

# Utilização de Estudos Empíricos em Engenharia de Software

Mariela Inés Cortés, Francisco José Rêgo Lopes  
Lázaro João Cá, Luiz Alexandre Castelo Branco Gomes,  
Mário Sérgio Rodrigues Falcão, Nádia Rafaela Costa

<sup>1</sup> Mestrado Acadêmico em Ciência da Computação.  
Universidade Estadual do Ceará. Fortaleza. Brazil.

mariela@larces.uece.br

**Resumo.** *O processo experimental subsidia na compreensão e no avanço no campo de conhecimento através do aprendizado iterativo, no entanto a sua aplicação nas pesquisas em engenharia de software tem sido largamente questionada. O amadurecimento na área pressupõe uma maior adoção de estudos empíricos na validação das pesquisas realizadas. Este artigo apresenta uma análise dos trabalhos publicados na Conferência Internacional de Engenharia de Software (ICSE), evento considerado referência na área de Engenharia de Software, avaliando de forma qualitativa e quantitativa a evolução da utilização de avaliações empíricas nos trabalhos publicados nos últimos dez anos. Embora a quantidade e a qualidade dos estudos empíricos venham evoluindo ao longo do tempo, os dados levantados apontam para falhas na utilização dos métodos empíricos, e espaço para melhoria.*

## 1. Introdução

A experimentação é o centro de todo processo científico, no entanto, a sua utilização na pesquisa em engenharia de software tem sido questionada tanto em qualidade quanto em quantidade. Alguns autores citam dificuldades na utilização de experimentação [Juristo and Moreno 2013] e outros argumentam em relação à utilidade da experimentação nesta área [Tichy 1998].

Em 2006 Zannier, Melnik e Maurer [Zannier et al. 2006] conduziram uma revisão sistemática para diagnosticar a evolução da aplicação de estudos empíricos em artigos publicados nos *proceedings* da *International Conference on Software Engineering* (ICSE) ao longo das suas 29 edições. O ICSE é a principal conferência internacional na área de Engenharia de Software. No estudo, os autores concluíram que, embora a quantidade de artigos publicados nos anais que apresentam avaliação empírica tenha aumentado, os dados indicam que a qualidade das avaliações realizadas não melhorou, alertando à comunidade para a correta utilização das avaliações empíricas em pesquisas na área.

O presente artigo empreende a avaliação da qualidade dos estudos apresentados no ICSE, com foco no mapeamento das últimas dez edições da Conferência, de modo a verificar a evolução em termos quantitativos e qualitativos, da utilização de estudos empíricos, através de uma inspeção nos trabalhos publicados. O artigo está organizado da seguinte forma: a Seção 2 apresenta a metodologia de pesquisa e os passos envolvidos em todo o processo. A Seção 3 apresenta os resultados obtidos, através de gráficos e tabelas. A Seção 4 discorre sobre os resultados obtidos, com o intuito de responder às questões de

pesquisa. A Seção 5 discute os trabalhos relacionados. A Seção 6 apresenta ameaças à validade do estudo e a Seção 7 apresenta as conclusões.

## 2. Metodologia de Pesquisa

Para a realização do presente trabalho foram utilizados princípios da revisão sistemática (RS). Segundo [Keele 2007], a RS é uma técnica que permite identificar, avaliar e interpretar a pesquisa relevante disponível a respeito de uma questão de pesquisa, assunto ou fenômeno de interesse. Esse empreendimento deve ser realizado dentro de uma metodologia justa, confiável, rigorosa e auditável. O processo é dividido em três fases, envolvendo planejamento, execução e relatório.

A fase de **planejamento** compreende a definição e detalhamento de todos os passos que serão necessários para realização da RS, gerando como resultado o protocolo de revisão. Nele constarão as questões de pesquisa que se quer responder, e as estratégias de busca, seleção, extração, sintetização e disseminação dos estudos.

Na fase de **execução** é realizada a revisão sistemática propriamente dita seguindo o protocolo definido. Esta fase compreende a seleção e avaliação da qualidade dos estudos, bem como a extração e sintetização de dados sobre os artigos coletados.

A última fase, **relatório**, compreende o empacotamento da pesquisa, com a confecção do relatório final e descrição dos mecanismos de disseminação.

### 2.1. Planejamento

No contexto da fase de planejamento, as próximas seções definem o protocolo de revisão seguido na realização do presente trabalho.

