



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Marconi Gomes da Rocha Filho

**Documentação de Ameaças à Validade em Estudos Empíricos de Engenharia
de Software no Brasil: Uma Revisão Sistemática**

Recife
2026

Marconi Gomes da Rocha Filho

**Documentação de Ameaças à Validade em Estudos Empíricos de Engenharia
de Software no Brasil: Uma Revisão Sistemática**

Monografia apresentada ao Curso de
Ciência da Computação da Universidade
Federal de Pernambuco, como requisito
para obtenção do título de Bacharel em
Ciência da Computação. Área de
concentração: Engenharia de Software.

Orientador(a): Sergio Castelo Branco Soares

Coorientador(a): Ivanildo Monteiro de Azevedo

Recife

2026

AGRADECIMENTOS

Agradeço primeiramente à minha família: Laurivânia, Djullie e Fifi, por, mesmo que fisicamente distantes, me apoiaram do início até o fim desta jornada. Obrigado por serem abrigo, apoio e carinho sempre que precisei; seja em território brasileiro ou estrangeiro, lhes amo.

Agradeço aos meus muitos amigos no que fiz no CIn. Alice, Amanda, Bruno, Conti, Eliab, Erick, Guilherme A., Guilherme S., Heitor, Igor, Íris, João, Josué, Luana, Lucas, Luiza, Luquet, Marcelo, Magnon, Melissa, Rodrigo, Teixeira, Thiago, Tiago, Tomás, Victor, Zezinho; Chegamos em tempos diferentes (e fui um dos últimos) todos ao fim, acreditamos uns nos outros, compartilhamos noites sem dormir, risadas, lágrimas, conhecimento e desespero durante esses anos. Tenho certeza de que sem cada um de vocês isso jamais teria sido possível, obrigado por terem se tornado o meu Recife seguro. Pelo ombro amigo, pelo suporte, pelas histórias, por tudo, levo vocês no coração para onde for.

Agradeço ao meu namorado, Marcos, que, quando estive frente às atividades acadêmicas pensando em desistir, ouviu minhas lamúrias, me ensinou a ter paciência, a desacelerar, e o balanço entre dedicação e descanso. Te amo.

Agradeço ao PET Informática, seus integrantes, egressos e tutores, que, com seu projeto de visitas, fez no ensino médio meus olhos brilharem em direção ao Centro de Informática da UFPE para ingressá-lo, onde anos depois tornei-me membro e pude fazer com muito orgulho o trabalho que um dia mudou minha vida.

Agradeço aos meus orientadores, Ivanildo e Sérgio, por serem a luz que deu clareza aos meus passos referente a este complexo trabalho durante meses. Levo seus ensinamentos comigo, meu legado é agora também partilhado convosco.

Agradeço aos meus mestres acadêmicos, professores que compartilharam com maestria vosso conhecimento, almejo um dia poder lecionar com a mesma destreza e dedicação que se propuseram quando fui seu aluno.

Agradeço aos meus colegas das parcerias academia/indústria no CIn, Apple Developer Academy e Projeto Motorola, para além de termos compartilhado momentos únicos, aprendi imensuravelmente sobre vida e carreira convosco.

Por fim, agradeço ao Governo Federal Brasileiro (2005, 2012), pela Lei das Cotas e pela Lei de institucionalização do Programa de Educação Tutorial, duas oportunidades que transformaram minha vida.

When everyone you thought you knew
Deserts your fight, I'll go with you
You're facing down a dark hall
I'll grab my light and go with you.
(Twenty One Pilots, 2018).

RESUMO

[Contexto] A Engenharia de Software Empírica usa e descreve métodos para investigar os muitos aspectos de software, como sua produção, processos e fenômenos relacionados, em que a porção empírica produz evidências confiáveis para o avanço do corpo de conhecimento científico dessa área, tal progresso só pode ser alcançado com rigorosos métodos de pesquisa a serem utilizados nos estudos que a compõem a fim de garantir alta confiabilidade e validade. Entre os diversos métodos empíricos estão as Ameaças à Validade, consideradas um elemento crítico em qualquer estudo empírico, entretanto recentes evidências (Schloegel et al., 2018) mostram que as Ameaças à Validade geralmente recebem atenção limitada em estudos empíricos. **[Objetivo]** Para quantificar, avaliar a qualidade da documentação e identificar a frequência das Ameaças à Validade no contexto da engenharia de software brasileira. **[Metodologia]** Foi desenvolvido um protocolo para conduzir uma Revisão Sistemática nos estudos empíricos primários publicados na trilha de pesquisa das últimas 5 edições de uma das maiores conferências de software brasileiras, o Simpósio Brasileiro de Engenharia de Software (SBES) a partir de 4 critérios de qualidade: Apresentação, Causas, Impactos e Mitigações. **[Resultados]** Os resultados mostram que, na trilha de pesquisa, 97,5% dos estudos apresentaram uma seção dedicada às Ameaças à Validade, entretanto há uma discrepância de 24% nas notas entre os critérios específicos. **[Conclusão]** Nas últimas 5 edições do SBES, outras trilhas que não a de pesquisa geralmente mostram práticas de documentação limitadas e tem a maior taxa de ausência de documentação. Na trilha de pesquisa, a simples presença dessa seção não garante que a documentação seja detalhada, a qualidade do relato de Ameaças à Validade é influenciada por múltiplas dimensões complementares. A descrição dos impactos parece desempenhar um papel importante na pontuação global de qualidade, além de que as dimensões parecem se comportar de forma independente ressaltando a existência de oportunidades de melhoria para padronizar os relatos.

Palavras-chave: Ameaças à Validade; Engenharia de Software; Revisão Sistemática; Metodologia Empírica.

ABSTRACT

[Context] Empirical Software Engineering uses and describes methods to investigate the many aspects of software, such as its production, processes, and related phenomena, where the empirical portion produces reliable evidence to advance the scientific body of knowledge of this field; such progress can only be achieved through rigorous research methods applied in the studies that compose it, in order to ensure high reliability and validity. Among the various empirical methods are Threats to Validity, considered a critical element in any empirical study, however recent evidence (Schloegel et al., 2018) shows that Threats to Validity generally receive limited attention in empirical studies. **[Objective]** To quantify, assess the quality of reports and identify the frequency of Threats to Validity in the context of Brazilian software engineering. **[Methodology]** A protocol was developed to conduct a Systematic Review of the primary empirical studies published in the research track of the last 5 editions of one of the largest Brazilian software conferences, the Brazilian Symposium on Software Engineering (SBES), based on 4 specific quality criteria: Presentation, Causes, Impacts, and Mitigations. **[Results]** The results show that, in the research track, 97.5% of the studies presented a section dedicated to Threats to Validity; however, there is a 24% discrepancy between the specific criteria scores. **[Conclusion]** In the last 5 editions of SBES, tracks other than the research track generally show limited reporting practices and have the highest rate of missing reports. In the research track, the mere presence of this section does not guarantee that the report is detailed; the quality of Threats to Validity reporting is influenced by multiple complementary dimensions. The description of impacts seems to play an important role in the overall quality score, and the dimensions seem to behave independently, highlighting the existence of opportunities for improvement to standardize the reports.

Keywords: Threats to Validity; Software Engineering; Systematic Review; Empirical Methodology.

LISTA DE ILUSTRAÇÕES

MONOGRAFIA

Figura 1 –	Visão geral do processo e contribuições esperadas.	23
Figura 2 –	Passo a passo do Protocolo de análise	31
Gráfico 1 –	Quantidade de Artigos por Ano (incluindo secundários)	34
Gráfico 2 –	Total de Artigos versus Método	35
Gráfico 3 –	Tipo de Estudo por Ano	36
Gráfico 4 –	Situação da Apresentação de Ameaças à Validade nos estudos	37
Gráfico 5 –	Pontuação de Qualidade das Ameaças à Validade	38
Gráfico 6 –	Mapa de Calor das Correlações	42

LISTA DE TABELAS

DISSERTAÇÃO

Tabela 1 –	Trilhas excluídas	33
Tabela 2 –	Distribuição da completude dos critérios de qualidade	39
Tabela 3 –	Interdependência entre os critérios	40
Tabela 4 –	Associações sem significância estatística	41

LISTA DE ABREVIATURAS E SIGLAS

SBES	Simpósio Brasileiro de Engenharia de Software
ICSE	International Conference on Software Engineering
TTVs	Threats to Validity
RS	Revisão Sistemática

LISTA DE FÓRMULAS

Fórmula 1 –	Percentual de Artigos sem documentação de Ameaças à Validade	29
Fórmula 2 –	Percentual de Artigos não empíricos	29
Fórmula 3 –	Percentual de Artigos por tipo num ano	29
Fórmula 4 –	Percentual Média das documentações de Ameaças à Validade	30
Fórmula 5 –	Correlação de Spearman	30

SUMÁRIO

1	INTRODUÇÃO	15
1.1	CONTEXTO	15
1.2	PROBLEMA	16
1.3	OBJETIVOS	17
1.4	QUESTÕES DE PESQUISA	17
1.5	CONTRIBUIÇÕES	19
2	TRABALHOS RELACIONADOS	19
2.1	PRINCIPAL TRABALHO	19
2.2	OUTROS TRABALHOS	20
3	METODOLOGIA	22
3.1	ETAPAS DA PESQUISA	22
3.2	JUSTIFICATIVA DE ESCOLHA DO SIMPÓSIO	24
3.3	ESTRATÉGIA DE BUSCA	24
3.3.1	Critérios de Inclusão	25
3.3.2	Critérios de Exclusão	25
3.3.3	Inclusão de estudos de metodologia mista	26
3.3.4	Estudos selecionados	26
3.4	CLASSIFICAÇÃO DOS ESTUDOS	27
3.5	ANÁLISE QUANTITATIVA	27
3.5.1	Métricas de Avaliação	28
3.5.2	Métodos de Análise dos dados	29
3.6	DESCRIÇÃO DO PILOTO	30
4	RESULTADOS	32
4.1	VISÃO GERAL DA TRILHA DE PESQUISA	33
4.2	COMPLETUDE DOS CRITÉRIOS DE QUALIDADE	37
4.3	DISCUSSÃO	43
5	AMEAÇAS À VALIDADE	46
5.1	VALIDADE INTERNA	46

5.2	VALIDADE EXTERNA	46
5.3	VALIDADE DE CONSTRUTO	47
5.4	VALIDADE DE CONCLUSÃO	48
5.5	TRABALHOS FUTUROS	49
6	CONSIDERAÇÕES FINAIS	50
	REFERÊNCIAS	51
	APÊNDICE A – ESTUDOS SELECIONADOS	53
	APÊNDICE B - CRITÉRIOS DE CLASSIFICAÇÃO DOS TIPOS DE ESTUDO	61
	APÊNDICE C - PROTOCOLO DE ANÁLISE DAS AMEAÇAS À VALIDADE NOS ESTUDOS SELECIONADOS	63

1 INTRODUÇÃO

1.1 CONTEXTO

Como pontuado por Verdecchia et al. (2023), a Engenharia de Software Empírica usa e descreve métodos com o objetivo de investigar os diversos aspectos do software, como sua produção, processos e fenômenos relacionados. A parte empírica deste campo é responsável por produzir evidências confiáveis para o desenvolvimento deste corpo científico de conhecimento. Entretanto, tal avanço só pode ser atingido com o uso de rigorosos métodos de pesquisa nos estudos que compõem este corpo.

A pesquisa empírica assegura a qualidade dos estudos por meio de uma série de métodos enfocados em validade e confiabilidade, sendo um deles a análise e documentação de Ameaças à Validade que, segundo descrito por Cruzes (2017), e mostrado por Gren (2018) é “um elemento crítico em qualquer estudo empírico, que deve ser contínuo durante todo o período do trabalho, desde o princípio nas primeiras concepções do estudo até o último passo dele.”

Evidências recentes sugerem que as Ameaças à Validade geralmente recebem atenção limitada e, em muitos casos, superficial nas pesquisas empíricas. Schloegel et al. (2024) apresentam que aproximadamente 80% das pesquisas publicadas nas principais conferências de Segurança Computacional e Engenharia de Software nem sequer incluem uma seção dedicada a discutir Ameaças à Validade. Em contrapartida, Lago et al. (2024) mostram que estudos premiados como “Artigo Destaque” do Grupo de Interesse Especial em Engenharia de Software da ACM (ACM SIGSOFT) na International Conference on Software Engineering (ICSE) mostram um padrão marcadamente distinto: apenas uma minoria (33%) desses artigos omitem uma seção específica para as Ameaças à Validade. Lago et al. (2024) ainda ressaltam que uma porção substancial dos estudos empíricos ainda falha ao explicitamente relatar e discutir suas Ameaças à Validade, reforçando a tendência geral que as documentações dessas ameaças ou são marginalizadas ou são completamente desconsideradas.

Essas evidências sugerem que, apesar da atenção completa e criteriosa sobre as Ameaças à Validade tender a ser o padrão em publicações altamente renomadas, ainda há inconsistências nas demais publicações da comunidade

acadêmica de Engenharia de Software Empírica. Similarmente, Verdecchia et al. (2023) e Gren (2018) criticam a maneira frequentemente superficial com a qual as Ameaças à Validade são tratadas, advogando pela sua reflexão antecipada durante toda a extensão e ciclo de vida de um estudo, ao contrário de se priorizá-las para as etapas finais do desenvolvimento dos trabalhos.

Apesar de existirem iniciativas para a padronização das Ameaças à Validade, como proposto através de taxonomias e estruturas prontas (Ampatzoglou et al., 2020), ou por meio de ferramentas de suporte como ValiDEPlan (Fonseca et al., 2019), ainda persiste uma lacuna entre a conscientização conceitual das Ameaças à Validade e a aproximação sistemática usada para aplicá-las na prática. Essa desconexão ressalta a necessidade de um guia metodológico mais claro e demanda uma mudança essencialmente cultural sobre o nível de rigor, consistência e engajamento sobre as preocupações com a validade dos estudos na Engenharia de Software Empírica (Siegmund et al., 2015; Teixeira et al., 2018).

Finalmente, trabalhos recentes de Azevedo et al. (2025) mostram que, apesar de existir um crescente interesse sobre a conscientização em relação às Ameaças à Validade em estudos de replicação, que são apenas uma porção dos estudos empíricos, ainda há uma desigualdade entre as documentações de Ameaças à Validade entre os tipos de estudos empíricos, o que indica um certo grau de imaturidade no desenho dos estudos de replicação na Engenharia de Software Empírica.

