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

Aluno: Eudis Oliveira Teixeira

Orientador: Sergio Castelo Branco Soares

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Agenda

- Introdução
- Revisão da Literatura
- Instrumento de Avaliação
- Metodologia de Avaliação da Qualidade
- Análise dos Dados e Discussão dos Resultados
- Considerações Finais

Motivação

- Importância de se realizar uma boa avaliação da qualidade dos estudos que se quer incluir na pesquisa
  - Agregar Evidências
  - Construir Conhecimento
  - Não comprometer os Resultados
  - Não há um padrão de avaliação da qualidade acordado para EE

Problemática

- Para Dyba, Kampenes e Sjoberg [21]
  - A **prática corrente** na condução e forma de divulgar os resultados de experimentos controlados em ES **é inadequado**.
  - Um estudo sem um adequado **tratamento estatístico** não é capaz de fornecer **informação suficiente para tirar conclusões** sobre a **aceitação ou rejeição das hipóteses** da pesquisa.
- Para Silva et al. [59], e Kitchenham et al. [40]
  - A **maioria das SLRs não avalia a qualidade dos estudos primários**.
- Para Higgins et al. [27]
  - Há muitos experimentos que não seguem **padrões confiáveis, válidos e repetíveis**.
- Para Jedlitschka et al. [29]
  - A **falta de padrões de *report* e a heterogeneidade de estilos** de divulgação é um grande problema para a integração de resultados de experimentos controlados em um corpo comum de conhecimento.

Questões de Pesquisa

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

Objetivo

- Analisar a **qualidade e evolução** dos experimentos controlados quanto aos:
  - **Mecanismos de suporte utilizados**
  - **Replicabilidade**
  - **Rigor estatístico**