#### 2.1.1. Questões de Pesquisa

Com o intuito de compreender a evolução da utilização de estudos empíricos nos artigos publicados nos *proceedings* do ICSE são propostas as seguintes questões de pesquisa:

- RQ1: A quantidade de estudos empíricos realizados tem incrementado ao longo dos últimos 10 anos do ICSE?
- RQ2: A qualidade dos estudos empíricos realizados tem incrementado ao longo dos últimos 10 anos do ICSE?

#### 2.1.2. Estratégias de Busca e Seleção

A busca manual realizada pelos pesquisadores compreendeu apenas os artigos publicados na trilha principal do ICSE, compreendendo as edições dos últimos 10 anos (2007 a 2016), dando continuidade à pesquisa apresentada em [Zannier et al. 2006]. Os artigos foram selecionados a partir dos *proceedings* do ICSE, excluindo *talks* de pesquisadores convidados, reportes de experiência, painéis e artigos provenientes de *workshops*. Inicialmente os autores identificaram e catalogaram os títulos dos artigos aceitos ao longo do período em estudo, estratificando-os por ano e atribuindo um número sequencial a cada estudo. O tamanho da população inicialmente considerada foi de 1226 artigos. Dado o

grande volume de publicações, uma amostra composta pelo 10% dos artigos de cada ano foi extraída. A extração das amostras foi feita de forma aleatória utilizando o serviço de geração de dados aleatórios *Random.org* [Random.Org ].

De posse da amostra de artigos, a seleção foi refinada da seguinte maneira: dois revisores, atuando de forma independente, verificaram a realização de algum tipo de avaliação empírica nos artigos. A partir das características dos estudos empíricos identificadas, e considerando o ponto de vista dos pesquisadores, os revisores determinaram qual tipo de estudo empírico foi empregado, dentre os seguintes: estudo de caso, experimento, *survey* ou revisão sistemática. Demais estudos que não se enquadram nos citados anteriormente são classificados como *outros*, tais como relato de experiência, metanálise, aplicação exemplo, discussão, etc.

### 2.1.3. Extração

A extração de dados dos artigos compreende especificamente a avaliação da sua qualidade, incluindo a existência de seção de análise de ameaças à validade. Esse processo será executado para os estudos que compreenderem experimentos, estudos de caso, revisões sistemáticas e *surveys*, segundo estabelecido em [Wohlin et al. 2012]

A avaliação da qualidade será empreendida considerando-se uma escala percentual de 0 a 100. Para obtenção do valor atribuído a cada estudo, serão aplicadas questões de avaliação específicas para cada tipo de estudo empírico apresentado, e atribuídas as pontuações 1 (quanto o artigo atende totalmente à questão), 0.5 (quando atende parcialmente) e 0 (quando não atende).

#### **Questões para Avaliação da Qualidade de um Estudo de Caso** [Runeson et al. 2012]

- O pesquisador teve envolvimento direto com o fenômeno?
- Objetivos, questões de pesquisa e as hipóteses (se aplicadas) são claros?
- O procedimento de coleta de dados foi definido?
- A metodologia de análise foi definida, com papéis e procedimentos de revisão?
- Existem conclusões claras da análise e recomendações para pesquisas futuras?
- Existe uma relação clara entre evidências a partir das observações e as conclusões?

#### **Questões para Avaliação da Qualidade de um Experimento** [Babbie 1999] [Malhotra 2015]

- O experimento apresenta hipóteses bem formuladas?
- As variáveis (dependentes e independentes) foram explicitadas?
- A amostra de participantes contém mais de 10 indivíduos?
- A amostragem da população foi selecionada de forma representativa?
- Foram realizados tratamentos (testes) aleatoriamente nos sujeitos participantes?
- Foram utilizadas técnicas da estatística descritiva na interpretação dos resultados?
- Foram utilizados testes estatísticos para rejeição da hipótese nula?

#### **Questões para Avaliação da Qualidade de uma Revisão Sistemática** [Keele 2007]

- Os critérios de inclusão e exclusão foram descritos e são apropriados?
- É provável que a RS tenha coberto todos os estudos relevantes?

- Os revisores avaliaram a qualidade/validade dos estudos incluídos?
- Os dados/estudos básicos foram adequadamente descritos?