1.2 PROBLEMA

Ameaças à Validade são amplamente reconhecidas como um componente fundamental de rigor metodológico na Engenharia de Software Empírica. Identificação, classificação, discussão e ações de mitigação apropriadas para as Ameaças à Validade são essenciais para garantir transparência, interpretabilidade e credibilidade aos resultados empíricos. Apesar de estruturas prontas existirem e serem amplamente utilizadas na literatura, estudos anteriores sugerem que a documentação das Ameaças à Validade é inconsistente, desigual ou insuficientemente detalhada.

Enquanto pesquisas recentes examinaram a documentação de Ameaças à Validade em contextos específicos, como estudos de replicação, ainda persiste uma

lacuna de evidências sistemáticas sobre como as Ameaças à Validade são documentadas em espaços de pesquisa empírica mais amplos. Em particular, nenhuma avaliação foi conduzida para perceber o grau de qualidade e compreensividade das Ameaças à Validade no fórum primário de pesquisas em Engenharia de Software, o Simpósio Brasileiro de Engenharia de Software (SBES).

Essa lacuna evidencia uma importante questão sobre o atual estado de maturidade metodológica neste fórum: Até que ponto os estudos empíricos do SBES abordam sistematicamente e rigorosamente as Ameaças à Validade?

1.3 OBJETIVOS

Avaliar sistematicamente a qualidade da documentação, rigor de classificação e explicitação de ações de mitigação das Ameaças à Validade em estudos empíricos publicados no Simpósio Brasileiro de Engenharia de Software.

Objetivos específicos

- Quantificar a prevalência da documentação de Ameaças à Validade em estudos empíricos publicados no SBES.
- Avaliar a completude e o rigor metodológico da documentação de Ameaças à Validade usando critérios específicos de qualidade.
- Analisar potenciais diferenças e associações entre tipos de estudos empíricos e a qualidade da documentação de Ameaças à Validade.

1.4 QUESTÕES DE PESQUISA

A pergunta principal de pesquisa é: **“Qual é a qualidade da documentação das Ameaças à Validade dos estudos empíricos publicados no SBES?”**

As perguntas a seguir refinam o inquérito principal especificando as dimensões pelas quais a qualidade da documentação das Ameaças à Validade foi examinada.

- **QP1:** Qual a proporção de estudos empíricos do SBES que explicitamente documenta Ameaças à Validade?

- **Justificativa:** Avaliar a presença ou ausência das Ameaças à Validade nos estudos empíricos é uma etapa fundamental para entender a maturidade geral das metodologias práticas deste fórum. A inclusão explícita de uma seção de Ameaças à Validade reflete uma conscientização do rigor metodológico e da transparência, servindo como um indicador inicial da qualidade da documentação. Ao quantificar a frequência das Ameaças à Validade ao longo dos estudos analisados, esta análise fornecerá uma base de mensuração sobre o quanto consistentemente as considerações sobre validade são abordadas nas publicações do SBES. Tal análise contribui para um maior entendimento da relação entre desenho do estudo e práticas de documentação de validade, dando base para futuras discussões sobre maturidade metodológica baseadas em evidências neste fórum.
- **QP2:** Qual a completude da documentação de Ameaças à Validade em termos de causas, impactos e estratégias de mitigação?
 - **Justificativa:** Além de meramente identificar a presença de Ameaças à Validade, é essencial avaliar a profundidade e compreensividade pela qual elas foram relatadas; A completude da documentação ao descrever as Ameaças à Validade, incluindo suas causas, potenciais impactos e estratégias de mitigação, serve como um melhor indicador de rigor metodológico do que o simples reconhecimento. Ao operacionalizar e medir a profundidade da documentação por meio de critérios de qualidade pré-definidos, esta análise permite uma avaliação estruturada da qualidade dos relatos de Ameaças à Validade dentre todos os estudos revisados. Integrar este indicador de qualidade no escopo geral da análise contribuirá para uma avaliação mais abrangente sobre a conscientização metodológica e transparência neste fórum.
- **QP3:** Como os critérios de qualidade das documentações de Ameaças à Validade se relacionam entre si nos estudos empíricos do SBES?
 - **Justificativa:** Além de tratar os critérios de qualidade como indicadores isolados, os critérios pré-definidos nesta análise foram combinados para criar uma pontuação de qualidade geral da documentação de Ameaças à Validade nos estudos. Avaliar como

estes critérios se relacionam entre si é essencial para determinar se eles genuinamente capturam dimensões diferentes das práticas de documentação ou se refletem uma única tendência em direção ao rigor metodológico. Ao analisar estas associações, bem como a relação de cada critério com a pontuação geral de qualidade, esta questão revela a estrutura interna da qualidade de documentação das Ameaças à Validade neste fórum. Ao identificar quais critérios tendem a co-ocorrer e quais critérios são abordados independentemente dos outros, produzirá indicadores sobre se os autores que relatam um aspecto também relatam os demais com rigor comparável, além de indicar se alguns desses critérios são sistematicamente priorizados ou não.

1.5 CONTRIBUIÇÕES

Neste sentido, este trabalho irá contribuir para um primeiro diagnóstico da qualidade da documentação das Ameaças à Validade no SBES a nível de simpósio, gerar novas evidências empíricas sobre a associação entre os desenhos de estudo e a maturidade da validade, e finalmente servir como base longitudinal para avaliação do rigor metodológico de trabalhos futuros.

2 TRABALHOS RELACIONADOS

Esta seção apresenta os trabalhos que se relacionam com o estudo que foi feito, mostrando um resumo do que a literatura recente apresentou e como tais trabalhos complementam, se relacionam e possuem lacunas em relação a este trabalho.

2.1 PRINCIPAL TRABALHO

Azevedo et al. (2025) realizaram uma revisão sistemática de literatura sobre 83 estudos de replicação publicados entre 2022 e 2024, recuperados da base SCOPUS, com o objetivo de analisar como as Ameaças à Validade são

documentadas em estudos de replicação na Engenharia de Software Empírica e seus respectivos estudos originais. A análise considerou as quatro dimensões de validade propostas pela taxonomia de Wohlin et al. (2012). Os resultados indicam que estudos de replicação documentam ameaças com maior frequência e detalhamento que os estudos originais, principalmente quanto à validade externa e interna, enquanto as validades de construto e de conclusão continuam sub-relatados. Experimentos controlados abordam as quatro dimensões de forma mais abrangente, enquanto *surveys* e estudos de caso mostram um tratamento mais estrito e associações significativas com Ameaças à Validade de conclusão.

Apesar desta pesquisa compartilhar similaridades com a de Azevedo et al. (2025), no sentido de avaliar a presença e as diferenças de documentação de Ameaças à Validade entre tipos de estudo, a pesquisa mencionada restringe seu escopo a estudos de replicação em base internacional, enquanto esta pesquisa abrangerá estudos empíricos primários independente de serem replicações e no contexto da Engenharia de Software brasileira, além de avaliar a completude e relações internas das características de documentação as quais não foram contempladas na análise daqueles autores.

2.2 OUTROS TRABALHOS

Lago et al. (2024) conduziram uma revisão sistemática da literatura em 91 artigos premiados com o *ACM SIGSOFT Distinguished Paper Award* na Conferência Internacional em Engenharia de Software (ICSE), no período de 2014 a 2023, com o objetivo de caracterizar sistematicamente o estado atual das práticas de análise e documentação de Ameaças à Validade na pesquisa em Engenharia de Software. Os autores extraíram dados sobre seis dimensões principais, incluindo a profundidade da reflexão, o uso de checklists ou diretrizes, as categorias de validade discutidas, a adequação da categorização ao tipo de estudo, dentre outros. Os resultados corroboram críticas apresentadas por outros autores: uma pequena porção dos artigos discutem as ameaças em profundidade, 33% não possuem uma seção dedicada para os relatos, e 61,5% não apresentam estratégias de mitigação. Os autores concluem que as Ameaças à Validade são tratadas majoritariamente como uma exigência tardia, não sendo consideradas com a devida importância desde o início do desenho da pesquisa, e finalmente propõem considerações destinadas a

pesquisadores, revisores e leitores quanto a uma chamada para ação da comunidade.

Apesar de muito semelhantes, principalmente na forma de avaliar a qualidade da documentação em um fórum delimitado de publicações (ICSE e SBES), esta pesquisa investiga o principal fórum brasileiro de Engenharia de Software, além de usar um modelo de avaliação composto de critérios com análises estatísticas, onde Lago et al. (2024) predominantemente apresentaram uma abordagem descritiva e qualitativa.

Verdecchia et al. (2023) apresentam um artigo de posicionamento que, com base nas suas experiências como pesquisadores, autores e revisores de artigos, refletem criticamente sobre o atual estado das seções de Ameaças à Validade na literatura de Engenharia de Software Empírica. Os autores identificam oito problemas recorrentes nas seções anteriormente mencionadas, entre eles o uso tratamento da seção como meros checklists e a adoção cega de uma taxonomia de categorização que pode não ser adequada para o método de pesquisa usado. Para cada problema os autores propõem uma visão de melhoria e defendem que as Ameaças à Validade sejam consideradas em todas as fases da pesquisa, desde o desenho do estudo, e que os métodos estruturados, como checklists, sirvam como um norte para discussões mais aprofundadas em vez de uma restrição metodológica.

Enquanto Verdecchia et al. (2023) formulam o problema de forma conceitual e opinativa, sem uma coleta de dados sistemática, este trabalho fornecerá evidências empíricas que permitirão verificar em que medida os problemas de documentação se manifestam no contexto brasileiro. Os critérios de qualidade aqui definidos, como apresentação das Ameaças à Validade e explicitação de ações de mitigação, tratam diretamente algumas das preocupações explicitadas por aqueles autores, permitindo avaliar quantitativamente fenômenos que o artigo de posicionamento descreve apenas qualitativamente.

Gren (2018) apresenta um artigo conceitual fundamentado na teoria de testes psicológicos, com o objetivo de provocar uma mudança de mentalidade na forma como a validade é tratada na pesquisa comportamental na Engenharia de Software. O autor argumenta que a taxonomia clássica popularizada por Wohlin et al. (2012) e

Runeson et al. (2012) frequentemente se converteram em um “paliativo rápido” ao final dos artigos, sem a devida profundidade teórica. Gren (2018) propõe um checklist integrado e hierárquico que detalha a validade de construto em seis aspectos, tendo a confiabilidade como pré-requisito. O autor também discute as raízes psicológicas de baixa qualidade das análises de validade e defende que estudos de validação sejam mais valorizados pela comunidade do que a mera construção de ferramentas.

Gren (2018) oferece uma fundamentação teórica para a crítica à superficialidade dos relatos de Ameaças à Validade, que é usada neste estudo como premissa para não só avaliar a presença mas também a qualidade da documentação. Diferente de apresentar uma proposta como o estudo de Gren (2018), este trabalho pretende mensurar empiricamente o rigor de apresentação e completude dos relatos, fenômenos cuja deficiência Gren (2018) antecipou em nível conceitual.

3 METODOLOGIA

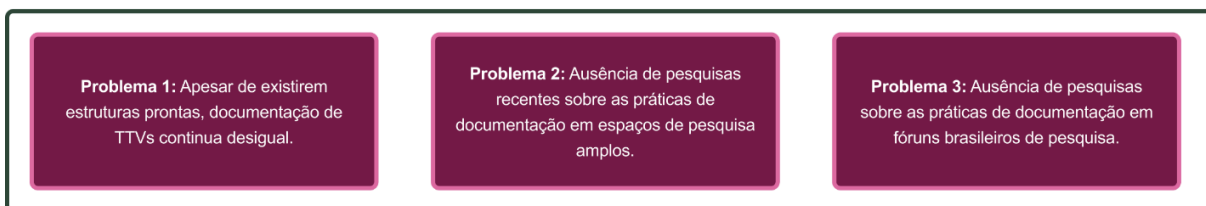
Este estudo conduz uma revisão sistemática para avaliar a qualidade das Ameaças à Validade documentadas em estudos empíricos publicados no SBES. Este estudo centra-se nos relatórios efetivamente apresentados pelos autores dos trabalhos analisados em vez de centrar-se nas possibilidades que os autores poderiam ter apresentado. A qualidade dos relatórios apresentados é avaliada através de um modelo composto que abrange a presença de documentação de Ameaças à Validade, o rigor de classificação utilizado, a profundidade da descrição e a explicitação de estratégias de mitigação. Este estudo se posiciona como um diagnóstico exploratório da maturidade metodológica do principal fórum brasileiro de pesquisa em Engenharia de Software.

3.1 ETAPAS DA PESQUISA

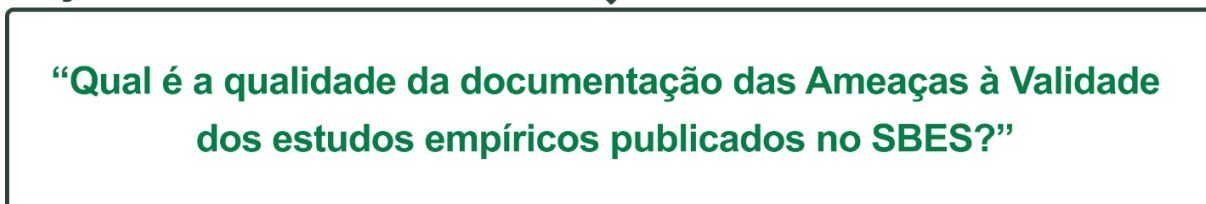
A pesquisa foi conduzida de acordo com as etapas na figura 1, seguida das suas contribuições esperadas com os resultados:

Figura 1 – Visão geral do processo e contribuições esperadas.