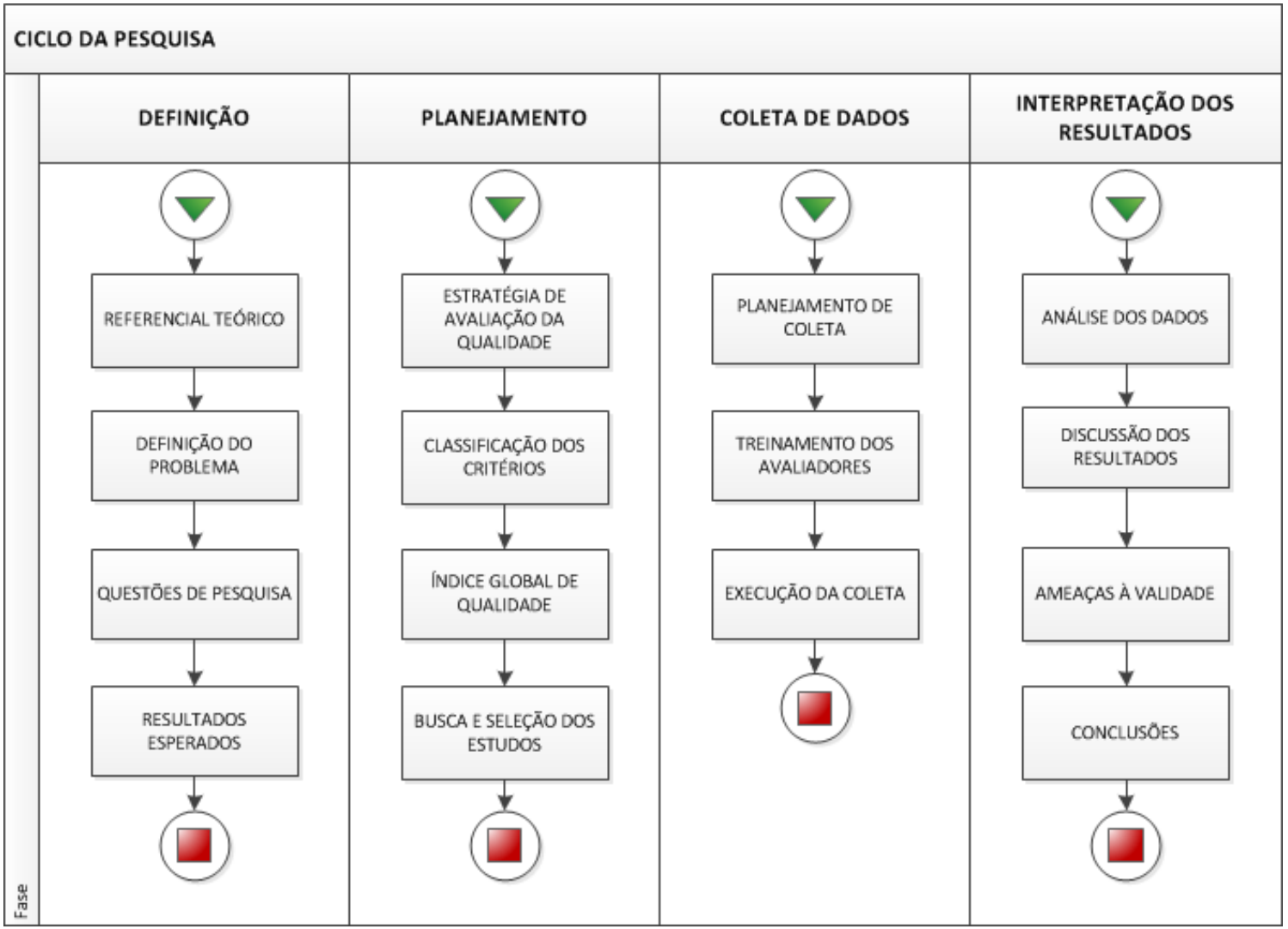
Contextualização

O estudo desenvolveu-se a partir do contexto de um mapeamento sistemático

- O mapeamento examinou 876 artigos.
- Identificar e analisar evidências sobre quais metodologias, processos, *guidelines*, ferramentas, técnicas e boas práticas foram utilizadas no suporte à execução dos estudos empíricos.
- 104 estudos categorizados como experimentos controlados.
- 84 estudos mencionaram ter usado mecanismo de suporte ao planejamento e/ou execução do experimento.

Introdução

Ciclo da Pesquisa



Mecanismos de Suporte Utilizados

- Nos últimos 30 anos, diversos estudos têm apresentado:
  - Ambientes de apoio
  - *Guidelines*
  - Ferramentas
  - Metodologias
  - Técnicas
  - Frameworks
  - Boas Práticas
- São mecanismos de suporte que permitem planejar, executar, analisar e empacotar estudos empíricos em ES.

Avaliação da Qualidade

- Em 2004, Kitchenham et al. [36] apresentou um ***Guideline*** sobre Engenharia de Software Baseado em Evidências (ESBE) com algumas **recomendações** de avaliação da qualidade para RSL's
- Em 2007 Kitchenham et al. [38] publicou novas diretrizes e uma lista com **50 perguntas** destinadas a avaliação da qualidade
- No entanto, **não há**, ainda, uma **definição de "qualidade"** no escopo da avaliação de estudos empíricos.
- A literatura de ESBE sugere ***checklists*** de qualidade para **diferentes tipos de EE**, no entanto, cada uma das fontes pesquisadas, identifica um **conjunto ligeiramente diferente de questões** e não há um **padrão acordado** para esse tipo de avalição

Experimentos Controlados

- Oferecem maior nível de controle do processo num relacionamento de causa (tratamentos) e efeito (resultados)
- Cinco fases bem definidas
  - Definição de Objetivos
  - Planejamento
  - Execução
  - Análise e Apresentação
  - Empacotamento

Definições

- Há algumas abordagens para verificação da qualidade: abordagens simples, listas de verificação e escalas de qualidade [27][34][52].
- Para essa análise escolheu-se uma escala de qualidade:
  1. São instrumentos baseados em uma série de itens de qualidade, organizados numericamente e por categoria;
  2. Fornece uma estimativa quantitativa da qualidade global do estudo;
  3. Retornam valores contínuos e por isso podem ser correlacionados com outras variáveis do estudo
  4. Indicada quando tem-se itens de avaliação objetivos e subjetivos.

Instrumento de Avaliação da Qualidade

- Revisão da literatura (*ad-hoc*) em busca de *checklists* e critérios para avaliação da qualidade de EE.
- Os critérios básicos incluídos da lista proposta por Dyba e Dingsoyr [20], que aborda as questões de qualidade:
  1. Relatório: Verifica se está claro e objetivo o contexto e resultados do estudo;
  2. Rigor : Verifica se o método do estudo adotou uma abordagem completa e adequada;
  3. Credibilidade: Verifica se a descoberta do estudo foi bem apresentada e significativa;
  4. Relevância: Verifica se a descoberta do estudo é útil para a indústria e comunidade científica de ES.