### Questões para Avaliação da Qualidade de um *Survey* [Keele 2007]

- O estudo define explicitamente seus objetivos?
- O estudo especifica e descreve minuciosamente a população alvo? (em termos de unidade de análise, critérios de exclusão / inclusão, fontes, etc.)
- O estudo especifica e descreve minuciosamente o método de amostragem? (por exemplo: métodos de amostragem não probabilísticos ou probabilísticos, etc.)
- O estudo fornece informações sobre sua taxa de resposta?
- O questionário do estudo está disponível?
- O estudo avalia formalmente a sua confiabilidade? (cálculo de erros).

## 2.2. Execução

A execução da RS contou com a participação de 6 pesquisadores, mestrandos sob a supervisão de um professor orientador, todos pesquisadores da área de engenharia de software. A partir dos *proceedings* do ICSE, foram encontrados 1226 estudos, dos quais foram selecionados 10%, resultando em 122, estratificados ano a ano, conforme Tabela 1. Esse conjunto foi então aleatoriamente sorteado entre os 6 pesquisadores.

Após essa distribuição cada pesquisador avaliou os artigos sob sua responsabilidade, analisando se estes incluíam estudos empíricos ou não. Em caso afirmativo, estudos empíricos foram classificados de acordo com os tipos listados na Seção 2.1.2. A mesma avaliação foi feita por um segundo pesquisador, de modo a classificar com maior precisão cada tipo de estudo. Os casos em que as opiniões divergiram foram reavaliados através de consenso entre os dois pesquisadores.

**Table 1. Distribuição dos Estudos**

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
<b>População</b>	89	104	70	96	129	150	189	99	199	101	<b>1226</b>
<b>Amostra</b>	9	10	7	10	13	15	18	10	20	10	<b>122</b>

## 3. Relatório

Após a finalização do processo de extração, os dados obtidos foram consolidados, considerando os tipos de estudo e ainda sua avaliação de qualidade.

### 3.1. Categorização por Tipo de Estudo Empírico

O extrato foi avaliado quanto à existência ou não de estudo empírico nos artigos e sua respectiva classificação do ponto de vista dos pesquisadores. A Tabela 2 apresenta os quantitativos dos diversos tipos de estudo encontrados, de forma estratificada por ano. Os tipos de estudo que não se enquadram em estudo de caso, experimentação, revisão sistemática e *survey* foram agrupados na coluna "Outros Tipos". A coluna "Sem Estudo Empírico" totaliza os artigos nos quais nenhum tipo de avaliação empírica foi realizada.

Observa-se a predominância do tipo "Experimentação", abrangendo 31,15% dos trabalhos, seguido por "Estudo de Caso", com um 15,57% trabalhos. "Surveys" e "Revisões Sistemáticas" são estudos com pouca representatividade no cômputo total, respondendo, respectivamente, por 8,20% e 0,82% dos artigos sob estudo. É importante notar que 14,75% dos artigos não apresentaram nenhum tipo de avaliação empírica.

**Table 2. Categorização por Tipo de Estudo**

Ano	Estudo de Caso	Experi- mentação	Revisão Sistemática	Survey	Outros Tipos	Sem Estudo Empírico
2007	2	1	0	0	4	2
2008	2	2	0	2	3	1
2009	1	1	0	0	4	1
2010	0	2	0	3	2	3
2011	4	6	0	0	1	2
2012	0	6	0	0	7	2
2013	2	8	0	2	6	0
2014	2	4	0	0	0	4
2015	4	3	0	2	8	3
2016	2	5	1	1	1	0
<b>TOTAL</b>	19 (15,57%)	38 (31,15%)	1 (0,82%)	10 (8,20%)	36 (29,51%)	18 (14,75%)

### 3.2. Qualidade dos Estudos

A avaliação de qualidade dos estudos, conduzida de acordo com os critérios descritos na Seção 2.1.3, foi feita para trabalhos do tipo "Experimentação", "Estudo de Caso", "Survey" e "Revisão Sistemática". A utilização desses critérios produziu um percentual indicativo de qualidade para cada um dos estudos, variando de 0% a 100%. A Tabela 3 exibe a média desse indicador de qualidade, classificada por ano e tipo de estudo.