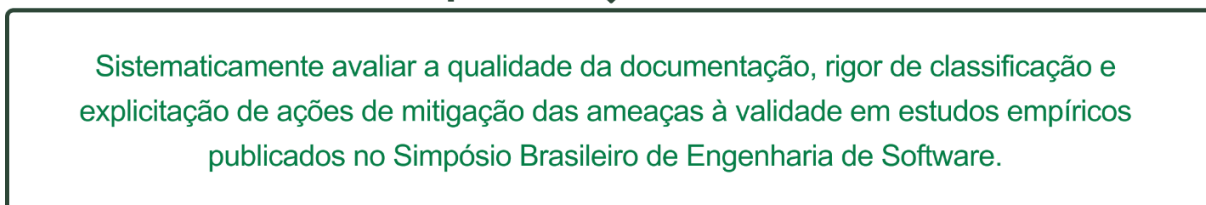
Problemas



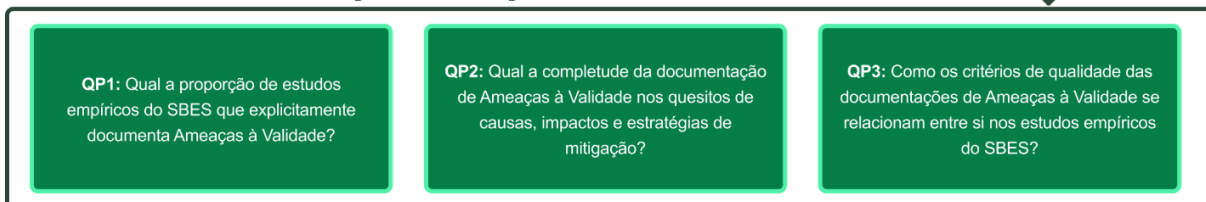
Objetivo



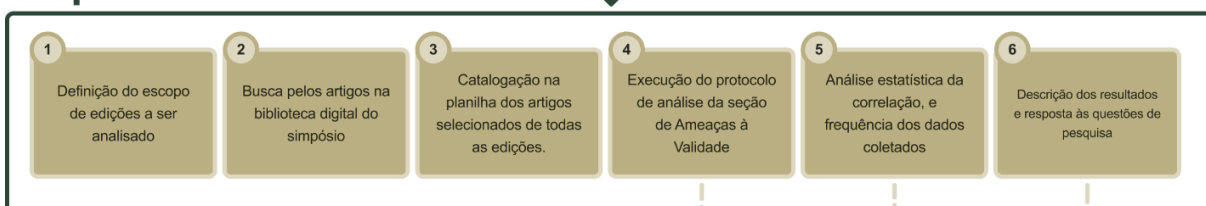
Questão Geral de Pesquisa



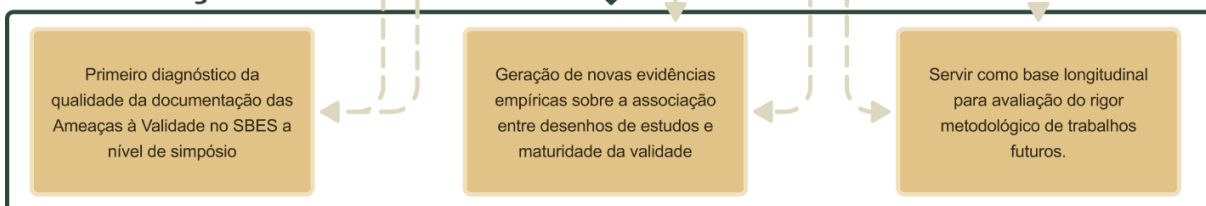
Questões de Pesquisa Específicas



Etapas



Contribuições



Fonte: Elaborado pelo autor (2026).

3.2 JUSTIFICATIVA DE ESCOLHA DO SIMPÓSIO

Este trabalho foca no SBES por conta do seu papel como o principal fórum nacional para as pesquisas em Engenharia de Software no Brasil e como um dos principais espaços para a área na comunidade Latino Americana. O SBES tem mantido historicamente uma ênfase em pesquisas empíricas, fazendo com que seja uma amostra representativa para o contexto de examinação das práticas relacionadas às Ameaças à Validade.

Indo além, o SBES serviu como um palco de discussões sobre rigor metodológico e preocupações com a validade, como evidenciado por trabalhos anteriormente publicados nele (Teixeira et al., 2018; Fonseca et al., 2019; Azevedo et al., 2025). Isso indica que a comunidade está ativamente engajada em debates que perpassam os temas de qualidade empírica e transparência de pesquisa.

Ao concentrar a análise no SBES, este estudo provém um diagnóstico a nível de simpósio sobre as práticas de documentação de Ameaças à Validade em uma comunidade bem-estabelecida de pesquisa. Tal avaliação contribui para o entendimento da maturidade metodológica das pesquisas na Engenharia de Software Empírica no Brasil, além de oferecer uma base estruturada para futuras comparações longitudinais ou entre simpósios.

3.3 ESTRATÉGIA DE BUSCA

Este estudo adota uma estratégia de identificação a nível de simpósio. Uma vez que todos os artigos publicados no SBES estão disponíveis publicamente na biblioteca digital da Sociedade Brasileira de Computação, o corpo de análise foi preenchido por artigos recuperados desta biblioteca, disponível em <https://sol.sbc.org.br/index.php/sbes/issue/archive>.

Uma vez que o escopo deste estudo é explicitamente delimitado a um fórum e conjunto de edições predeterminados, não é necessário uma *string* de busca. Em vez disso, o processo de identificação consiste em sistematicamente listar todos os artigos das edições predeterminadas seguido de um filtro para manter apenas aqueles que satisfazem as condições de inclusão e não satisfazem as condições de exclusão.

Desta forma garantimos a cobertura de toda a amostra alvo dentro dos limites do fórum, eliminando riscos associados com pesquisas baseadas em palavras chaves.

3.3.1 Critérios de Inclusão

Os estudos publicados foram incluídos na análise quando satisfizeram todos os seguintes critérios:

1. O estudo foi publicado na trilha de pesquisa do Simpósio Brasileiro de Engenharia de Software.
2. O estudo pertence a uma das edições do SBES entre 2021 e 2025, período selecionado para prover uma análise contemporânea das práticas de documentação de Ameaças à Validade.

3.3.2 Critérios de exclusão

Os estudos publicados foram excluídos da análise quando satisfizeram pelo menos uma das seguintes condições:

1. O estudo não foi publicado nas edições dentro do período temporal especificado.
2. Estudos que, após análise prévia, não foram considerados adequados para as diretrizes desta análise devido à sua inerente natureza. Como por exemplo, quando percebe-se que na trilha de ferramentas as Ameaças à Validade e limitações são mais relacionadas à ferramenta em si do que ao estudo sendo realizado. Por isso, incluir tais estudos introduziria grande potencial de prejudicar os resultados finais da análise devido a sua alta taxa de ausência de documentações das Ameaças à Validade. Portanto, decidiu-se considerar apenas estudos publicados e apresentados na trilha de Pesquisa do SBES.
3. Estudos que não envolvam a coleta e análise de dados qualitativos ou quantitativos. Esta seção inclui artigos puramente teóricos, artigos de opinião, estruturas conceituais sem validação empírica, e propostas metodológicas sem validação empírica.

4. Estudos secundários e/ou terciários como revisões sistemáticas de literatura, estudos de mapeamento, *surveys* de literatura ou meta-análises.
5. Artigos curtos, demonstrações de ferramentas, artigo de simpósio de doutorado, relatórios da indústria, resumos de apresentações, tutoriais e resumos estendidos.

3.3.3 Inclusão de estudos de Metodologia Mista

Estudos de Métodos Mistos não foram excluídos desta análise. Na engenharia de software empírica, o desenho dos estudos frequentemente combina métodos qualitativos e quantitativos a fim de investigar fenômenos complexos, ao excluir tais estudos, resultaria em uma representação incompleta das práticas de pesquisa empírica neste fórum.

Dado que Ameaças à Validade são relevantes independentemente se o estudo adota um desenho exclusivamente quantitativo, qualitativo ou de métodos mistos, a sua inclusão é metodologicamente justificada. Estudos de Métodos Mistos estão sujeitos a preocupações quanto à validade que podem expandir-se em múltiplas dimensões (interno, externo, construto e conclusão), fazendo com que sejam particularmente relevantes para uma avaliação compreensiva das práticas de documentação.

Para garantir consistência analítica, Estudos de Métodos Mistos foram classificados de acordo com critérios pré-definidos e foram tratados como uma categoria distinta na análise dos dados. Desta forma, foi evitado o viés metodológico enquanto se preserva a integridade das comparações entre diferentes tipos de estudos empíricos.

3.3.4 Estudos selecionados

A lista de estudos selecionados que atendem aos critérios de inclusão impostos anteriormente pode ser encontrada no apêndice A.

3.4 CLASSIFICAÇÃO DOS ESTUDOS

Neste trabalho, os tipos de estudos foram classificados de acordo com o checklist disponível no apêndice B, independentemente da auto-classificação do estudo. Assim, é possível conduzir uma análise padronizada por meios específicos e apropriadamente embasar os resultados que responderão às questões de pesquisa. Esta lista de tipos de estudos empíricos foi baseada nos trabalhos de Easterbrook et al. (2008) e refinada para mais apropriadamente atingir os objetivos desta pesquisa, como por exemplo, adicionar relatos de experiências que são comumente encontrados nos estudos empíricos de engenharia de software.

A tabela no apêndice B, para cada linha, o tipo de estudo é descrito na coluna esquerda e a lista de critérios de características na coluna direita. O estudo analisado precisa possuir a maioria das características da coluna direita para ser considerado um estudo do tipo da coluna esquerda.

É sabido que alguns tipos de estudos, como notavelmente Etnografias, podem resultar em um número nulo de aparições dado o checklist de classificação apresentado e a amostra escolhida. Isto reflete a preferência específica de alguns tipos de estudos em relação a outros na comunidade científica brasileira.

3.5 ANÁLISE QUANTITATIVA

A análise quantitativa usada neste estudo tem como objetivo caracterizar e examinar padrões nas documentações de Ameaças à Validade nos estudos selecionados. Inicialmente, estatísticas descritivas foram usadas para sumarizar a presença de documentação das ameaças, além de construir os indicadores gerais de qualidade derivados dos critérios de avaliação específicos descritos na subseção 3.5.1.

Para explorar em maior profundidade potenciais relações no conjunto de dados, técnicas de análise estatística inferencial foram utilizadas para avaliar possíveis associações entre os tipos de estudos empíricos e a qualidade da documentação das Ameaças à Validade. Quando apropriado, procedimentos estatísticos não-paramétricos foram considerados, dado o tamanho da amostra e a natureza ordinal de alguns indicadores de qualidade.

Esta análise foca em identificar padrões, variações e potenciais diferenças nas práticas de documentação de Ameaças à Validade; Estes resultados tem como

objetivo prover uma base de avaliação estruturada e resultando em evidências empíricas das tendências de documentação e nível de rigor metodológico.

3.5.1 Métricas de Avaliação

A qualidade da documentação de Ameaças à Validade foi definida por quatro critérios de qualidade à medida que a seção do estudo for avaliada. Cada critério varia entre 0 e 1 em pontuação, e são descritos a seguir:

- 1. Apresentação das Ameaças à Validade:** O estudo analisado apresenta suas Ameaças à Validade?
 - a. Sim (1), documentou as ameaças identificadas usando o método de categorização de Cook e Campbell (1979) ou algum método formal equivalente e amplamente conhecido na Engenharia de Software Empírica.
 - b. Parcialmente (0,5), documentou as ameaças identificadas usando o método de categorização de Cook e Campbell (1979) mas utilizou o método incorretamente (como por exemplo categorização incorreta das Ameaças à Validade), ou documentou as ameaças identificadas entretanto sem o uso de uma classificação formal.
 - c. Não (0), não houve seção dedicada para a documentação das Ameaças à Validade.
- 2. Relatório de causas:** O estudo analisado documenta as causas das Ameaças à Validade identificadas?
 - a. Sim (1), documentou todas as causas para todas as ameaças identificadas.
 - b. Parcialmente (0,5) documentou as causas para algumas, mas não todas, as Ameaças à Validade identificadas.
 - c. Não (0), não documentou nenhuma causa para as Ameaças à Validade identificadas.
- 3. Relatório de impactos:** O estudo analisado documenta os impactos das Ameaças à Validade identificadas no estudo?

- a. Sim (1), documentou todos os impactos no estudo para todas as ameaças identificadas.
- b. Parcialmente (0,5) documentou os impactos no estudo para algumas, mas não todas, as Ameaças à Validade identificadas.
- c. Não (0), não documentou nenhum impacto no estudo para as Ameaças à Validade identificadas.

4. Relatório de ações de mitigação: O estudo analisado documenta ações de mitigação para Ameaças à Validade identificadas?

- a. Sim (1), documentou ações de mitigação para todas as ameaças identificadas.
- b. Parcialmente (0,5) documentou ações de mitigação para algumas, mas não todas, as Ameaças à Validade identificadas.
- c. Não (0), não documentou nenhuma ação de mitigação para as Ameaças à Validade identificadas.

3.5.2 Métodos de Análise dos dados

Foi usada análise de frequência simples para a análise de porcentagem de frequência das primeiras impressões sobre os dados, descritas pelas seguintes fórmulas:

Fórmula 1 - Percentual de Artigos sem documentação de Ameaças à Validade

$$PAsd = \frac{\sum Asd}{TEt}$$

Fonte: Elaborado pelo Autor (2026).

Fórmula 2 - Percentual de Artigos não empíricos

$$PAne = \frac{\sum Ane}{TEt}$$

Fonte: Elaborado pelo Autor (2026).

Fórmula 3 - Percentual de Artigos por tipo num ano

$$PAa = \frac{\sum ATa}{TEa}$$

Fonte: Elaborado pelo Autor (2026).

Onde Asd é o número de artigos sem documentação de Ameaças à Validade da trilha sendo analisada, Ane é o número de artigos não empíricos da trilha, TEt é o número de artigos totais da trilha, ATa é o número de artigos por tipo na edição sendo analisada e TEa é o total de artigos na edição sendo analisada.

Para o Gráfico 5, foi usada a pontuação média e, como cada critério pode receber apenas 0, 0,5 ou 1, o cálculo foi feito como descrito na Fórmula 4.

Fórmula 4 - Pontuação Média das documentações de Ameaças à Validade.

$$Pontuação\ Média = \frac{\sum pontuações\ do\ critério}{números\ de\ estudos}$$

Fonte: Elaborado pelo autor (2026).