Instrumento de Avaliação da Qualidade

- Organizado de acordo com as dimensões de garantia de qualidade de um experimento, propostos por Kitchenham et al. [36], e sugerido por Dieste et al. [18]:
  - (1)Contexto, (2)Delineamento experimental
  - (3)Condução do experimento e coleta de dados
  - (4)Análise dos dados, (5)Interpretação, e (6) Apresentação dos resultados
- Adicionados novos critérios relacionados às fases de condução de um experimento [65]
  - Definição de Objetivos, Planejamento,
  - Execução, Análise, Apresentação e Empacotamento
- Categorizados em critérios gerais e específicos

Instância do Instrumento de Avaliação da Qualidade

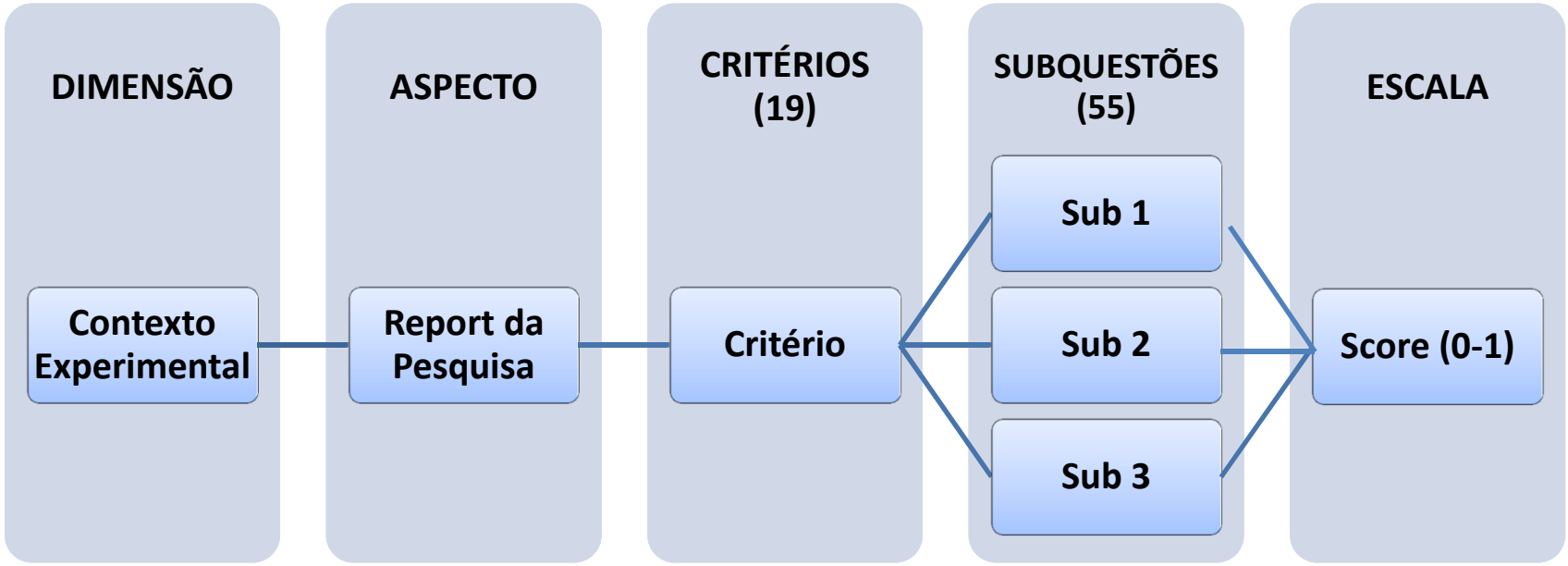


Figura 1 - Instância do instrumento de qualidade
Fonte: Elaboração pelo autor

Avaliação do Instrumento de Avaliação da Qualidade

- Realização de Estudo Piloto
- Métrica Utilizada foi o *Cohen's Kappa* [13]
- Interpretação sugerida por Landis JR e Koch GG [43]