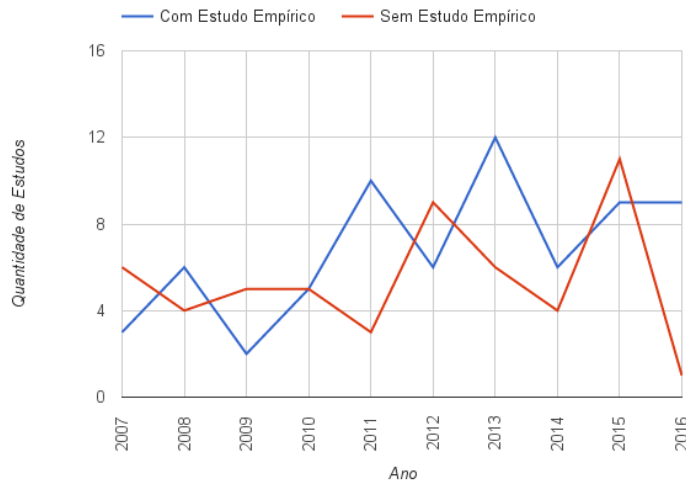
**Table 3. Média da Qualidade dos Estudos**

Ano	Estudo de Caso	Experimentação	Revisão Sistemática	Survey
2007	79,17%	31,25%	-	-
2008	33,33%	28,13%	-	40,63%
2009	66,67%	31,25%	-	-
2010	-	62,50%	-	41,67%
2011	81,25%	43,75%	-	-
2012	-	65,63%	-	-
2013	87,50%	30,47%	-	40,63%
2014	87,50%	78,13%	-	-
2015	85,42%	58,33%	-	59,38%
2016	75,00%	66,25%	62,50%	68,75%
<b>MÉDIA</b>	74,48%	49,57%	62,50%	50,21%

## 4. Discussão

Para responder a questão de pesquisa RQ1: "A quantidade de estudos empíricos realizados tem incrementado ao longo dos últimos 10 anos do ICSE?", foi traçado o gráfico que mostra a evolução do quantitativo de artigos que apresentam algum dos quatro tipos de estudo empírico para os quais se pode, nesse trabalho, medir sua qualidade. Avaliando a Figura 1 é possível constatar que o quantitativo de trabalhos que apresentam um dos quatro tipos de estudo empírico vem em tendência clara de crescimento, com a conseguinte

tendência de redução dos artigos que não apresentam esse tipo de estudo. Mais especificamente, no presente estudo, a porcentagem de artigos sem avaliação empírica foi de 14,75%, enquanto que na pesquisa realizada sobre a amostra nas edições anteriores do ICSE [Zannier et al. 2006] essa porcentagem representava o 30% da amostra.



**Figure 1. Evolução da Quantidade de Trabalhos**

Em relação à questão de pesquisa RQ2: "A qualidade dos estudos empíricos realizados tem incrementado ao longo dos últimos 10 anos do ICSE?" foram gerados gráficos a partir dos dados extraídos, de forma a ilustrar sua evolução.

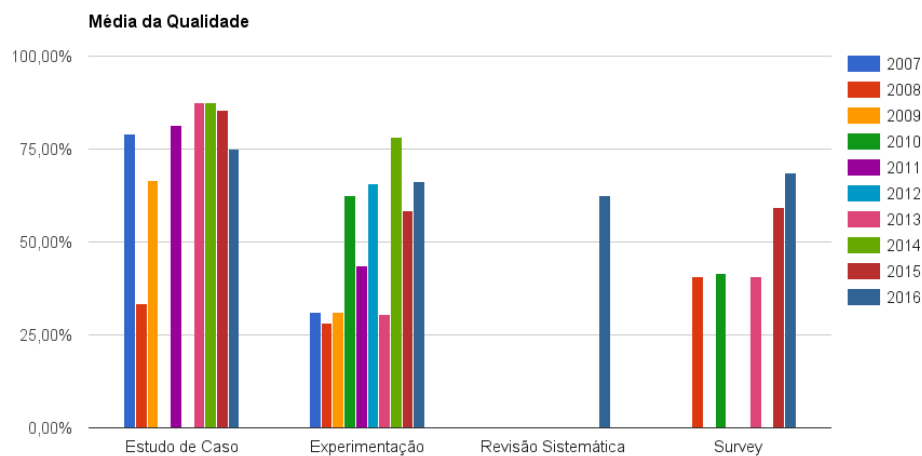
A Figura 2 mostra um quadro geral da qualidade média dos artigos, em um percentual que varia de 0 a 100. É possível observar que, em linhas gerais, existe variação na qualidade dos estudos ao longo do tempo. A **experimentação** se mostra o tipo de estudo empírico com maior qualidade ao longo dos anos sob estudo.