Na análise de correlação feita no Gráfico 6, foi usada a correlação de Spearman, que mede a força e a direção da associação entre duas variáveis. Este método foi usado porque é o ideal para dados quantitativos que não seguem uma distribuição normal, para analisar a correlação entre os critérios de qualidade entre si, e entre a pontuação total, como mostrado na Fórmula 5:

Fórmula 5 - Correlação de Spearman

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

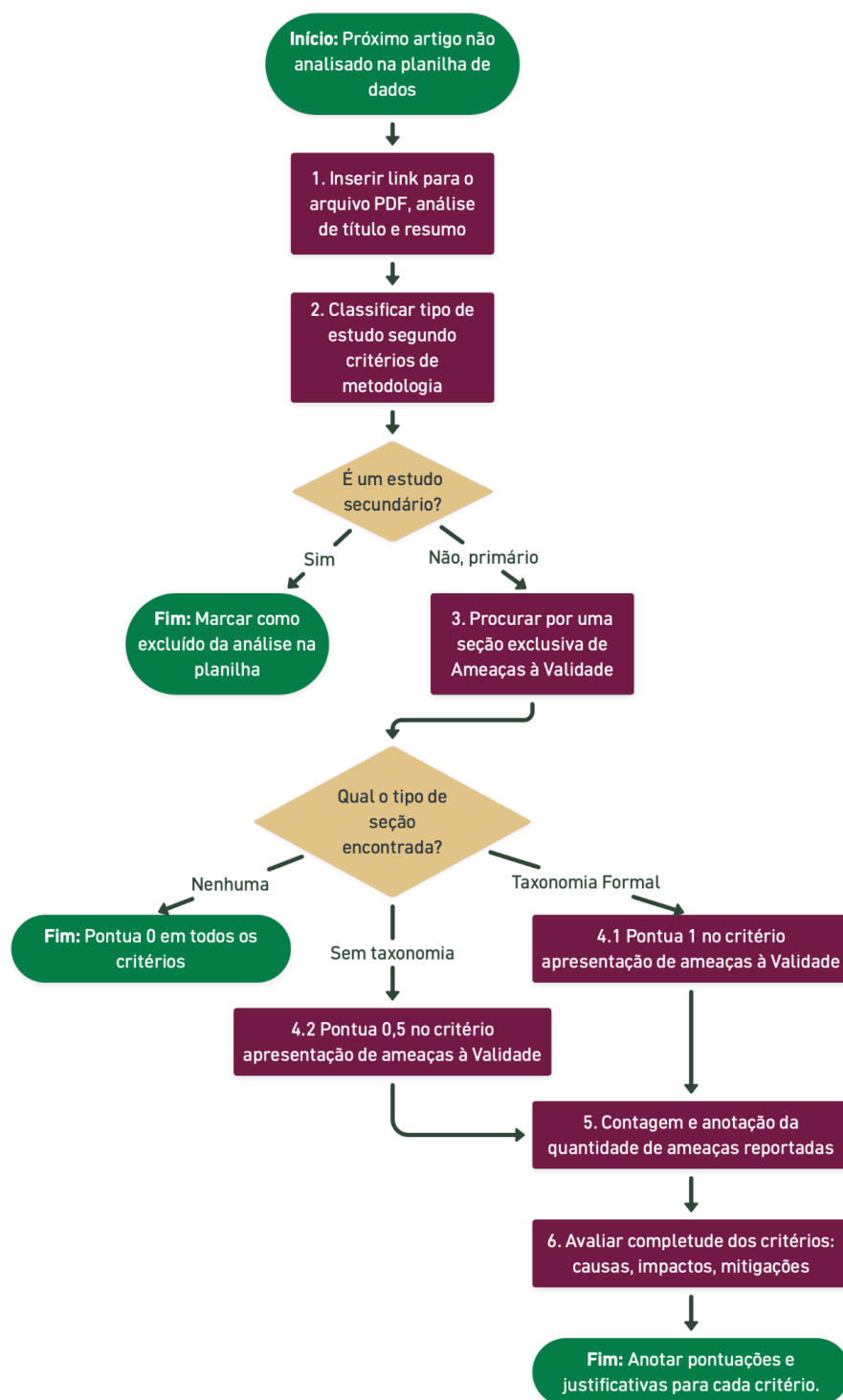
Fonte: Elaborado pelo autor (2026).

Onde d_i é a diferença entre os "postos" (posições) das duas variáveis para cada par e n é o número de observações ou pares de dados.

3.6 DESCRIÇÃO DO PILOTO

Na fase inicial do estudo, um piloto foi desenvolvido para maturar o processo de avaliação dos estudos selecionados com base nos critérios de qualidade definidos na subseção 3.4.1. O piloto foi executado preliminarmente na edição mais recente do SBES (2025) e subsequentemente aplicado nas edições anteriores com as etapas refinadas pelas lições aprendidas na execução. As etapas são descritas a seguir na figura 2, e detalhadamente no protocolo em anexo no apêndice C.

Figura 2 – Passo a passo do Protocolo de análise



Fonte: Elaborado pelo autor (2026).

4 RESULTADOS

No escopo inicial desta revisão, que contemplava todas as trilhas do SBES entre 2021 e 2025, uma análise preliminar realizada durante a execução do protocolo indicou que algumas trilhas apresentavam características que inviabilizavam uma análise robusta das Ameaças à Validade dos artigos (Veja a Tabela 1). Em particular, as trilhas Educação, Ferramentas, Indústria, Ideias Inovadoras e Resultados Emergentes apresentaram uma elevada incidência de artigos sem seção dedicada a Ameaças à Validade, além da presença de estudos não empíricos em algumas delas.

A decisão de restringir o escopo foi motivada principalmente pelas evidências observadas nas edições de 2024 e 2025. Nas trilhas Educação e Ideias Inovadoras e Resultados Emergentes, foi identificada uma elevada proporção de artigos que não apresentavam qualquer discussão sobre Ameaças à Validade. Em 2025, 8 dos 20 artigos da trilha Educação (40%) e 13 dos 32 artigos da trilha Ideias Inovadoras e Resultados Emergentes (41%) não continham relatos de Ameaças à Validade. Em 2024, esse cenário também se mostrou expressivo, com 4 dos 19 artigos da trilha Educação (21%) e 9 dos 22 artigos da trilha Ideias Inovadoras e Resultados Emergentes (41%) sem qualquer discussão sobre Ameaças à Validade. Além disso, foram identificados estudos não empíricos nessas trilhas, correspondendo a 10% (2/20) dos artigos da trilha Educação e 3% (1/32) da trilha Ideias Inovadoras e Resultados Emergentes em 2025, bem como 11% (2/19) e 14% (3/22), respectivamente, em 2024.

Nas trilhas Ferramentas e Indústria, na análise preliminar em 2025, observou-se que os artigos não apresentavam seções estruturadas de Ameaças à Validade compatíveis com os objetivos desta pesquisa. Como a avaliação proposta depende da análise da apresentação das ameaças, da descrição de suas causas, impactos e estratégias de mitigação, a ausência dessas informações impossibilitava a aplicação dos critérios de qualidade definidos no protocolo. Em conjunto, esses fatores comprometeriam a realização da análise entre os estudos selecionados.

Tabela 1 – Trilhas excluídas

Ano	Trilha	Total de Artigos	Não possui Ameaças à Validade	Artigos não empíricos
2025	Educação	20	8 (40%)	2 (10%)
2025	Ferramentas	22	-	-
2025	Indústria	6	-	-
2025	Ideias Inovadoras e Resultados Emergentes	32	13 (41%)	1 (3%)
2024	Educação	19	4 (21%)	2 (11%)
2024	Ideias Inovadoras e Resultados Emergentes	22	9 (41%)	3 (14%)

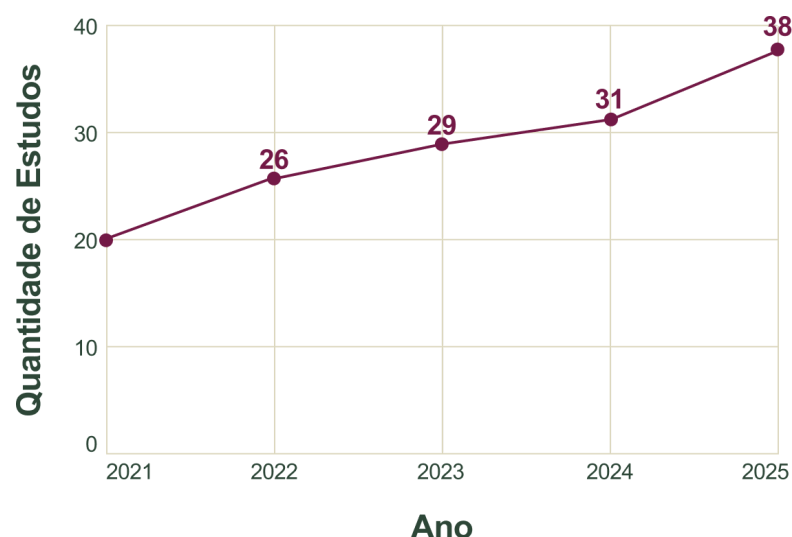
Fonte: Elaborado pelo autor (2026).

Dessa forma, optou-se por concentrar a revisão exclusivamente na trilha de Pesquisa do SBES. Após a identificação desse padrão nas edições de 2024 e 2025, decidiu-se aplicar o mesmo recorte às edições de 2021, 2022 e 2023, realizando a coleta de dados apenas na trilha de Pesquisa. Como consequência, as trilhas Educação, Ferramentas, Indústria e Ideias Inovadoras e Resultados Emergentes foram excluídas da análise em todo o período investigado. Essa decisão permitiu preservar a consistência metodológica da amostra e assegurar uma base de estudos mais adequada para a investigação da qualidade dos relatos de Ameaças à Validade em estudos empíricos de Engenharia de Software.

4.1 VISÃO GERAL DA TRILHA DE PESQUISA

O Gráfico 1 apresenta a distribuição temporal dos 144 estudos totais, incluindo os secundários. Observa-se um crescimento contínuo no número de estudos publicados na trilha de Pesquisa do SBES ao longo do período analisado, passando de 20 estudos em 2021 para 38 em 2025. Esse aumento corresponde a um crescimento de aproximadamente 90% no volume de estudos incluídos na amostra, sugerindo uma ampliação da produção científica empírica publicada na trilha de Pesquisa do evento nos últimos anos.

Gráfico 1 – Quantidade de Artigos por Ano (incluindo secundários)

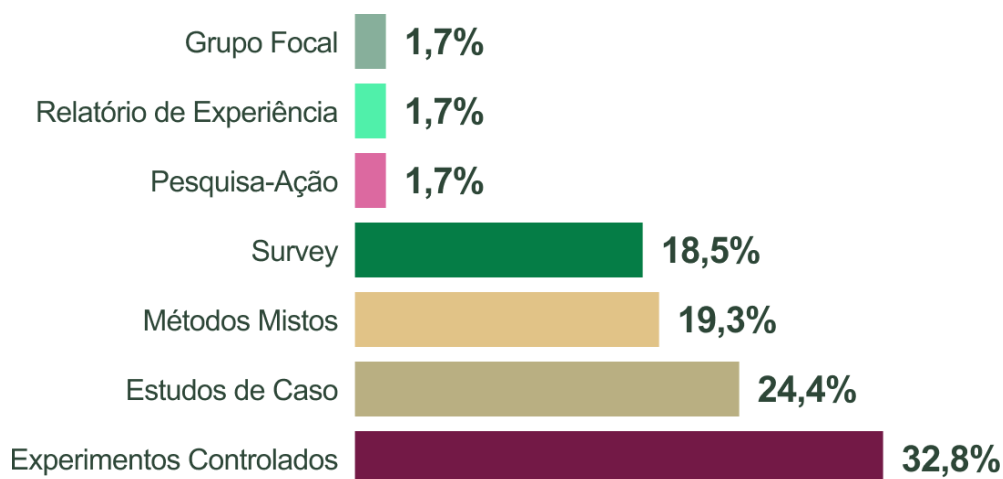


Fonte: Elaborado pelo autor (2026).

A aplicação dos critérios de inclusão e exclusão resultou na seleção de 119 estudos primários para análise, enquanto 25 estudos secundários foram excluídos por não atenderem ao escopo definido neste estudo. A distribuição dos estudos selecionados por método de pesquisa é apresentada no Gráfico 2. Observa-se uma predominância de Experimento Controlado com 39 estudos (32,8%) e Estudos de Caso com 29 estudos (24,4%), seguidos por Estudos de Métodos Mistos com 23 estudos (19,3%) e Surveys com 22 estudos (18,5%). Em conjunto, essas quatro categorias representam aproximadamente 95% dos estudos analisados, evidenciando sua ampla adoção na pesquisa empírica publicada no SBES.

Por outro lado, métodos como Pesquisa-Ação, Relatórios de Experiência e Grupos Focais apresentaram baixa representatividade, com apenas dois estudos cada (1,7%), enquanto nenhum estudo foi classificado como Etnografia. Esses resultados evidenciam que a maior parte dos estudos analisados está concentrada em quatro categorias metodológicas: Experimentos Controlados, Estudos de Caso, Surveys e Estudos de Métodos Mistos, enquanto os demais métodos aparecem apenas de forma pontual ou não foram identificados na amostra.

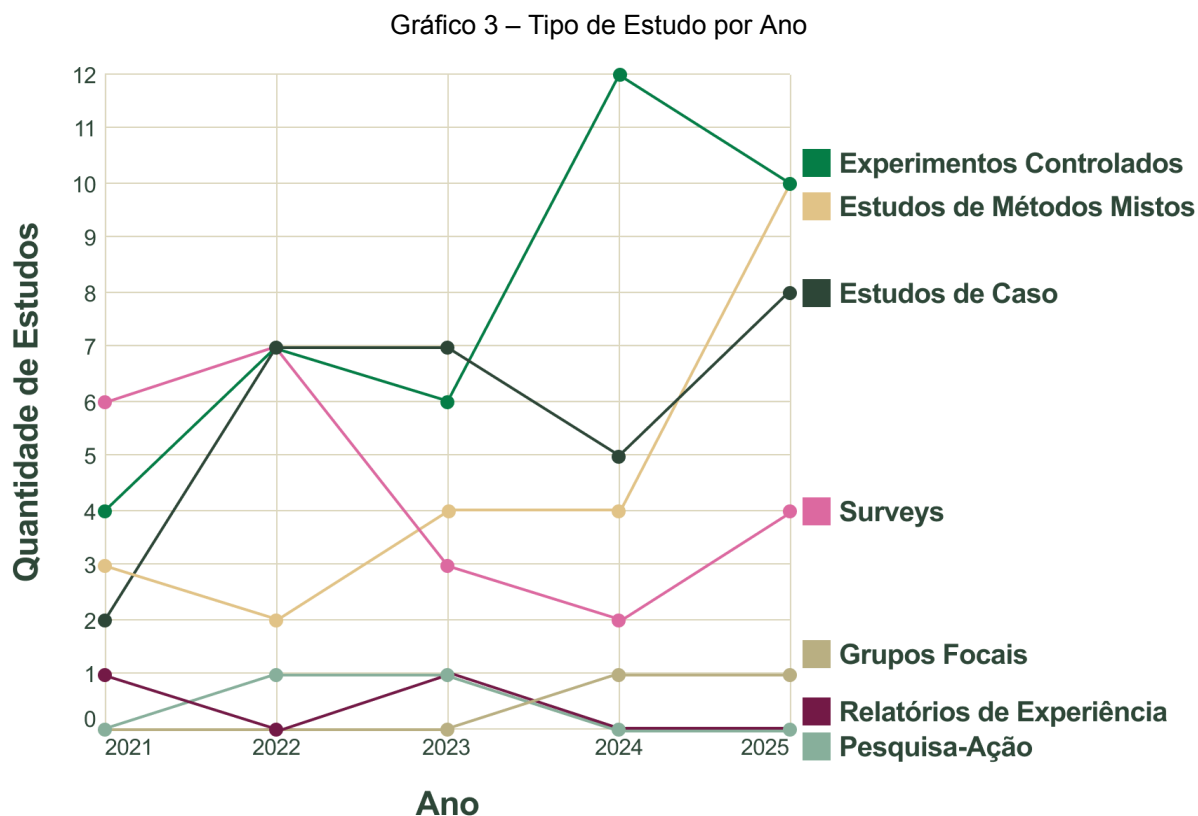
Gráfico 2 – Total de Artigos versus Método



Fonte: Elaborado pelo autor (2026).

