	1	1	1	1	0	1	0	1	0,5	1	1	1	0	1
SQ22	1	0	1	1	0	1	0	0	0	1	1	0	1	1	0
SQ23	0	0	1	0,5	0,5	0,5	0	1	0	0,5	0,5	0	0,5	0	0
SQ24	0,5	0,5	1	1	1	1	1	1	1	1	1	0,5	1	1	0,5
SQ25	1	1	1	1	1	1	1	1	0,5	1	0,5	0	1	0,5	0,5
SQ26	0	0,5	0,5	0,5	1	1	0,5	1	0	1	0,5	0	1	0,5	0,5
SQ27	1	1	1	1	1	1	1	1	0,5	1	1	0,5	1	0	1
SQ28	1	0	0,5	1	1	1	1	1	1	1	1	1	1	0,5	1
SQ29	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
SQ30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ31	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0
SQ32	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
SQ33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ34	0	0,5	0,5	1	1	0	1	1	1	1	1	1	0	0	1
SQ35	0,5	0	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ36	0	0	0,5	1	0,5	1	0	0,5	0	0	1	0	1	0	0
SQ37	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
SQ38	1	0,5	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ39	0,5	0,5	0,5	0,5	1	1	1	1	1	1	1	0,5	1	0,5	1
SQ40	0	0	0	0	0,5	0	0	1	0	0,5	1	0	0,5	0	0
SQ41	0,5	0,5	0,5	1	1	0,5	1	1	1	1	0,5	0	1	0,5	1
SQ42	0,5	1	1	1	1	1	0,5	0,5	1	1	1	1	1	0,5	1
SQ43	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
SQ44	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ45	1	0,5	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ46	1	0,5	1	1	1	1	1	0,5	0,5	1	0,5	1	1	1	1
SQ47	0,5	0,5	0,5	1	1	1	1	1	0,5	1	0,5	1	1	1	1
SQ48	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ49	0,5	0,5	0,5	1	1	1	1	0,5	1	1	1	1	1	0,5	0,5
SQ50	0,5	0	0,5	1	1	1	1	0,5	1	1	1	0	0,5	1	1
SQ51	1	0	1	1	1	1	0,5	1	1	1	1	1	1	0,5	1
SQ52	1	1	0,5	1	1	1	1	1	1	0,5	0	1	1	1	1
SQ53	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ54	0,5	0,5	1	0,5	1	1	1	0,5	1	0,5	1	0	1	1	0,5
SQ55	1	0	0	0,5	0	0	1	0	0	0	0	0	0	0	1

Paper	PS721	PS728	PS729	PS741	PS745	PS752	PS764	PS766	PS769	PS770	PS774	PS792	PS836	PS848
Year	2007	2007	2007	2008	2008	2008	2009	2009	2009	2009	2009	2010	2011	2012
Source	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ	ESEJ
Mecan	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ2	0,5	0,5	0,5	1	1	1	1	1	0	0	1	0,5	1	0,5
SQ3	1	0,5	0,5	1	1	0,5	0,5	0,5	1	0,5	0,5	0,5	0	0,5
SQ4	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ5	1	1	1	1	1	1	1	1	1	1	1	1	0	1
SQ6	1	1	0	1	1	1	1	1	0	1	1	1	0	0
SQ7	1	0,5	1	1	1	0	0	0	1	0	1	0	1	1
SQ8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ9	1	1	0,5	0,5	1	0,5	1	1	1	1	1	1	1	1
SQ10	1	1	1	1	1	1	1	1	1	1	1	1	0,5	1
SQ11	1	1	1	1	1	0,5	1	1	1	1	1	1	0	1
SQ12	1	0,5	1	1	1	0	0	0	0	0	0	0	0,5	0
SQ13	1	1	1	0,5	0,5	1	1	1	1	1	0,5	0,5	0	0,5
SQ14	1	1	1	1	1	1	0,5	1	0	0,5	1	0	0	0
SQ15	0	0	1	0	0	0	0	0	0	0	0	0	0	0
SQ16	0,5	1	0,5	1	0	0,5	0	1	0	0,5	1	0,5	0	0,5
SQ17	0	0	1	0	1	0	0	1	0	0	0	0	0	0
SQ18	1	0,5	1	1	1	1	1	1	1	1	1	1	0	1
SQ19	1	1	1	1	1	1	1	1	0,5	1	1	1	0	1
SQ20	1	0,5	1	0,5	1	1	1	1	1	0,5	1	1	0	1
SQ21	1	1	1	1	1	0	1	1	1	1	1	1	0	0
SQ22	0	1	1	1	1	1	1	1	1	1	1	0	0	0
SQ23	0	0,5	0,5	0	1	1	0,5	1	0	0,5	1	1	0	0
SQ24	0,5	0,5	0,5	0,5	0,5	1	1	1	1	1	1	1	1	1
SQ25	0,5	1	0,5	1	0,5	1	1	1	1	1	1	1	1	1
SQ26	0,5	1	0,5	0,5	0,5	0	0	0	0	0	0	0	0	0
SQ27	1	1	1	2	1	1	1	1	1	1	1	1	0	1
SQ28	0,5	1	1	1	1	0	1	0,5	0	0,5	0,5	0	0	0,5
SQ29	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SQ30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ31	0	0	0	1	1	0,5	1	0	0	0	0	0	0	0
SQ32	0	0	0	1	1	0	0	0	0	0	0	0	0	0
SQ33	1	1	1	1	1	0,5	0,5	0,5	1	1	1	1	1	1
SQ34	1	1	1	0	1	0	0	0	1	1	0	1	1	1
SQ35	1	1	1	1	1	1	1	1	1	1	1	1	0,5	1
SQ36	0	0	0	0	0	0	1	1	0	1	1	0	1	0
SQ37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ38	0,5	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ39	1	1	1	1	1	0	0	0	1	0	0	0	0	0
SQ40	0	1	0,5	0	0	1	0	1	0	0	0	0	0	0
SQ41	1	1	1	0,5	1	1	0	0	1	0	1	1	1	1
SQ42	0	1	1	0	1	1	1	1	1	1	1	1	0	1
SQ43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SQ44	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ45	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ46	1	1	0,5	1	1	1	1	1	0,5	1	1	1	1	1
SQ47	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ48	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SQ49	1	1	1	1	0,5	0	1	0	1	1	0,5	1	1	0,5
SQ50	1	1	1	0,5	0,5	0	0	1	1	1	1	1	0	1
SQ51	0,5	1	1	1	1	1	1	1	1	1	1	1	0,5	1
SQ52	1	0,5	0,5	1	1	0,5	1	0	0	0,5	1	1	1	1
SQ53	1	1	0	1	0,5	0	1	1	0,5	1	1	0,5	1	1
SQ54	1	1	1	1	1	0	1	1	1	1	1	0,5	1	1
SQ55	1	0	1	0	0	0	0	0	0	0	1	0	1	0