Além desse gráfico geral, foram traçados outros, individuais para cada um dos tipos de estudos empíricos considerados. Contudo, como só foi encontrado um único artigo do tipo "Revisão Sistemática", não foi possível traçar o gráfico para esse tipo. Para os outros três, cada gráfico exibe a sua evolução em termos de qualidade, ao longo do tempo. Também foi traçada a linha de tendência.

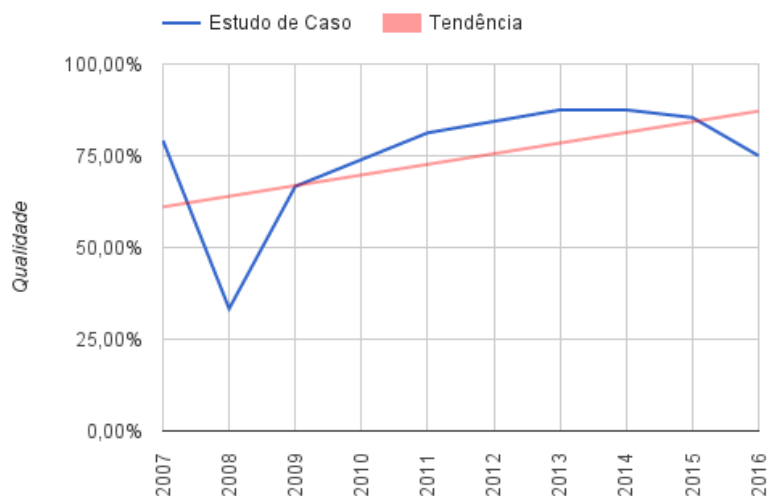
Embora as figuras 3, 4 e 5 exibam uma variação considerável da qualidade dos trabalhos ao longo dos anos, as linhas de tendência, desenhadas em cor vermelha, mostram que a qualidade vem crescendo. Desta forma a resposta a RQ2 é: sim, é evidente um incremento na qualidade dos estudos ao longo do tempo.

Por outro lado, a despeito dessa tendência de melhoria, somente 4 dos 122 estudos empíricos analisados apresentaram qualidade 100% segundo a avaliação dos pesquisadores, sendo os quatro do tipo **experimentação**.

Do ponto de vista quantitativo, verifica-se que existem artigos escritos que não apresentam qualquer tipo de verificação empírica. Sob o prisma da qualidade, a avaliação é a mesma: tem deixado a desejar, principalmente quando se constata que poucos trabalhos apresentam 100% de qualidade.