O Gráfico 3 apresenta a distribuição dos estudos selecionados por tipo de pesquisa ao longo do período analisado. Observa-se que Experimentos Controlados, Estudos de Caso, Surveys e Estudos de Métodos Mistos estiveram presentes em todos os anos investigados, constituindo os métodos mais frequentemente empregados pelos pesquisadores. Os Experimentos Controlados foram o método predominante ao longo do período, seguidos pelos Estudos de Caso, Estudos de Métodos Mistos e Surveys.

Em contraste, Pesquisa-Ação, Relatórios de Experiência e Grupos Focais apresentaram baixa ocorrência, com apenas dois estudos cada ao longo dos cinco anos analisados, enquanto nenhum estudo foi classificado como Etnografia. Pelo baixo número de ocorrências desses estudos, não foi possível fazer uma análise estatística relevante sobre a qualidade dos relatos por tipo de estudo. Por fim, também se observa um crescimento expressivo dos Estudos de Métodos Mistos em 2025, ano em que foram identificados 10 estudos desse tipo, representando quase metade de todos os trabalhos classificados nessa categoria.



Fonte: Elaborado pelo autor (2026)

A análise dos estudos selecionados no Gráfico 4 mostrou que a grande maioria dos trabalhos apresentou uma seção dedicada às Ameaças à Validade. Dos 119 estudos primários analisados, 116 (97,5%) apresentaram algum relato de Ameaças à Validade, enquanto apenas 3 estudos (2,5%) não relataram qualquer ameaça. Esse resultado indica que a discussão sobre Ameaças à Validade está amplamente presente nos estudos empíricos publicados na trilha de Pesquisa do SBES, embora ainda existam casos pontuais em que tais informações não são apresentadas.

Gráfico 4 – Situação da Apresentação de Ameaças à Validade nos estudos



Fonte: Elaborado pelo autor (2026).

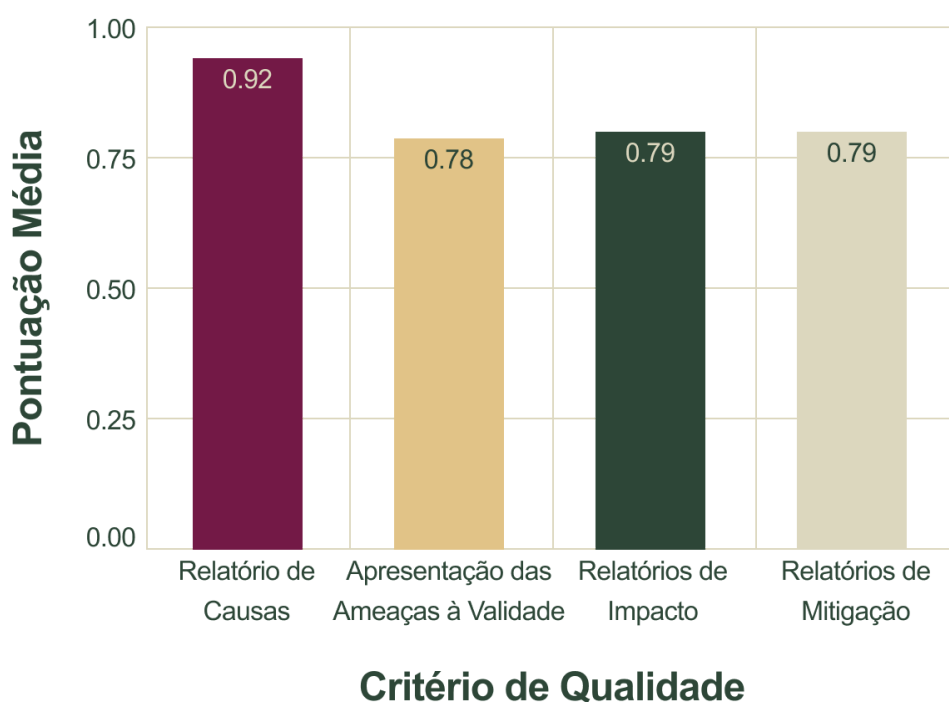
Embora a presença de Ameaças à Validade tenha sido observada em 97,5% dos estudos analisados, esse resultado não implica necessariamente que os relatos sejam completos ou de alta qualidade. Pelo contrário, ele indica que a comunidade reconhece a importância de discutir limitações metodológicas, mas ainda permanece a necessidade de investigar como essas ameaças são descritas, categorizadas e mitigadas.

4.2 COMPLETUDE DOS CRITÉRIOS DE QUALIDADE

O Gráfico 5 apresenta a pontuação média obtida pelos estudos analisados em cada um dos quatro critérios utilizados para avaliar a qualidade dos relatos de Ameaças à Validade. Observa-se que o critério Relatório de causas alcançou a maior média ($\approx 0,92$), indicando que a maioria dos estudos descreve adequadamente as causas associadas às ameaças identificadas. Esse resultado sugere que os autores tendem a justificar de forma consistente a origem dos fatores que podem comprometer a validade de seus estudos.

Os demais critérios apresentaram médias relativamente próximas entre si, variando entre aproximadamente 0,78 e 0,80. A apresentação de Ameaças à Validade registrou a menor média ($\approx 0,78$), indicando que, embora os estudos frequentemente reportem Ameaças à Validade, nem sempre utilizam corretamente classificações formais ou amplamente reconhecidas na Engenharia de Software. Os critérios Relatório de impactos e Relatório de mitigação apresentaram médias semelhantes ($\approx 0,79$), sugerindo que os estudos geralmente discutem os possíveis efeitos das ameaças sobre os resultados e as ações adotadas para reduzi-las, ainda que nem sempre de forma completa.

Gráfico 5 – Pontuação de Qualidade das Ameaças à Validade



Fonte: Elaborado pelo autor (2026).

De maneira geral, os resultados indicam que os relatos de Ameaças à Validade nos estudos analisados apresentam um nível satisfatório de qualidade em todas as dimensões avaliadas. Entretanto, a diferença observada entre o relato das causas e os demais critérios sugere oportunidades de melhoria principalmente na apresentação correta das ameaças, bem como na descrição mais detalhada de seus impactos e estratégias de mitigação. Essas evidências reforçam a importância de diretrizes de relato que promovam maior padronização e completude na documentação das Ameaças à Validade em estudos empíricos de Engenharia de Software.

A Tabela 2 apresenta a distribuição dos estudos em relação aos quatro critérios de qualidade adotados para avaliar o relato de Ameaças à Validade. De forma geral, observa-se que a maioria dos estudos alcançou a classificação Sim em todos os critérios, indicando uma tendência positiva quanto à documentação das Ameaças à Validade.

Tabela 2 – Distribuição da completude dos critérios de qualidade

Distribuição	Sim	Parcialmente	Não
Relatório de causas	103/119 (86,6%)	13/119 (10,9%)	3/119 (2,5%)
Apresentação de Ameaças à Validade	71/119 (59,7%)	45/119 (37,8%)	3/119 (2,5%)
Relatórios de impactos	74/119 (62,2%)	39/119 (32,8%)	6/119 (5,0%)
Relatórios de mitigação	75/119 (63,0%)	39/119 (32,8%)	5/119 (4,2%)

Fonte: Elaborado pelo autor (2026).

O Relatório de causas apresentou o melhor desempenho entre os critérios avaliados, com 103 dos 119 estudos (86,6%) fornecendo descrições completas das causas das ameaças identificadas. Apenas 13 estudos (10,9%) apresentaram relatos parciais e 3 estudos (2,5%) não relataram informações sobre as causas das ameaças. Esse resultado sugere que os pesquisadores tendem a dedicar maior atenção à explicação dos fatores que podem comprometer a validade de seus estudos. Em contraste, a apresentação de Ameaças à Validade apresentou uma distribuição menos concentrada na classificação completa. Apenas 71 estudos (59,7%) categorizaram corretamente suas ameaças de acordo com algum esquema formal de classificação, enquanto 45 estudos (37,8%) apresentaram categorização parcial e 3 estudos (2,5%) não forneceram categorização adequada. Esse resultado indica que, embora a maioria dos estudos adote alguma forma de classificação das ameaças, uma parcela ainda apresenta inconsistências ou limitações nesse processo.

Os relatórios de impactos apresentaram desempenho ligeiramente superior ao da apresentação, com 74 estudos (62,2%) fornecendo descrições completas dos impactos das ameaças identificadas. Outros 39 estudos (32,8%) relataram os impactos apenas parcialmente, enquanto 6 estudos (5,0%) não forneceram informações sobre esse aspecto. Esses resultados sugerem que a descrição das consequências potenciais das ameaças é relativamente comum, embora nem sempre seja realizada de maneira abrangente.

Quanto aos relatórios de mitigação, 75 estudos (63,0%) apresentaram descrições completas das estratégias adotadas para reduzir ou controlar as ameaças identificadas, enquanto 39 estudos (32,8%) forneceram apenas relatos parciais e 5 estudos (4,2%) não relataram qualquer ação de mitigação. Embora a proporção de relatos completos seja ligeiramente superior à observada para impactos, os resultados indicam que ainda existe espaço para melhorar o detalhamento e a abrangência das estratégias de mitigação apresentadas.

Em conjunto, os resultados indicam que o aspecto mais consolidado do relato de Ameaças à Validade é a descrição de suas causas, enquanto a apresentação formal, a descrição de seus impactos e o detalhamento das ações de mitigação apresentam níveis semelhantes de completude, com aproximadamente um terço dos estudos fornecendo apenas relatos parciais. Esses achados evidenciam oportunidades de melhoria na qualidade e na consistência dos relatos das Ameaças à Validade em estudos empíricos da Engenharia de Software.

A Tabela 3 apresenta as correlações de Spearman estatisticamente significativas identificadas entre os critérios de qualidade avaliados e a pontuação total de qualidade do relato de Ameaças à Validade. De modo geral, os resultados indicam associações positivas entre os critérios analisados, sugerindo que estudos com melhor desempenho em uma dimensão tendem também a apresentar maior qualidade global de relato.

Tabela 3 – Interdependência entre os critérios

Correlações de Spearman significativas ($p < 0,05$ e $n=119$)	Coeficiente ρ	Magnitude da Correlação
Causas × Impactos	0,428	Moderada
Apresentação × Pontuação total	0,558	Moderada-forte
Impactos × Pontuação total	0,644	Forte
Mitigações × Pontuação total	0,532	Moderada-forte
Causas × Pontuação total	0,443	Moderada

Fonte: Elaborado pelo autor (2026).

A única correlação significativa observada entre critérios individuais foi entre relato de causas e relato de impactos ($\rho = 0,428$), caracterizando uma associação de magnitude moderada. Esse resultado sugere que estudos que descrevem de forma

mais completa as causas das Ameaças à Validade também tendem a apresentar descrições mais detalhadas de seus potenciais impactos. Em relação a pontuação total de qualidade, todos os critérios, com exceção das correlações não significativas relatadas separadamente, apresentaram associações positivas. A correlação mais elevada foi observada entre relato de impactos e pontuação total ($\rho = 0,644$). Entretanto, como a pontuação total incorpora o próprio critério de impactos, essa associação deve ser interpretada com cautela, pois pode refletir, em parte, um efeito de correlação parte-todo inerente à composição da medida agregada. As correlações entre apresentação de Ameaças à Validade e pontuação total ($\rho = 0,558$) e entre relato de mitigação e pontuação total ($\rho = 0,532$) apresentaram magnitudes moderadas a fortes, evidenciando que esses critérios também contribuem para a qualidade dos relatos. De forma semelhante, o relato de causas apresentou correlação moderada com a pontuação total ($\rho = 0,443$), reforçando sua relevância para a completude da documentação das Ameaças à Validade.

Os resultados sugerem que a qualidade do relato das Ameaças à Validade é influenciada por múltiplas dimensões complementares. Entretanto, a descrição dos impactos parece desempenhar um papel particularmente importante na composição da pontuação global, seguida pela apresentação de Ameaças à Validade e pelo detalhamento das estratégias de mitigação.

A Tabela 4 apresenta as correlações não significativas entre os pares de categorias analisadas no estudo. Em todos os casos, os coeficientes de correlação de Spearman são próximos de zero, indicando ausência de associação relevante entre as variáveis avaliadas.