Tabela 1 - Interpretação sugerida para os valores de *Kappa*

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

Fonte: Elaboração pelo autor

Tabela 2 - Resultados do Estudo Piloto

Papers	Avaliadores			
	1	2	3	4
PS089	B	C	B	B
PS100	D	D	D	C
PS242	B	C	B	B
PS256	C	C	D	C
PS836	D	D	D	D
KAPPA GERAL			**0.623**	
P-valor geral			< 0.001	

Classificação dos Itens de Qualidade

- Mecanismo de avaliação para mensurar a realidade proposta
- Escala do Tipo *Likert* [44]
- Escolheu-se uma escala de 3 pontos
  - Manter o equilíbrio na avaliação [12]
  - Confiabilidade da avaliação [11]

Tabela 3 – Legenda da escala *Likert*

0	Não Atende	Quando não há nada no trabalho que atenda ao critério avaliado
0,5	Atende Parcialmente	Quando há evidências que atendam parcialmente ao requisito avaliado
1	Atende Totalmente	Quando o trabalho apresenta atendimento total ao critério avaliado

Fonte: Elaboração pelo autor

Metodologia de Avaliação

Índice Global de Qualidade

- Forma de avaliar a qualidade
- Indicador quantitativo da qualidade de um estudo
- O cálculo do indicador é feito da seguinte forma:
  - CLASSIFICAÇÃO = (NT / TOTAL POSSÍVEL) X 100 = N (%).

Tabela 4 - Classificação da Qualidade segundo Beecham et al. [9]

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

Fonte: Elaboração pelo autor

Busca e Seleção dos Estudos

- Busca manual a partir dos três principais veículos internacionais da comunidade de ESE.
  - ESE
  - ESEM
  - ESEJ
- Artigos completos e não repetidos publicados no período de 1997 a 2012
- 876 papers do Mapeamento Sistemático, 104 experimentos controlados e 84 explicitaram o uso de mecanismos de suporte.

Alocação dos *Papers* aos Avaliadores

Processo de Extração dos Dados

- Cinco pesquisadores (dois alunos de MSc e três de PhD).
- Agrupados por conveniência em quatro duplas.
- As duplas foram formadas pela junção do autor com um aluno de MSc ou PhD.
- Papers foram organizados em planilha de acordo como ano de publicação.
- Função randômica do Excel para atribuir cada dupla aos grupos de estudo.
- Cada dupla avaliou 26 papers.
- Cada *paper* foi avaliado por no mínimo dois pesquisadores.
- A análise dos conflitos foi tratada em sessões de reunião de discussão.
- Em desacordos foi considerada a opinião de um terceiro pesquisador.
- O processo foi realizado sob a supervisão do orientador.

Coleta de Dados e Execução do Estudo

- Discussão dos procedimentos para coleta dos dados e execução do estudo (gerado um guia de extração).
- Localização da página, parágrafo ou seção do artigo, que motivou a escolha por determinado valor da escala.
- As duplas receberam as planilhas com o instrumento de avaliação ao mesmo tempo.
- Processo de coleta em aproximadamente 30 dias.
- Verificação do preenchimento e de dados discrepantes.
- Disponibilização dos dados brutos.

Índice de Confiabilidade dos Avaliadores

- Antes da discussão dos conflitos, foi medido o índice de concordância entre as duplas utilizando o coeficiente *Kappa*
- Interpretação sugerida por Landis JR e Kotch GG [43]

Tabela 5 - Índice de Concordância *Kappa*

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

Fonte: Elaboração pelo autor

Estatística Descritiva

- Análise estatística foi realizada usando o pacote IBM SPSS e uma significância estatística = 0,05
- Descrição dos dados e distribuição das frequências das variáveis

Tabela 6 - Frequências dos Experimentos Controlados

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

Fonte: Elaboração pelo autor

Índice de Qualidade por Critério

- Cálculo do Índice = (NT / Total Possível)X100 = N (%).