Apêndice D – Guia para realização do processo de extração dos dados

Análise da Qualidade de Experimentos Controlados no Contexto da Engenharia de Software Empírica

Guia para Extração de Dados

Eudis Teixeira[1], Liliane Sheyla, Alex Nery[1], Waldemar Ferreira Nt[1], Adauto Almeida[1], Sergio Soares[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

Resumo: Este trabalho apresenta um conjunto de instruções para a realização do processo de extração de dados, visando identificar, analisar e interpretar evidências da comunidade de Engenharia de Software Empírica - ESE para obter uma visão sobre qualidade e reportagem das pesquisas empíricas que envolvem experimentos controlados em ES, publicadas durante todo o período de existência dos veículos de divulgação investigados.

1. Introdução

Estudos empíricos têm se mostrado um importante instrumento para o avanço científico da área de Engenharia de Software. Como principais tipos de estudos empíricos encontrados nesta área, podemos destacar: **experimento controlado**, quasi-experimento, *survey*, etnografia, pesquisa-ação, estudo de caso e estudos sistemáticos da literatura (revisão e mapeamento). Nesse cenário, é essencial a realização de pesquisas para obter uma visão pormenorizada desses estudos, identificar os desafios e desenvolver recursos para conduzir a realização destes na ES.

Assim, esta pesquisa tem como objetivo realizar um estudo empírico no intuito de obter uma visão de como eram e como está os trabalhos da comunidade de engenharia de software empírica desde seu nascimento até hoje em específico quanto à **qualidade e reportagem** dos experimentos controlados em engenharia de software, publicados durante todo o período de existência dos locais pesquisados, a saber: as conferências EASE (Evaluation and Assessment in Software Engineering) e ESEM (Empirical Software Engineering and Measurement), além do jornal ESEJ

(Empirical Software Engineering). Este documento apresenta um conjunto de instruções necessárias para a realização da extração de dados dos 104 artigos encontrados para esta pesquisa.

2. Pesquisadores Envolvidos

Esta etapa do mapeamento será executada por cinco pesquisadores, divididos em quatro duplas. Todo o processo será realizado sob a orientação do Professor Doutor Sergio Soares que também será o responsável pela resolução de prováveis divergências entre as duplas.