Tabela 4 – Associações sem significância estatística

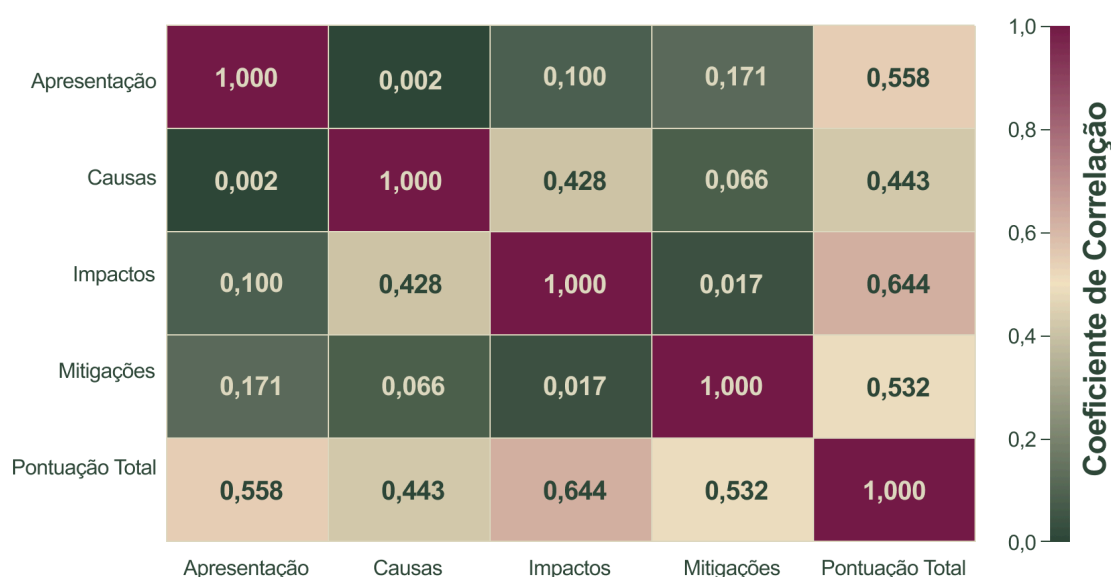
Correlações não-significativas	Coeficiente ρ
Apresentação × Causas	0,002
Apresentação × Impactos	0,100
Apresentação × Mitigação	0,171
Causas × Mitigações	0,066
Impactos × Mitigações	0,017

Fonte: Elaborado pelo autor (2026).

Especificamente, a relação entre Apresentação e Causas apresentou $\rho = 0,02$, sugerindo uma associação praticamente inexistente entre a forma como os elementos são categorizados e as causas associadas. De modo semelhante, a correlação entre Causas e Mitigações ($\rho = 0,066$) também indica uma relação muito fraca, não evidenciando padrão consistente que conecte diretamente as causas identificadas às estratégias de mitigação propostas. As correlações entre a Apresentação e Impactos ($\rho = 0,100$) e entre Apresentação e Mitigação ($\rho = 0,171$) também foram fracas e não estatisticamente significativas. Os resultados indicam que a simples apresentação formal das ameaças à validade não implica necessariamente um relato mais aprofundado de seus impactos ou de suas estratégias de mitigação. Por fim, a relação entre Impactos e Mitigações ($\rho = 0,017$) reforça esse cenário de independência entre as variáveis, sugerindo que os impactos relatados não apresentam associação com as respectivas estratégias de mitigação.

O mapa de calor do Gráfico 6 revela com clareza quais fatores exercem maior influência sobre a Pontuação Total. Esse diagnóstico visual permite direcionar a atenção para os impactos e categorizações, onde residem os maiores motores de mudança para o indicador principal.

Gráfico 6 – Mapa de Calor das Correlações



Fonte: Elaborado pelo autor (2026).

De forma geral, os resultados indicam que, neste conjunto de dados, as categorias analisadas parecem se comportar de maneira relativamente independente entre si, sem evidências de relações sistemáticas fortes ou moderadas entre os pares investigados.

4.3 DISCUSSÃO

Esta seção discute os resultados obtidos à luz das questões de pesquisa que orientam este estudo. Para cada questão, os achados empíricos são interpretados em relação ao contexto da trilha de Pesquisa do SBES e ao corpo de conhecimento existente sobre práticas de reporte de Ameaças à Validade em Engenharia de Software Empírica.

QP1: Qual a proporção de estudos empíricos do SBES que explicitamente documenta Ameaças à Validade?

Os resultados mostram que a discussão sobre Ameaças à Validade está amplamente consolidada nos estudos empíricos publicados na trilha de Pesquisa do SBES. Dos 119 estudos analisados, 116 (97,5%) relatam uma seção de Ameaças à Validade, evidenciando que a grande maioria dos pesquisadores reconhece a importância de explicitar as ameaças metodológicas e potenciais fontes de vies. Esse percentual sugere um nível de maturidade da comunidade em relação às práticas de relato científico, especialmente quando comparado a cenários em que a discussão de validade é tratada apenas de forma superficial ou é completamente omitida. Enquanto Lago et al. (2024) observaram que a ausência de relatos explícitos de Ameaças à Validade ainda é um problema recorrente em estudos empíricos, os resultados encontrados no SBES indicam que essa prática está amplamente consolidada. Contudo, a elevada frequência de relatos não implica necessariamente alta qualidade ou completude, reforçando a necessidade de avaliar não apenas a presença, mas também a profundidade e a consistência das informações relatadas.

A simples presença de uma seção dedicada às Ameaças à Validade não garante que as ameaças sejam adequadamente categorizadas, justificadas ou acompanhadas de estratégias de mitigação. Dessa forma, embora os resultados indiquem que a prática de relatar Ameaças à Validade esteja amplamente disseminada no SBES, eles também reforçam a necessidade de avaliar a profundidade e a qualidade desses relatos.

Outro aspecto relevante é que apenas três estudos não apresentaram qualquer discussão sobre Ameaças à Validade. Embora esse número seja reduzido, sua existência demonstra que ainda há espaço para fortalecer diretrizes editoriais e recomendações para autores, incentivando a adoção consistente dessa prática em todos os estudos empíricos.

QP2: Qual a completude da documentação de Ameaças à Validade em termos de causas, impactos e estratégias de mitigação?

A análise da completude dos relatos revelou diferenças importantes entre os critérios avaliados. O relato das causas apresentou o melhor desempenho, com 86,6% dos estudos fornecendo descrições completas. Esse resultado sugere que os pesquisadores tendem a identificar e justificar adequadamente os fatores que podem comprometer a validade de seus estudos, demonstrando preocupação em explicar a origem das ameaças relatadas. Por outro lado, os critérios relacionados aos impactos e às estratégias de mitigação apresentaram níveis inferiores de completude. Apenas 62,2% dos estudos descreveram completamente os impactos das ameaças identificadas, enquanto 63,0% relatam de forma completa as respectivas estratégias de mitigação. Em ambos os casos, aproximadamente um terço dos estudos apresentou apenas descrições parciais. Esses resultados sugerem que muitos autores conseguem identificar a existência de uma ameaça, mas nem sempre discutem de forma detalhada suas possíveis consequências ou as ações adotadas para reduzir seus efeitos.

A apresentação de Ameaças à Validade também se destacou como um ponto de atenção. Apenas 59,7% dos estudos classificaram corretamente suas ameaças segundo esquemas formais amplamente utilizados na literatura. A elevada

proporção de classificações parciais (37,8%) sugere que ainda existe dificuldade em aplicar de maneira consistente os modelos tradicionais de categorização de Ameaças à Validade como o de Wohlin et al. (2012) e Cook e Campbell (1979). Esse resultado pode refletir tanto limitações de conhecimento metodológico quanto a ausência de orientações mais específicas para o relato dessas informações.

Em conjunto, os achados indicam que os estudos analisados apresentam níveis satisfatórios de completude, mas ainda existem oportunidades de melhoria. Particularmente, a descrição dos impactos das ameaças, das estratégias de mitigação e a correta categorização das Ameaças à Validade representam aspectos que poderiam ser beneficiados pela adoção de diretrizes de relato mais estruturadas e padronizadas.

QP3: Como os critérios de qualidade das documentações de Ameaças à Validade se relacionam entre si nos estudos empíricos do SBES?

Os resultados da análise de correlação sugerem que os diferentes critérios de qualidade do relato de Ameaças à Validade não evoluem necessariamente de forma conjunta. Entre os critérios individuais avaliados, apenas a relação entre relato de causas e relato de impactos apresentou correlação estatisticamente significativa ($\rho = 0,428$), indicando uma associação moderada. Esse resultado sugere que estudos que explicam melhor as origens das ameaças também tendem a discutir de maneira mais detalhada suas possíveis consequências.

Por outro lado, as correlações entre apresentação e causas, causas e mitigação, apresentação e mitigação, apresentação e impactos, impactos e mitigação não foram estatisticamente significativas. Isso indica que o bom desempenho em um critério específico não implica necessariamente qualidade equivalente nos demais critérios. Por exemplo, um estudo pode apresentar uma categorização adequada das ameaças sem fornecer uma descrição aprofundada de suas causas, ou pode discutir os impactos de uma ameaça sem detalhar as ações utilizadas para mitigá-la.

Quando analisada a relação entre cada critério e a pontuação total de qualidade, todas as associações significativas foram positivas. O relato de impactos

apresentou a correlação mais forte com a qualidade geral ($\rho = 0,644$), seguido pela apresentação de Ameaças à Validade ($\rho = 0,558$) e pelo relato de mitigação ($\rho = 0,532$). Esses resultados sugerem que a qualidade geral do relato de Ameaças à Validade depende da combinação de múltiplas dimensões, mas que a discussão dos impactos das ameaças desempenha um papel particularmente relevante na caracterização de relatos mais completos. Os resultados indicam ainda que os critérios avaliados representam dimensões complementares, e não redundantes, da qualidade do relato de Ameaças à Validade. Essa independência relativa reforça a importância de avaliar cada critério individualmente, uma vez que a presença de um elemento bem documentado não garante que os demais aspectos também tenham sido adequadamente relatados. Consequentemente, iniciativas voltadas à melhoria da qualidade dos relatos devem abordar simultaneamente a apresentação, a explicação das causas, a descrição dos impactos e o detalhamento das estratégias de mitigação.

5 AMEAÇAS À VALIDADE

Assim como mostrado na introdução desta pesquisa, é crítico para a transparência e replicabilidade de todo estudo empírico identificar e documentar as Ameaças à Validade do seu estudo, para isso, foi utilizada a taxonomia de Wohlin et al. (2012) para categorizar as Ameaças à Validade deste estudo nas subseções a seguir.

5.1 VALIDADE INTERNA

Há uma ameaça decorrente dos motivos não considerados pelas métricas deste estudo que podem diretamente influenciar a qualidade da documentação das Ameaças à Validade nos estudos analisados e, consequentemente, afetar a sua pontuação final e as correlações de tipo de estudo e pontuação geral, como limites de páginas, senioridade dos autores envolvidos e ano de publicação.

5.2 VALIDADE EXTERNA

Existe uma Ameaça à Validade causada pelo uso de um conjunto específico de estudos, aqueles publicados no SBES, que pode não representar adequadamente o estado das práticas de documentação na comunidade brasileira de engenharia de software empírica, uma vez que o fórum selecionado, apesar de ser o principal no tópico, não representa o todo da comunidade científica deste campo de estudo. Esta é uma ameaça inerente ao estudo, uma vez que seria impraticável incluir outras conferências no espaço de tempo e pesquisadores disponíveis para fazer uma análise mais completa e mitigar os efeitos dessa ameaça, de qualquer forma considera-se os resultados relevantes pelo espaço temporal analisado e a relevância do SBES.

5.3 VALIDADE DE CONSTRUTO

Há uma Ameaça à Validade causada pela indistinção de categoria entre experimentos controlados e quasi-experimentos sob a categoria de “Experimentos Controlados” no checklist, que tem o potencial de impactar os resultados das correlações entre pontuação de qualidade geral da documentação para esse tipo de estudo, já que a categoria não representa unicamente um tipo específico de estudo; Entretanto esta foi uma decisão de desenho do estudo na hora de construir o checklist de tipos, onde fazer a distinção entre os dois tipos iria frequentemente sobrepor características de duas categorias e fazer com que diversos estudos fossem classificados como “Estudos de Métodos Mistos” e, por isso, não foram tomadas ações de mitigação quanto a esta ameaça.

Existe uma Ameaça à Validade causada pelo uso de uma categoria de estudo guarda-chuva na lista de tipos de estudos, a categoria de “Estudos de Métodos Mistos”, para os estudos que não se encaixaram em apenas uma categoria, além do fato de ser desconsiderada a auto-classificação que os autores fazem sobre seus estudos. Isto pode impactar os resultados finais da categoria uma vez que é possível acontecerem categorizações errôneas dos estudos nela incluídos.

Existe uma Ameaça à Validade causada pelo uso de um conjunto específico de tipos de estudos com critérios predefinidos, os quais podem não representar todos os tipos de estudos empíricos publicados recentemente no SBES, e por isso não trazer um resultado completamente representativo para todos os tipos de estudos. Apesar disso, nota-se pelos resultados que todos os tipos de estudos

tiveram pelo menos um estudo incluído em cada categorização, mostrando um grau de maturidade na construção do checklist para abarcar corretamente até os tipos de estudos mais raros presentes na amostra analisada.

Por usar um mesmo valor máximo para todos os critérios de qualidade específicos, os resultados finais de correlação podem ser impactados uma vez que não é concreto que todos os critérios específicos possuam o mesmo grau de importância nos relatos de Ameaças à Validade. Em trabalhos futuros, será feita uma análise sobre qual critério é mais importante para o entendimento das Ameaças à Validade pelos leitores, sendo assim possível ajustar os valores máximos de pontuação para cada um dos critérios a fim de ter uma análise mais balanceada e mitigar os efeitos desta ameaça.

Algumas Ameaças à Validade podem ser consideradas inerentes do desenho do estudo de forma incorreta, afetando assim o critério de qualidade “Documentação de ações de mitigação” já que, por isentar ameaças inerentes de uma ação de mitigação, este critério pode obter incorretamente uma nota mais alta que os demais critérios. Por motivos de tempo da pesquisa foi decidido continuar com uma simples descrição, mas é reconhecido que é necessário desenvolver mais especificamente o conceito de “ameaça inerente ao desenho do estudo” para que os resultados deste critério não sejam indevidamente mais altos que os demais.

5.4 VALIDADE DE CONCLUSÃO

Assim como qualquer revisão sistemática, uma ameaça existente se deve a subjetividade da interpretação de texto sobre as seções de Ameaças à Validade dos estudos, isso pode impactar diretamente as pontuações dos critérios de qualidade específicos e consequentemente a conclusão final, uma vez que diferentes pesquisadores podem ter diferentes interpretações sobre o mesmo texto de documentação. A replicação da metodologia utilizada neste estudo para posteriores edições do SBES ou de outras conferências, com um número maior de pesquisadores a fazer a análise, pode mitigar os efeitos desta ameaça.