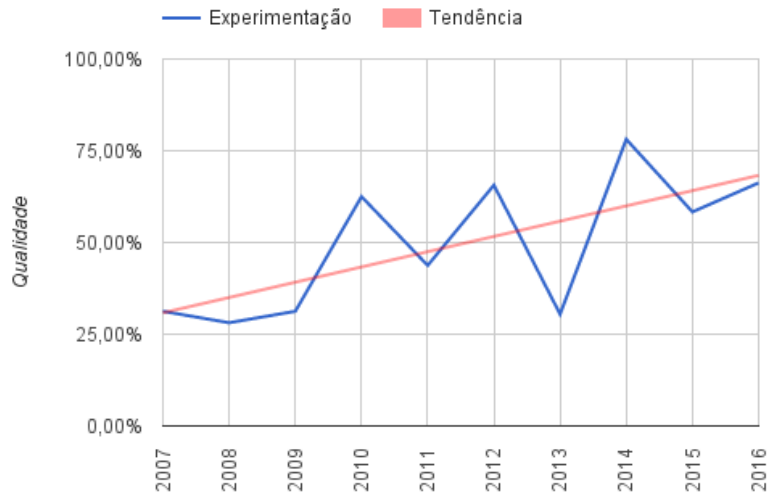


**Figure 2. Evolução da Qualidade dos Trabalhos**

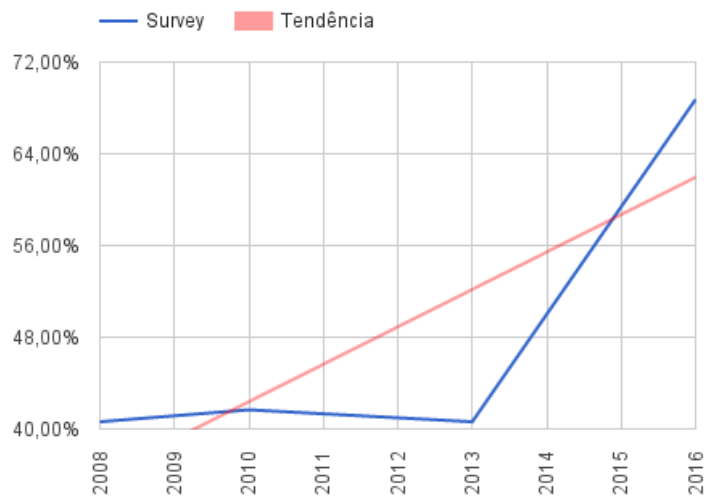


**Figure 3. Evolução da Qualidade do Estudo de Caso**

O estudo de Zannier, Melnik e Maurer [Zannier et al. 2006] não utiliza o termo "qualidade" na avaliação dos trabalhos, mas sim utiliza o que chama de "solidez" e a avalia em cada estudo considerando os seguintes critérios: 1) Deixa claro 4 parâmetros: tipo de estudo, tipo de amostragem, população alvo e população utilizada, e tipo de avaliação; 2) Implementa de forma adequada o uso de um método de análise dependendo das escalas de medida; e 3) Quando apropriado tem hipóteses bem definidas. De acordo com esses critérios, Zannier, Melnik e Maurer [Zannier et al. 2006] apontaram que a solidez dos trabalhos apresentados nos 29 anos de ICSE que foram por eles analisados não melhorou. Esta análise contrasta com a do presente estudo, o qual exhibe uma melhoria na qualidade.



**Figure 4. Evolução da Experimentação**



**Figure 5. Evolução do Survey**

## 5. Trabalhos Relacionados

Ao pesquisar sobre trabalhos relacionados foi identificado o grande interesse nos estudos empíricos para Engenharia de Software. Dentre eles, destaca-se o trabalho de Almeida [Almeida 2011], que realiza um mapeamento sistemático com o objetivo de levantar mecanismos para nortear os estudos empíricos em Engenharia de Software. Além deste, o trabalho de Gomes [Neto et al. 2011], que estabelece critérios relevantes sobre como classificar estudos empíricos em uma importante conferência de Engenharia de Software, também mostra-se relevante. Esses textos foram tomados como base para os *checklists* descritos na Seção 2.1 (Protocolo de Revisão).



Referente à metodologia usada, o presente artigo apresenta uma metodologia de pesquisa semelhante a do trabalho [Gonçalves et al. ], onde os autores investigaram a evolução dos trabalhos enviados à Conferência *Software Agents, Environments and Applications School (WESAAC)*, analisando os artigos publicados em suas 9 edições e adotando uma estratégia de pesquisa que combina estudo de escopo e boas práticas de revisão sistemática, com objetivo de compreender a evolução da comunidade Multi-Agent Systems.

Por fim, o artigo de Zannier, Melnik e Maurer [Zannier et al. 2006], que inspirou a presente pesquisa, apresenta uma avaliação da qualidade e quantidade dos artigos apresentados nas primeiras 29 edições do ICSE, tomando por base um extrato de 5% do total dos artigos. O trabalho aqui apresentado tem seu período de avaliação iniciado imediatamente após o último ano avaliado pelo trabalho de Zannier, Melnik e Maurer com objetivo de avaliar a sua evolução e atualizar as projeções. Cabe registrar que a forma de avaliação da qualidade dos trabalhos nesse estudo está calcada em critérios objetivos que permitem mensurar a qualidade, sendo em sua essência diferente da apresentada no trabalho de Zannier, Melnik e Maurer.

## 6. Ameaças à Validade

Ainda que a validação da classificação dos estudos tenha sido feita por dois pesquisadores, a decisão final dos conflitos foi tomada em consenso e não por um terceiro autor fora do contexto, o que pode levar a parte da amostra a receber uma classificação errônea.

