INTRODUÇÃO À ENGENHARIA DE SOFTWARE EXPERIMENTAL

AULA BASEADA EM MATERIAL DO PROF. MANOEL MENDONÇA, no Livro do Wholin e em outros artigos (Kitchenham, Pfleeger etc)

Material compilado a partir de:

- **Victor R. Basili, “The Role of Experimentation in Software** Engineering”, Keynote Speech at ICSE´1996;
- **Shari L. Pfleeger, “Evaluating Software Technologies”,** Tutorial at SBES´2002;
- **Guilherme H. Travassos, “Experimental Software** Engineering: An Introdution”, ESELAW´2005;
- **Erika Nina Höhn, “Revisão Sistemática”, USP/ICMC.**
- **Márcio Barros, Guilherme Travassos et al. “Métodos** Estatísticos aplicados em Engenharia de SoftwareExperimental”, SBES'2006.
- (K,P&P) **Kitchenham, Pickard and Pflee**ger, Case Studies for Method and Tool Evaluation, IEEE Software, July 1995

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Definindo o Tipo de Estudo Experimental
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Tipos de Estudos Experimentais
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

O Que é Engenharia de Software ?

- Engenharia de software é a disciplina que estuda o desenvolvimento e a manutenção de software em escala industrial.
  - Técnicas
  - Metodologias
  - Processos
  - Ferramentas
- ... para gerência, desenvolvimento, manutenção, reengenharia (etc.) de software

Algumas Necessidades Fundamentais em Engenharia de Software ...

- Adotar novas tecnologias
- Testar se uma nova tecnologia é útil
- Avaliar o impacto de uma tecnologia

COMO DECIDIR O QUE FAZER?

- *Perguntar a um perito*
- *Pesquisar na literatura*
- *Seguir a prática da indústria, ou ....*

SER EXPERIMENTAL

- Fazer uma revisão sistemática
- Fazer um levantamento de campo (*survey)*
- Fazer um estudo de caso
- Fazer um experimento controlado

MÉTODO CIENTÍFICO

- Quais são as suas metas e qual é a sua situação?
- Existe evidência na literatura e/ou na indústria e como esta evidência se aplica à SUA situação?
- Se não existe evidência suficiente, que tipo de avaliação experimental você deve fazer?

O Uso da abordagem científica para o desenvolvimento, evolução e manutenção de software é o que chamamos de:

Engenharia de Software Experimental

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Tipos de Estudos Experimentais
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

O Paradigma Experimental

- O Paradigma experimental de uma disciplina evolui pela aplicação do ciclo: modele, experimente, aprenda;
- Normalmente começa com a observação e o registro do que é observado, e evolui para a manipulação de variáveis controláveis e a observação de seu efeito em variáveis de interesse.

Modelos, Experimentação e Aprendizado: um paradigma experimental

- Para entender uma disciplina é necessário a construção de **modelos, não só de produtos mas também de processos e** domínios de aplicação;
- Para testar se a compreensão está correta é preciso testar esses modelos, isto implica em **experimentação;**
- Ao se analisar resultados experimentais, **aprendemos e** encapsulamos esse conhecimento em modelos mais sofisticados;
- Este paradigma experimental é usado em muitas áreas de conhecimento: física, medicina, química, manufatura, etc.

Experimentação x Disciplinas

- As diferenças na aplicação do paradigma experimental nos vários campos de conhecimento são ditadas pelos objetos de estudo, as propriedades do sistemas que os contêm, as relações entre os objetos e o sistema, e a cultura da disciplina;
- Isto impacta em:
  - como modelos são construídos
  - como experimentação é feita.
- Praticamente todas as disciplinas científicas têm um campo próprio que estuda como fazer experimentação nessa disciplina.

O Paradigma Experimental em Física

- Física visa entender e predizer o comportamento do universo físico;
- Há dois grupos bem definidos de pesquisadores, os teóricos e os experimentalistas, e progride a partir do interrelacionamento entre estes dos grupos;
- Teóricos constroem modelos para explicar o universo baseados em teorias sobre variáveis essenciais e sua interação determinada em experimentos anteriores;
- Suas teorias predizem o resultado de eventos mensuráveis;
- Experimentalistas observam, medem, e experimentam para provar ou refutar uma hipótese ou teoria; também exploram novos domínios.

O Paradigma Experimental em Medicina

- Também possui dois grupos bem definidos: os práticos (ou profissionais) e os pesquisadores; existe um claro relacionamento entre eles.
- Os pesquisadores visam entender o funcionamento do corpo humano para prever os efeitos de procedimentos e drogas;
- Os práticos aplicam o conhecimento ganho para definir processos de tratamento do corpo humano;
- Começou como uma forma de arte e só evoluiu quando começou com o ciclo de observação, construção de modelos, experimentação e aprendizado;
- Dificuldades:
  - Estudos variam de experimentos controlados a estudos de caso
  - Variância do ser humano dificulta a interpretação de resultados
  - É trabalho e complexo para se obter dados
- Nem por isto a medicina deixou de evoluiu ao longo do tempo !