Tabela 7 - Frequências dos critérios melhor avaliados

Critérios	Médias
15. São mencionadas as ameaças à validade e como essas ameaças afetam os resultados e conclusões?	92,78%
3. Houve a descrição de pesquisas similares e como estas se relacionam com o estudo?	87,50%
17. As significâncias estatísticas são mencionadas com os resultados?	87,01%

Fonte: Elaboração pelo autor

Tabela 8 - Frequências dos critérios pior avaliados

Critérios	Médias
9. Os autores descreveram ou apresentaram alguma medida de concordância entre os avaliadores?	9,61%
11. As perdas e desistências de participantes ocorridas desde a seleção até o final do estudo foram descritas?	13,94%
19. A pesquisa apresenta ou indica a disponibilidade dos dados brutos?	16,34%

Fonte: Elaboração pelo autor

Análise dos Indicadores

Tabela 9 - Análise dos índices das variáveis

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

Fonte: Elaboração pelo autor

- Distribuição do Índice Global

Tabela 10 - Teste *One-Sample Kolmogorov-Smirnov*

Testes	Índice Global	Rigor Estatístico	Replicabilidade
p Value	0,584	0,123	0,909

- Testes Paramétricos
  - Anova, Post Hoc e Teste T

Resposta às Questões de Pesquisa

Q1. Qual é a evolução da qualidade dos estudos em relação aos mecanismos de suporte utilizados?

1. Houve evolução estatisticamente significativa em relação ao grupo que usou algum mecanismos de suporte.
2. Não houve evolução estatisticamente significativa em relação ao período avaliado (1997 a 2012).

Tabela 11 – Procedimentos – Diferença de Média por período avaliado

Mecanismo de Suporte	Procedimentos Estatísticos	*P Value*
Grupos (yes/not)	Teste T	*0,002*
Períodos Avaliados	Anova	*0,895*

Fonte: Elaboração pelo autor

Resposta às Questões de Pesquisa

Q2. Qual é a evolução da qualidade dos estudos em relação à replicabilidade?

Não houve evolução estatisticamente significativa na qualidade da variável ao longo do período avaliado (1997 a 2012)

Tabela 12 - Anova – Diferença de Média por período avaliado

Replicabilidade	Procedimentos Estatísticos	*P Value*
Períodos Avaliados	Anova	*0,663*

Fonte: Elaboração pelo autor

Resposta às Questões de Pesquisa

Q3. Qual é a evolução da qualidade dos estudos em relação ao rigor estatístico?

Não houve evolução estatisticamente significativa na qualidade da variável ao longo do período avaliado (1997 a 2012)

Tabela 13 - Anova – Diferença de Média por período avaliado

Rigor Estatístico	Procedimentos Estatísticos	*P Value*
Períodos Avaliados	Anova	*0,189*

Fonte: Elaboração pelo autor

Resposta às Questões de Pesquisa

Q4. Qual é a evolução da qualidade dos estudos no período avaliado?

Não houve evolução estatisticamente significativa na qualidade da variável ao longo do período avaliado (1997 a 2012).

Tabela 14 - Anova – Diferença de Média por período avaliado

Qualidade Global	Procedimentos Estatísticos	*P Value*
Períodos Avaliados	Anova	*0,819*

Fonte: Elaboração pelo autor

Resposta às Questões de Pesquisa

Q5. Qual é a evolução da qualidade dos estudos por veículo de divulgação avaliado?

Houve evolução estatisticamente significativa entre o *Journal* (ESEJ) e os outros dois locais (EASE e ESEM)

Tabela 15 - Anova – Diferença de Média por local avaliado

Qualidade Global	Procedimentos Estatísticos	*P Value*
Locais Avaliados	Anova	*0,000*

Tabela 16 - *Tukey Test* – Diferença de Média por locais avaliados

Média dos Locais	Locais Avaliados	*P Value*
EASE (55,21%)	ESEM	0,934
	ESEJ	**0,040**
ESEM (54,25%)	EASE	0,934
	ESEJ	**0,000**
ESEJ (57,15%)	EASE	**0,040**
	ESEM	**0,000**

Fonte: Elaboração pelo autor

Resposta às Questões de Pesquisa

Q5 - Variável Replicabilidade por local avaliado

Houve evolução estatisticamente significativa entre o *Journal* (ESEJ) e os outros dois locais (EASE e ESEM)

Tabela 17 - ANOVA – Diferença de média do índice de replicabilidade por local

Replicabilidade	Procedimentos Estatísticos	*P Value*
Locais Avaliados	Anova	*0,000*