A análise qualitativa da pesquisa depende de fatores relevantes que devem ser resguardados e averiguados corretamente com o intuito de não depreciarem a pesquisa ou diminuïrem seu coeficiente de acurácia. Entre eles a forma como os artigos são selecionados, buscando alcançá-los de forma aleatoriamente ponderada em razão de que não tendam para um grupo em específico e conseqüentemente conclusões erradas sejam extraídas dos dados levantados.

Em relação à validade externa, um fator expressivo diz respeito à quantidade mínima de dados da amostra. Neste caso, um percentual de 10% do total dos estudos encontrados foi definido. Apesar do processo ter sido realizado de forma aleatória, é possível que a amostra não represente fielmente o todo.

Embora tenha sido buscada a objetividade da avaliação de qualidade dos trabalhos através da definição de questões claras para ponderá-la com base em definições extraídas de referências reconhecidas na área, a análise de cada questão envolve a subjetividade do pesquisador, o que pode influenciar no valor final do percentual de qualidade do estudo.

Em relação à validade interna, houve replicação na análise envolvendo a primeira etapa do processo relativo à quantidade de artigos com estudos empíricos, porém não houve replicação na análise dos trabalhos na fase relativa à qualidade dos estudos.

## 7. Conclusões

O presente trabalho buscou responder às questões sobre quantidade e qualidade dos estudos empíricos realizados pelo ICSE nos últimos dez anos. De acordo com os resultados apresentados, ainda existe um número significativo de artigos que não apresentam nenhum tipo de validação empírica. No entanto, apresenta-se uma melhora nesta situação em relação aos resultados apresentados em [Zannier et al. 2006], já que a percentagem cai de 30% a 14%.

Em relação aos artigos que apresentaram estudo empírico de acordo com a especificação estabelecida para cada caso, de forma geral, evidenciaram uma tendência à melhoria quanto às suas respectivas questões de qualidade. No entanto, e apesar do aumento na qualidade dos estudos empíricos apresentados nos artigos, a qualidade na maioria dos casos ainda é questionável: o tipo de estudo que apresentou uma melhor avaliação em relação à qualidade foi o "estudo de caso", atingindo um 74%. Por outro lado, apenas quatro estudos experimentais foram considerados em completa conformidade com seu tipo, atingindo o 100% de qualidade. Desta forma, de acordo com os dados apresentados, é possível verificar que técnicas de experimentação em ES tem despertado o interesse para a sua utilização em pesquisas da área específica. No entanto, a comunidade precisa focar mais no emprego adequado de tais métodos de forma a garantir a confiabilidade nos resultados das pesquisas e subsidiar o crescimento da área.

## References

- Almeida, Adauto, e. a. (2011). Mecanismos para guiar estudos empiricos em engenharia de software: Um mapeamento sistematico. In *Proceedings of the Experimental Software Engineering Latin American Workshop*. ESELAW.
- Babbie, E. (1999). Métodos de pesquisa de survey. *EditoraUFMG*, 2(5):100–154.
- Gonçalves, E., Cortés, M., de Oliveira, M., Veras, N., Falcão, M., and Castro, J. Retrospective, relevance, and trends of software agents, environments and applications school (wesaac).
- Juristo, N. and Moreno, A. M. (2013). *Basics of software engineering experimentation*. Springer Science & Business Media.
- Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report*. EBSE.
- Malhotra, R. (2015). *Empirical Research in Software Engineering: Concepts, Analysis, and applications*. CRC Press, 1st edition.
- Neto, P. A. d. M. S., Cruzes, D. S., de Almeida, E. S., et al. (2011). 25 years of software engineering in brazil: An analysis of sbes history. In *Software Engineering (SBES), 2011 25th Brazilian Symposium on*, pages 4–13. IEEE.
- Random.Org. True random number service. <http://random.org>. Acessado em 16/11/2016.
- Runeson, P., Host, M., Rainer, A., and Regnell, B. (2012). *Case Study Research in Software Engineering: Guidelines and Examples*. Wiley Publishing, 1st edition.
- Tichy, W. F. (1998). Should computer scientists experiment more? *IEEE Computer*, 31(5):32–40.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.
- Zannier, C., Melnik, G., and Maurer, F. (2006). On the success of empirical studies in the international conference on software engineering. In *Proceedings of the 28th international conference on Software engineering*, pages 341–350. ACM.