Esta análise foi conduzida com um número extremamente limitado de pesquisadores para executar a análise da documentação das Ameaças à Validade, o que pode impactar diretamente os resultados obtidos nas pontuações específicas e gerais, além dos resultados de categorização dos tipos de estudos. Seria necessária

a replicação deste estudo em outros fóruns, contextos, e números de pesquisadores disponíveis para gerar evidências empíricas mais amplas sobre o estado atual das práticas de documentação na engenharia de software brasileira.

O uso de apenas 5 anos de edições do SBES com apenas uma trilha de oito tipos de estudos pode ser considerado um escopo demasiadamente estreito para a análise estatística e por isso pode trazer uma amostra muito pequena, com pouco poder de generalização dos resultados. De qualquer forma, considera-se que é uma base empírica de guia bem construída para indicar o estado das práticas de documentação no contexto analisado, bem como uma base de replicação sólida para estudos futuros com amostras e poder de generalização maiores.

5.5 TRABALHOS FUTUROS

As direções de continuidade desta seção derivam diretamente das ameaças identificadas e discutidas anteriormente, de modo que cada proposta responde a uma limitação identificada neste estudo.

A primeira direção, como mencionado na subseção 5.3, propomos balancear os pesos dos critérios de qualidade específicos por meio de pesquisas futuras para analisar o nível de entendimento e importância de cada critério com pesquisadores experientes da área, a fim de melhor balancear futuras análises utilizando a base deste estudo.

Segundo, propomos operacionalizar mais especificamente o conceito de “Ameaça à Validade inerente ao desenho do estudo” por meio de uma definição precisa e de regras de decisão validadas junto a pesquisadores experientes, aplicando-as a replicações futuras deste estudo em outros fóruns e edições a fim de corrigir a possível tendência de superestimação do critério de documentação de ações de mitigação.

Indo além, propomos replicar o processo de extração e pontuação com mais pesquisadores independentes, reportando medidas de concordância entre eles e comparando a categorização obtida pelo checklist com a auto-classificação declarada pelos autores, para que seja possível mitigar os efeitos da ameaça discutida na seção 5.4

Finalmente, propomos expandir a análise a edições futuras e anteriores do SBES, além de outros fóruns nacionais e internacionais como SBQS e ESEM,

permitindo descobrir tendências temporais e testar em que medida os resultados observados neste trabalho se generalizam em outros contextos.

6 CONSIDERAÇÕES FINAIS

Os resultados deste estudo revelam que a comunidade do SBES incorporou amplamente a prática de relatar Ameaças à Validade em estudos empíricos. Entretanto, a qualidade desses relatos ainda apresenta variações importantes, especialmente em relação à apresentação formal das ameaças e ao detalhamento de seus impactos e estratégias de mitigação. Além disso, a limitada correlação entre vários dos critérios analisados sugere que a qualidade do relato é multidimensional e requer atenção específica a cada um de seus componentes. Esses achados reforçam a necessidade de diretrizes e recomendações que promovam relatos mais completos, consistentes e padronizados, contribuindo para aumentar a transparência, a reprodutibilidade e a credibilidade da pesquisa em Engenharia de Software Empírica.

REFERÊNCIAS

- AMPATZOGLOU, Apostolos *et al.* Guidelines for Managing Threats to Validity of Secondary Studies in Software Engineering. *In*: FELDERER, Michael; TRAVASSOS, Guilherme Horta (Orgs.). Contemporary Empirical Methods in Software Engineering. Cham: Springer International Publishing, 2020. p. 415–441.
- AZEVEDO, Ivanildo *et al.* Reimagining Studies' Replication: A Validity-Driven Analysis of Threats in Empirical Software Engineering. *In*: Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <<https://sol.sbc.org.br/index.php/sbes/article/view/37052>>
- Cook, T. D., & Campbell, D. T.. Quasi- experimentation: Design and analysis issues for field settings. Chicago: Rand-McNally. 1979.
- CRUZES, Daniela; BEN OTHMANE, Lotfi. Threats to Validity in Empirical Software Security Research. *In*: *Threats to Validity in Empirical Software Security Research*. p. 275–300, nov. 2017.
- EASTERBROOK, Steve *et al.* Selecting Empirical Methods for Software Engineering Research. *In*: SHULL, Forrest; SINGER, Janice; SJØBERG, Dag I. K. (Orgs.). Guide to Advanced Empirical Software Engineering. London: Springer, 2008. p. 285–311.
- FONSECA, Liliane da Silva; TEIXEIRA, Eudis Oliveira; SOARES, Sergio. ValiDEPlan — Validity-Driven Software Engineering Experiments Planning Tool. *In*: Porto Alegre, RS, Brasil: SBC, 2019. Disponível em: <https://sol.sbc.org.br/index.php/cbsoft_estendido/article/view/7665>
- GREN, Lucas. Standards of validity and the validity of standards in behavioral software engineering research: the perspective of psychological test theory. *In*: ESEM '18. New York, NY, USA: Association for Computing Machinery, 2018. Disponível em: <<https://doi.org/10.1145/3239235.3267437>>
- KITCHENHAM, Barbara A. *et al.* Preliminary Guidelines for Empirical Research in Software Engineering. IEEE Transactions on Software Engineering, v. 28, n. 8, p. 721–734, 2002.
- KNAFLIC, Cole Nussbaumer. Storytelling com dados: um guia de visualização de dados para profissionais de negócios. Rio de Janeiro: Alta Books, 2018.
- SCHLOEGEL, Moritz *et al.* SoK: Prudent Evaluation Practices for Fuzzing, 2024.
- LAGO, Patricia *et al.* Threats to Validity in Software Engineering – hypocritical paper section or essential analysis? *In*: ESEM '24. New York, NY, USA: Association for Computing Machinery, 2024. Disponível em: <<https://doi.org/10.1145/3674805.3686691>>
- MY BLOOD. Intérprete: Twenty One Pilots. Compositor: Tyler Joseph. *In*: TRENCH. Intérprete: Twenty One Pilots. Hollywood, Califórnia, Estados Unidos da América: Fueled By Ramen, 2018. 1 CD, faixa 4.

RUNESON, Per *et al.* Case Study Research in Software Engineering: Guidelines and Examples. Hoboken, NJ: John Wiley & Sons, 2012.

SIEGMUND, Janet; SIEGMUND, Norbert; APEL, Sven. Views on internal and external validity in empirical software engineering. *In: ICSE '15.* Florence, Italy: IEEE Press, 2015.

TEIXEIRA, Eudis; FONSECA, Liliame; SOARES, Sergio. Threats to validity in controlled experiments in software engineering: what the experts say and why this is relevant. *In: SBES '18.* New York, NY, USA: Association for Computing Machinery, 2018. Disponível em: <<https://doi.org/10.1145/3266237.3266264>>

VERDECCHIA, Roberto *et al.* Threats to validity in software engineering research: A critical reflection. *Inf. Softw. Technol.*, v. 164, n. C, dez. 2023.

WOHLIN, Claes *et al.* Experimentation in Software Engineering. Berlin / Heidelberg: Springer, 2012.

APÊNDICE A – ESTUDOS SELECIONADOS

Edição do SBES	Nome e Link do Artigo
2025	FOSTER Taxonomy-based Open Science-Related Practices in Software Engineering: Review and Observations
2025	Measuring how changes in code readability attributes affect code quality evaluation by Large Language Models
2025	LLMs as Test Generators: A Comparative Benchmarking Study
2025	Investigating the Impact of GitHub Discussions on Maintainers' Workload and Community Dynamics
2025	Socio-Technical Smell Dynamics in Code Samples: A Multivocal Review of Their Emergence, Evolution, and Co-Occurrence
2025	Explorando a Influência da Segurança Psicológica na Qualidade do Software em Equipes Ágeis
2025	Combining TSL and LLM to Automate REST API Testing: A Comparative Study
2025	Aged to Perfection? Analyzing the Impact of Years of Experience on Code Quality
2025	Building the Landing Zone for Site Reliability Engineering: A Multivocal Literature Review
2025	Rumo a um Catálogo de Anti-padrões de Observabilidade: Uma Revisão da Literatura
2025	Evaluating Fine-tuning Approaches for Duplicate Bug Report Detection
2025	Generating Proto-Personas through Prompt Engineering: A Case Study on Efficiency, Effectiveness and Empathy
2025	An Ontology-based Approach for Developing Systems to Identify Who Knows What in Software Organizations
2025	Designing Mutation Operators for Android Device Components: A View Through Bluetooth and Location API's
2025	Threat Modelling and Vulnerability Assessment for IoT Solutions: A Case Study
2025	A Comparative Study of LLMs for Gherkin Generation
2025	Bugs in the Shadows: Static Detection of Faulty Python Refactorings

2025	Identifying and Addressing Test Smells in JavaScript: A Developer-Centric Study
2025	Choosing the Right GitWorkflow: A Comparative Analysis of Trunk-based vs. Branch-based Approaches
2025	A Real-time Data Synchronization Approach for High-availability Micro Applications
2025	Continuously Managing Microservice Granularity: An Evidence-Based Industrial Approach
2025	“Frustrating, Stressful, and Overwhelming”: Insights into Software Practitioners’ Productivity from Stack Exchange Discussions
2025	Science Communication of Brazilian Software Engineering Research on LinkedIn
2025	LSTM-based Forecasting for Non-linear Workloads in On-premises Software Systems
2025	Attentionsmelling: Using Large Language Models to Identify Code Smells
2025	GenIA-E2ETest: A Generative AI-Based Approach for End-to-End Test Automation
2025	An Exploratory Study on the Lifecycle of Code Clones During Code Review
2025	404: Civility Not Found? Evaluating the Effectiveness of Small Language Models in Detecting Incivility in GitHub Conversations
2025	Evaluating the Potential of Large Language Models in Security-Related Software Requirements Classification
2025	PRemo: A Dataset of Emotions Found on Pull Request Discussions
2025	Automatic Identification of Machine Learning-Specific Code Smells
2025	Investigating How Psychological Safety Can Be Fostered in Agile Teams
2025	Understanding How Brazilian Software Organizations Collect and Utilize User Feedback: A Survey
2025	Diversity Matters: Perceived Inclusion and Discrimination by Brazilian Tech Professionals
2025	An Empirical Study on the Effectiveness of Iterative LLM-Based Improvements for Static Analysis Issues
2025	Unveiling the Relationship Between Continuous Integration and Code Review: A Study with 10 Closed-source Projects
2025	Survey-Based Insights into the Replication Crisis and the 3R in Software Engineering

2025	Understanding Psychological Safety in Agile Software Development Teams
2024	Robotic-supported Data Loss Detection in Android Applications
2024	Code smell severity classification at class and method level with a single manually labeled imbalanced dataset
2024	Towards Effective Collaboration between Software Engineers and Data Scientists developing Machine Learning-Enabled Systems
2024	Investigating Accountability in Business-intensive Systems-of-Systems
2024	Unveiling the Landscape of System Thinking Modeling Tools Use in Software Engineering
2024	Revisiting Aristotle vs. Ringelmann: The influence of biases on measuring productivity in Open Source software development
2024	A Comparison Between Hierarchical and Non-Hierarchical Software Clustering
2024	How do Agile Organizations Manage Risks: An Analysis of the State of Practice in Brazil
2024	Addressing the Synchronization Challenge in Cypress End-to-End Tests
2024	Investigating Benefits and Limitations of Migrating to a Micro-Frontends Architecture
2024	Challenges and Solutions of Free and Open Source Software Documentation: A Systematic Mapping Study
2024	Views on Agile Leadership for Software Teams: A Case Study with Leaders and Non-Leaders
2024	Assessing Python Style Guides: An Eye-Tracking Study with Novice Developers
2024	Applying Design Science Research to Develop a Process Reference Model for IT Workforce Outsourcing: Preliminary Findings
2024	Guiding the Way: Facilitating Requirements Elicitation with Selection Universe Approach
2024	Categorizing IoT Software Systems Security Vulnerabilities Through Literature Studies
2024	Explorando a detecção de conflitos semânticos nas integrações de código em múltiplos métodos
2024	The Skill Gap in Software Industry: A Mapping Study
2024	Iterative Deepening URL-Based Search: Enhancing GUI Testing for Web Applications

2024	How does Technical Debt Evolve within Pull Requests? An Empirical Study with Apache Projects
2024	Industrial Practices of Requirements Engineering for ML-Enabled Systems in Brazil
2024	Rumo a uma Taxonomia de Observabilidade para Aplicações Baseadas em Microsserviços
2024	Refinando a Precisão da Detecção de Conflitos: Uma Análise do CSDiff com Abordagem Focalizada
2024	Avaliação de Métodos para Elicitação e Especificação de Requisitos de Usabilidade com Histórias de Usuário: Um Experimento Controlado
2024	Exploring the Role of Job Satisfaction and Work-Life Balance: A Study on the Mental Health Among Software Engineers
2024	Comparative Analysis of Large Language Model Tools for Automated Test Data Generation from BDD
2024	Promise+: expandindo a base de dados de requisitos de software Promise exp
2024	Eyes on Code Smells: Analyzing Developers' Responses During Code Snippet Analysis
2024	Detecting Code Smells in JavaScript: An Annotated Dataset for Software Quality Analysis
2024	A self-adaptive IoT architecture to support intelligent environments
2024	On the Employment of Machine Learning for Recommending Refactorings: A Systematic Literature Review
2023	A feasibility study of usability and UX evaluation technologies in multi-touch context: A quantitative and qualitative analysis
2023	A Thematic Synthesis on Empathy in Software Engineering based on the Practitioners' Perspective
2023	An Agile Management Model for Distributed Software Development Teams
2023	An Assessment of Machine Learning Algorithms and Models for Prediction of Change-Prone Java Methods
2023	An empirical study of the relationship between refactorings and merge conflicts in Javascript code
2023	Analyzing a Semantics-Aware Bug Seeding Tool's Efficacy: A qualitative study with the SemSeed tool
2023	Analyzing the Impact of CI Sub-practices on Continuous Code Quality in Open-Source Projects: An Empirical Study
2023	Applying Spectrum-Based Fault Localization to Android Applications