Tabela 18 – *Tukey Test* – Diferença de média do índice de replicabilidade por local

Locais Avaliados	Locais por comparação	*P Value*
EASE	ESEM	0,535
	ESEJ	**0,001**
ESEM	EASE	0,535
	ESEJ	**0,000**
ESEJ	EASE	**0,001**
	ESEM	**0,000**

Fonte: Elaboração pelo autor

Resposta às Questões de Pesquisa

Q5 - Variável Rigor Estatístico por local avaliado

Houve evolução estatisticamente significativa entre o *Journal* (ESEJ) e os outros dois locais (EASE e ESEM)

Tabela 19 - ANOVA – Diferença de média por rigor estatístico

Rigor Estatístico	Procedimentos Estatísticos	*P Value*
Locais Avaliados	Anova	*0,000*

Tabela 20 - *Tukey Test* – Diferença de média por rigor estatístico

Locais por Comparação	Locais Avaliados	*P Value*
EASE	ESEM	0,334
	ESEJ	**0,001**
ESEM	EASE	0,334
	ESEJ	**0,003**
ESEJ	EASE	**0,001**
	ESEM	**0,003**

Fonte: Elaboração pelo autor

Limitações e Ameaças à Validade

- De acordo com a taxonomia apresentada por Wohlin et al. [67], as ameaças foram categorizadas em:
  - Ameaças à validade Interna
  - Ameaças à validade Externa
  - Ameaças à validade de Constructo
  - Ameaças à validade de Conclusão

Limitações e Ameaças à Validade

Ameaças à validade interna

- Confiabilidade do Instrumento
  - Treinamento aos avaliadores
  - Estudo Piloto
- Subjetividade do Instrumento
  - Revisão da Lista de Critérios
- Experimentos que realizaram mais de um tipo de estudo
  - Quatro *papers* realizaram mais de um tipo de estudo

Ameaças à validade externa

- Materiais Usados
  - Critérios Genéricos
  - Critérios Específicos

Limitações e Ameaças à Validade

Ameaças à validade de constructo

- Conceito de Qualidade [27][41][48]
  - Minimizar o viés
  - Maximizar a validade interna e externa
  - Interpretação correta dos Resultados

Ameaças à validade de conclusão

- Procedimentos Estatísticos
  - Pacote estatístico amplamente utilizado (IBM SPSS)
  - Significância Estatística = 0,05
  - Cohens Kappa
  - Distribuição dos Dados (Teste KS)
  - Anova, Post Hoc, Teste T

Recomendações para Trabalhos Futuros

- Investigar a validade do constructo do instrumento de qualidade apresentado
- Avaliação de qualidade de outros tipos de estudos
- Realizar uma Pesquisa Qualitativa
- Analisar o Contexto e Delineamento Experimental
- Replicação do Estudo
- Desenvolver um modelo/guia/processo que apoie a avaliação de qualidade de EE em ES

Conclusões e Contribuições

- Agregar conhecimentos sobre como foi e como estão os estudos empíricos em ES.
- Evolução do instrumento de avaliação utilizado.
- Compreensão sobre os conceitos que norteiam a construção de um modelo/processo/guia de suporte à avaliação da qualidade de EE.

Referências

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

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

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

[18] DIESTE, O. et al. **Quantitative determination of the relationship between internal validity and bias in software engineering experiments: consequences for systematic literature reviews**.

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

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

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

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

[34] KHAN, K. S. et al. **Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews**. [S.l.]: NHS Centre for Reviews and Dissemination, 2001.

[36] KITCHENHAM, B. A. et al. **Preliminary guidelines for empirical research in software engineering**. Software Engineering, IEEE Transactions on, 2002. v. 28, n. 8, p. 721–734.

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

[40] KITCHENHAM, B. et al. **Trends in the Quality of Human-Intensive Software Engineering Experiments: A Quasi-Experiment**. IEEE, 2013.

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

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

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

[59] SILVA, F. Q. DA et al. **Six years of systematic literature reviews in software engineering: An updated tertiary study**. Information and Software Technology, 2011. v. 53, n. 9, p. 899–913.

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

Obrigado!

| Questionamentos? |

Eudis Teixeira

eot@cin.ufpe.br

Centro de Informática U·F·P·E

Universidade Federal de Pernambuco