Assim, as duplas responsáveis pelo processo de extração dos dados serão:

- Dupla 1: Eudis e Liliane
- Dupla 2: Eudis e Alex
- Dupla 3: Eudis e Neto
- Dupla 4: Eudis e Adauto

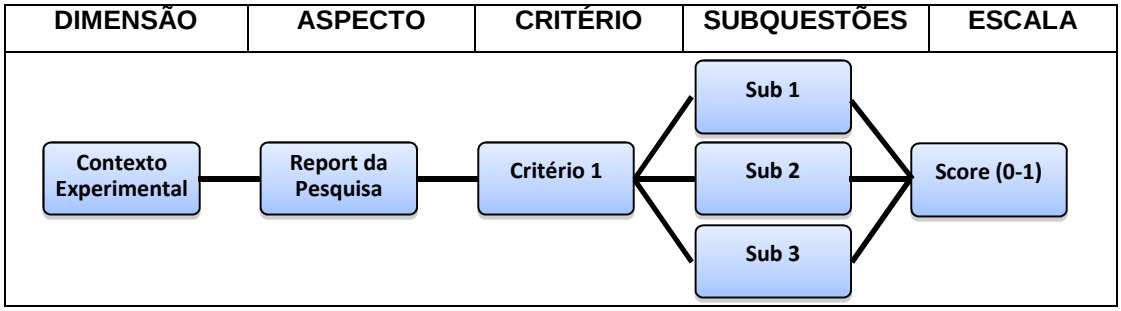
3. Processo de Extração dos Dados

O processo de extração de dados de cada artigo desta pesquisa será realizado por quatro duplas de pesquisadores, assim, cada artigo será analisado por dois pesquisadores. Para realizar a extração das informações de cada artigo, será utilizada uma planilha Excel com o Instrumento de Avaliação da Qualidade. É importante ressaltar que a planilha de avaliação possui uma coluna de observações para a indicação da localização exata da página, parágrafo ou seção do artigo, que motivou a escolha do avaliador por determinado valor da escala de qualidade em cada critério avaliado**.**

Seguem abaixo algumas observações importantes que devem ser levadas em consideração por todos os pesquisadores durante o processo de extração dos dados:

- Cada artigo será analisado por dois pesquisadores, por isso é importante que cada pesquisador faça a extração em planilhas separadas e após a conclusão de uma etapa, cada dupla faz a composição dos resultados em uma única planilha. Caso ocorram divergências, outro pesquisador será responsável por avaliar o artigo e resolver a divergência.
- Uma instância da estratégia de análise da qualidade é mostrada abaixo para que possamos ter uma ideia da constituição do instrumento de qualidade.

Tabela 1
Instância da Escala Utilizada



- Ao observarem na planilha de extração irão verificar que contém 19 critérios e cada um desses terá outras subquestões. Cada uma das subquestões deverá ser respondida de acordo com a escala e legenda abaixo:

Tabela 2
Escala do Tipo Likert Utilizada

ESCALA	**VALORES**		
Likert 3	Não Atende	Atende Parcialmente	Atende Totalmente
Score (0-1)	0	0,50	1

Tabela 3
Legenda da Escala Utilizada

Não Atende (0) → Quando não há nada no trabalho que atenda à subquestão avaliada.
Atende Parcialmente (0,5) → Deve ser concedido no caso em que há evidências que atendam parcialmente ao requisito avaliado.
Atende Totalmente (1) → Quando o trabalho apresenta atendimento total à subquestão avaliada**.**

- Ao perceber que um artigo não é um experimento controlado ou possa ser repetido, os dados devem ser registrados normalmente, porém, deve ser relatado o fato na coluna de observações para que posteriormente seja discutido pelo grupo e se for o caso excluído.
- A princípio todos os artigos incluídos são trabalhos completos, porém, caso encontre um

short paper, resumo, apresentação ou algo que não seja um trabalho completo, é preciso reportar o fato ao tutor da dupla para averiguação.

Por fim, é importante mencionar que qualquer dúvida encontrada durante o processo de extração precisa ser reportada a todos os envolvidos, para que todos estejam cientes das dúvidas levantadas e de suas respectivas soluções.

4. Cronograma básico de Atividades

O processo de extração dos dados será realiza em duas etapas, com duração de 15 dias cada. Em cada etapa um conjunto de artigos será analisado por cada dupla e ao final será realizada uma **Reunião Interna**, aonde cada dupla irá se reunir para fazer a junção dos resultados em uma única planilha e resolverem possíveis divergências. Após a primeira reunião poderá haver novos ajustes quanto aos procedimentos anteriormente explicitados e caso necessário será realizada novas reuniões com o Professor Sergio Soares (revisor interno) para dirimir dúvidas e/ou discutir resultados.

ETAPA 1 **Início: 01/11/2013** **Fim: 15/11/2013**	• Total de Artigos da etapa: 52 • Total de Artigos de cada dupla: 13 • Reunião Interna: **16/11**. • Horários da Reunião: **D1**: 8h – 9h | **D2**: 9h05 – 10h05 | **D3**: 10h10 – 11h10 | **D4**: 11h15 – 12h15
ETAPA 2 **Início: 18/11/2013** **Fim: 02/12/2013**	• Total de Artigos da etapa: 52 • Total de Artigos de cada dupla: 13 • Reunião Interna: **03/12**. • Horários da Reunião: **D1**: 8h – 9h | **D2**: 9h05 – 10h05 | **D3**: 10h10 – 11h10 | **D4**: 11h15 – 12h15