2023	Architectural Technical Debt - A Systematic Mapping Study
2023	Assessing the Readability of ChatGPT Code Snippet Recommendations: A Comparative Study
2023	Bit-wise All Edge-pairs Coverage
2023	Cytession: Automated GUI Testing for Web Applications
2023	Do you see what happens around you? Men's Perceptions of Gender Inequality in Software Engineering
2023	DoME: An Architecture for Domain Model Evolution at Runtime Using NLP
2023	Effects of Modularization on Developers' Cognitive Effort in Code Comprehension Tasks: A Controlled Experiment
2023	Empirical evidence on technical challenges when adopting continuous practices
2023	Evolution of Teamwork Quality Instruments in Agile Software Development: A Systematic Literature Review
2023	Exploring Psychological Safety in Software Engineering: Insights from Stack Exchange
2023	Flaky Tests in UI: Understanding Causes and Applying Correction Strategies
2023	Help! I need somebody. A Mapping Study about Expert Identification in Software Development
2023	How The Retry Pattern Impacts Application Performance: A Controlled Experiment
2023	Hybrid Software Development with Scrum: Perceptions of Brazilian Software Practitioners
2023	Investigating Transparency in Software Ecosystems
2023	On the Experiences of Practitioners with Requirements Elicitation Techniques
2023	On the Support for Designing a Conversational Software API: An Action Research Study: An Action Research Study
2023	On the Understanding of the Role of Continuous Experimentation in Technology-Based Startup
2023	Similar Bug Reports Recommendation System using BERT
2023	Software Merge: A Two-Decade Systematic Mapping Study

2023	Understanding an organizational change and development intervention applied in a Global Software Industry: A case study: A Case Study
2022	A Configurable Test Case Prioritization Technique for Early Fault Detection and Low Test Case Spreading
2022	A Technique to Test Refactoring Detection Tools
2022	Assessing the Impact of Code Samples Evolution on Developers' Questions
2022	Benefits and Difficulties of Gender Diversity on Software Development Teams: A Qualitative Study
2022	Semantic conflict detection with overriding assignment analysis
2022	Developers in focus! Developer Experience Analysis in a Collaborative Modeling Tool
2022	Understanding the relationships between the perceptions of burnout and instability in Software Engineering
2022	Flying over Brazilian Organizations with Zeppelin: A Preliminary Panoramic Picture of Continuous Software Engineering
2022	ICT Practitioners' Perception of Working from Home During the Covid-19 Pandemic: Exploring Gender Differences
2022	Machine Learning for Change-Prone Class Prediction: A History-Based Approach
2022	Perceptions of Technical Debt and Its Management Activities - A Survey of Software Practitioners
2022	PriorTD: A Method for Prioritization Technical Debt
2022	Put Your Hands In The Air! Reducing Manual Effort in Mutation Testing
2022	Smell Patterns as Indicators of Design Degradation: Do Developers Agree?
2022	The Private Life of Merge Conflicts
2022	The Role of Non-Technical Skills in the Software Development Market
2022	The who, what and how of the current research at the Brazilian Symposium on Software Engineering
2022	Towards Merge Conflict Resolution by Combining Existing Lines of Code
2022	Towards Mobile UX Evaluation and Improvement in Startups: A Case Study on the Adoption of User Experience Questionnaire and Hierarchical Task Analysis

2022	Towards a Fault Taxonomy for Microservices-Based Applications
2022	Towards a common understanding of sustainable software development
2022	UX Requirements Matters: Guidelines to Support Software Teams on the Writing of Acceptance Criteria
2022	A characterization study of testing contributors and their contributions in open source projects
2022	A Decade of Internationalization of the Brazilian Symposium on Software Engineering: The Good, the Bad, and the Ugly
2022	Use Cases for Software Development Analytics: A Case Study
2022	Validating an Interactive Ranking Operator for NSGA-II to Support the Optimization of Software Engineering Problems
2021	Human Factors and their Influence on Software Development Teams - A Tertiary Study
2021	A Mining Software Repository Extended Cookbook: Lessons learned from a literature review
2021	A Survey on the Use of UML in the Brazilian Industry
2021	Towards a Technique to Detect Weaknesses in C Programs
2021	Deep Reinforcement Learning Based Android Application GUI Testing
2021	Knowledge-based Risk Management: A Systematic Literature Review
2021	A Quasi-Experiment to Investigating the Impact of the Strategy Design Pattern on Maintainability
2021	How Do Brazilian Software Development Teams Deal with Working From Home After a Year of the COVID-19 Pandemic?
2021	Breaking one barrier at a time: how women developers cope in a men-dominated industry
2021	Experiences and Practices in GUI Functional Testing: A Software Practitioners' View
2021	Factors That Affect Merge Conflicts: A Software Developers' Perspective
2021	How do Code Smell Co-occurrences Removal Impact Internal Quality Attributes? A Developers' Perspective
2021	Trust yourself! Or maybe not: factors related to overconfidence and uncertainty assessments of software effort estimates

2021	On the Interplay of Smells Large Class, Complex Class and Duplicate Code
2021	Ants Doing Legwork: Investigating Motivators for Software Development Career Abandonment
2021	Evolving Delta-Oriented Product Lines: A Case Study on Feature Interaction, Safe and Partially Safe Evolution
2021	Textual merge based on language-specific syntactic separators
2021	Estimativa do Tempo de Resolução de Issues no GitHub Usando Atributos Textuais e Temporais
2021	On the relation between technical debt indicators and quality criteria in Stack Overflow discussions
2021	Learning Quick Fixes from Code Repositories

APÊNDICE B – CRITÉRIOS DE CLASSIFICAÇÃO DOS TIPOS DE ESTUDO

Tipo de Estudo	Critérios de característica
Experimentos Controlados (incluindo quasi-experimentos)	- Uma hipótese clara e testável; - Uma teoria deduzida a partir da hipótese; - Uma ou mais variáveis independentes sendo manipuladas; - Especificação de como as variáveis serão medidas; - Variáveis, que quando não independentes, não afetem o experimento; - Combinação de mudanças de variáveis (tratamentos); - (Para experimentos controlados) Sujeitos alocados randomicamente aos tratamentos; - (Para quasi-experimentos) Sujeitos não alocados randomicamente aos tratamentos.
Estudos de Caso	- Uma pergunta de pesquisa preocupada com o “como” ou “por que” um certo fenômeno acontece no contexto da vida real; - Uma proposição de estudo que indica precisamente o que tem a intenção de mostrar; - O pesquisador é um observador neutro com mínimo a nenhum impacto no ambiente de pesquisa; - Uso de amostragem intencional (no caso de serem múltiplos casos); - Uma unidade bem definida de análise para a coleta de dados.
Pesquisas Survey	- Uma coorte exploratória para geração da hipótese e identificação de tendências; - Questões a serem respondidas por uma amostra humana; - Uma pergunta de pesquisa que indaga sobre a natureza da amostra populacional em foco; - Uma unidade bem definida de análise para a coleta de dados.
Etnografias	- Consideração explícita das pré-concepções do pesquisador envolvido que podem afetar a pesquisa; - Uma pergunta de pesquisa que foca nas práticas culturais de uma comunidade em particular; - Acesso aos participantes da comunidade em particular sendo estudada; - Meios para identificar os potenciais membros representantes da comunidade em particular.
Pesquisas-Ação	- Um “Dono do Problema” que colabora para identificá-lo; - O pesquisador é um agente ativo de mudança para resolver o problema por si só

	ou inserido numa equipe; - Uma solução final que é específica ao problema e não é intencionada a generalizar a solução para uma classe ampla de problemas similares; - Um problema real e importante que precisa de solução; - Uma seção de “conhecimentos obtidos” com a solução para o problema.
Grupos Focais	- Uma pergunta de pesquisa focada em explorar motivações, percepções compartilhadas ou construção de consenso; - Um script semi-estruturado ou não-estruturado de moderação, ou um conjunto de questões; - O autor atuou como moderador; - Amostragem intencional dos participantes “possuidores de informação relevante”; - Resultados de análise qualitativa verbal ou textual.
Relatos de Experiência	- Autores atuando como praticadores, educadores ou participantes ativos que diretamente desenharam e implementaram a metodologia sendo usada no estudo; - Descrição da aplicação da prática, técnica, aplicação específica, ferramenta, método pedagógico, ou projeto em um contexto autêntico da vida real; - Avalia a experiência retrospectivamente de forma sem a necessidade de se basear em controles estritamente experimentais, atribuição aleatória de atividades, ou testes formais utilizando variáveis isoladas; - Uso de dados empíricos para embasar as observações (logs, retorno dos participantes, avaliações de performance, etc).
Estudos de Métodos Mistos	- Coleta de dados e técnicas de análise associadas com dados qualitativos e quantitativos; - Duas ou mais características chave de mais de um dos tipos de estudos citados acima.

**APÊNDICE C – PROTOCOLO DE ANÁLISE DAS AMEAÇAS À VALIDADE NOS
ESTUDOS SELECIONADOS**

**Protocol for a Systematic Literature Review on Threats to Validity Reports in
Brazilian Empirical Software Engineering Studies**

Marconi G. R. Filho
Ivanildo Azevedo
Sérgio Soares

*Centro de Informática
Universidade Federal de Pernambuco
Recife - PE, Brazil*

1 TRACK OF CHANGES

Author	Date	Description of Change
Marconi Filho	11/11/2025	Initial proposal of protocol
Marconi Filho	11/12/2025	Adds sections for 'Background', 'Research Questions', 'Inclusion Criteria' and 'Exclusion Criteria'
Marconi Filho	15/12/2025	Adds 'Search Criteria', 'Study Classification Checklist' and 'Selected Studies' sections.
Marconi Filho	09/02/2026	Adds 'Data Analysis' section.
Marconi Filho e Ivanildo Azevedo	19/02/2026	Review and adjustment of consistency details in the protocol
Ivanildo Azevedo	25/02/2026	Adjustment of proposal details
Marconi Filho	02/03/2026	Adds 'Step-by-step Description' section

2 ABSTRACT

Context: This protocol describes the methodology for a Systematic Review (SR) performed to evaluate how threats to validity (TTVs) are documented, detailed, and planned mitigation in empirical software engineering studies published in the last 5 editions of the Brazilian Symposium on Software Engineering (SBES). The choice of SBES derives from its relevance within the Brazilian and Latin American software engineering research communities and its strong focus on empirical studies, but the choice of a reduced scope comes from this study's context which is framed as an undergraduate thesis and the small number of researchers available to run the analysis.

Objective: The search aims to quantify the frequency to which TTVs are documented, identify the most commonly reported types of threats based on Cook and Campbell's (1979) classification, assess the report completeness of them

according to predefined quality criteria, and analyze the relation between mitigation actions taken and study types following a type categorization list by specific criteria.

Method: To address quantitative research questions the protocol employs frequency analysis, counting techniques, and statistical tests to examine possible associations between study types and validity-related indicators. Quality indicators will be derived using a three-level ordinal scale assessing the use of formal classification schemes of TTVs, presentation of threats, discussion of causes, impacts, and mitigation strategies.

Expected Results: By using quantitative metrics, this search aims to provide a comprehensive overview of current reporting practices, highlight gaps in the reports of TTVs, as well as supporting future improvements in the methodological rigor of empirical research in software engineering.

3 STEP-BY-STEP DESCRIPTION

The initial phase of the pilot study is executed by one of the researchers and consists of a sequential review of the proceedings from the latest edition of the Brazilian Symposium on Software Engineering (SBES 2025). The procedure for literature screening, classification, and data extraction was conducted and described as it follows:

1. Initial Screening and Data Registration: For each paper in the archival sequence, the researcher first records the study's source link (PDF) in the data sheet. Then, a preliminary screening of the title and abstract is performed to understand the study's general context and methodological framework.

1.1 SBES 2023 and earlier: These editions have no clear separation between tracks, so in the website for the symposiums' edition an inner section regarding the research track's accepted papers is accessed. All the names of the accepted papers are exactly matched with the name on the archival's website, then pasted on the sheet (excluding journal-first articles).

2. Study Type Categorization and Filtering: Following the initial screening, studies are categorised based on their research design. This classification requires a

targeted review of the methodology section (and supplementary sections if needed) to determine the study's category according to this protocol's checklist.

- Studies classified as secondary or tertiary research are excluded from further phases due to being on the exclusion criteria;
- Primary studies are retained and advanced to the rigorous TTV extraction phase.

3. TTVs Section analysis: For each retained primary study, the researcher searches for explicit reports of TTVs. A "TTVs Presentation" score is assigned based on the presence and nomenclature of such section:

- **Score 0:** If the study entirely lacks a section dedicated to TTVs or limitations. Further phases are not conducted in this case and the study gets 0 in all other criterion; The researcher may proceed to search for the keywords 'limitations' or 'threats' on the paper and note the findings on the notes.
- **Score 0,5:** If the study lacks a formally designated TTVs section but contains a generalized "Limitations" section, has not a dedicated section but a fairly long and descriptive subsection of TTVs/Limitations, or if the study does not use the four-category taxonomy proposed by Cook and Campbell (1979); When any of those are true, the analysis will proceed to step 5 considering each limitation as a TTV for all purposes of this protocol.
- **Proceed to Detailed Analysis:** If a dedicated TTVs section is identified with structured threats according to a well-known taxonomy, like the one proposed by Cook and Campbell (1979), regardless of whether the authors explicitly cited the framework, the researcher proceeds to a comprehensive analysis of the section. This involves extracting the identified threats across each category, alongside their reported causes, impacts, and corresponding mitigation actions.

4. Assessment of TTVs Presentation: For studies undergoing detailed analysis, the researcher analyses the conceptual presentation of the reported TTVs by cross-checking the authors' categorization against the description established on the framework by Cook e Campbell (1979):

- **Score 0,5:** Is assigned to the "TTVs Presentation" criterion if the study shows misalignments or incorrectly categorized the reported TTVs.
- **Score 1:** Is assigned to the "TTVs Presentation" criterion if the study accurately reports the specific TTVs within their appropriate theoretical categories.

5. TTV counting: The researcher reads the entirety of the TTVs section counting how many threats were reported and adding the recorded data to the Data Sheet.

6. Report completeness criterion: Beyond presentation, the completeness of the reported threats is evaluated. The read TTVs section content is systematically compared against the scoring checks described in the “Quantitative Analysis Procedure” section of this protocol to assign the report completeness score for each component (causes, impacts and mitigations).

6.1 Actual report: The data gathered is considered only by what the author reported and not what could be reported on its study.

6.2 Mitigations reporting: Recommendations inside the TTVs section that focus on inherent threats of the study will also be considered as mitigation actions, for example a Survey with Brazilian population has an inherent limitation that it may not generalize the findings, so future studies are recommended to replicate this study in other populations such as Europe; Or a study that focus on a non-exhaustive list of cases, and acknowledges that not all cases can be analyzed, is also discarded from mitigation reporting.

6.3 Justification notes: For each perceived deficiency on criterion reporting the researcher should note the reason on the “justification” section of the analysed study on the data sheet.