O Paradigma Experimental em Engenharia de Software

- Como tantas outras disciplinas, a engenharia de software necessita de um ciclo próprio de construção de modelos, experimentação e aprendizado;
- Engenharia de software (também) é uma disciplina de laboratório;
- Deve existir práticos cujo papel é construir cada vez “mais arato” e “mais rápido” sistemas cada vez “melhores”, utilizando o conhecimento disponível;
- Deve existir pesquisadores que tentem entender a natureza dos processos e produtos de software e da relação entre os dois no desenvolvimento e manutenção de sistemas;
- Comparado com outras disciplinas, a Engenharia de Software é uma disciplina muito nova (1967), e a área de experimentação ainda está na sua infância.

O Paradigma Experimental em Engenharia de Software

- A relação entre práticos e pesquisadores é altamente simbiótica:
  - Pesquisadores precisam de laboratórios para observar e manipular variáveis, a indústria é o ambiente ideal;
  - Práticos precisam entender como melhor construir e manter seus sistema e os pesquisadores são quem melhor podem auxiliar nesta tarefa.

O Estado da Disciplina Experimental em ES (1)

Onde está o estado da disciplina de modelagem, experimentação e construção de modelos em ES?

- No começo ...
- O Principal conferência mundial é a “International Symposium on Empirical Software Engineering and Measurement” (ESEM, antigo ISESE). O próximo evento (5º.) será em Banff, Canada.
- O Principal evento “nacional” é o “Experimental Software Engineering Latin-American Workshop”(ESELAW).

O Estado da Disciplina Experimental em ES (2)

- Uso de Modelos
  - Modelos empíricos de custo e ocorrência de defeitos
  - Modelos de processos
  - Modelos de produtos (uso intensivo de modelos matemáticos)
- Pouca Experimentação
  - Teóricos e práticos veem seus modelos com autoevidentes e que não precisam ser testados
  - Para qualquer modelo e tecnologia precisa-se testar as condições em que eles funcionam adequadamente.

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Tipos de Estudos Experimentais
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

Natureza da Engenharia de Software

- Software é desenvolvimento e não produção
  - As fábricas de software quebram um pouco este paradigma
- A maioria das tecnologias são intensivamente humanas;
- Software, domínio, e culturas variam muito. Os software não são iguais.
  - Existe um número enorme de variáveis envolvidas;
  - Seus efeitos são mal compreendidos e modelados;
- Atualmente:
  - Existe pouca compreensão dos limites de tecnologias
  - Existe pouca análise e experimentação controlada

Paradigmas de Pesquisa em ES

- Paradigma Analítico
  - Baseado em matemática
  - Propõe uma teoria formal ou um conjunto de axiomas
  - Deriva matematicamente um conjunto de resultados
  - Está no cerne da ciência da computação e expõe a herança matemática de nossa área
- Paradigma Experimental
  - Observa o mundo ou soluções existentes;
  - Propõe um modelo de comportamento ou solução melhor;
  - Mede e analisa modelos experimentalmente
  - Valida (ou refuta) hipóteses e modelos
  - Repete o processo para evoluir o conhecimento

Causa

Efeito

Observação

Tratamento

Resultado

Variável independente

Execução do experimento

Variável dependente

O Quê o Paradigma Experimental Envolve

- Observação
- Projeto experimental
- Coleta de dados
- Análise qualitativa ou quantitativa
- Avaliação do objeto de estudo (processo ou produto)

Análise Qualitativa x Análise Quantitativa

- Análise Quantitativa
  - Medição controlada (normalmente intrusiva)
  - Objetiva
  - Orientada a Verificação
- Análise Qualitativa
  - Observação naturalística (normalmente não intrusiva)
  - Entrevistas e questionários (normalmente intrusivas)
  - Subjetiva
  - Orientada a descoberta

Tipos de Estudos em ESE

- Um estudo é o ato de descobrir algo desconhecido ou de testar uma hipótese, pode incluir todos os tipos de análise quantitativa e qualitativa.
- Estudos Experimentais
  - Voltado ao teste de hipóteses
  - São geralmente quantitativos
  - Experimento controlados ou quasi-experimentos
- Estudos Observacionais
  - Voltado à compreensão e descoberta
  - Geralmente são mais qualitativos que quantitativos
  - Pesquisa qualitativa ou semi-qualitativa, entrevistas e levantamentos

Devem responder a duas questões

- O quê estudar e porquê estudar?
  - Estudos de fatores humanos
  - Estudos de projetos e produtos
  - Estudo de métodos e técnicas
  - Estudos da organização e de seus processos
- Que tipo de estudo experimental realizar?
  - Estudos In Vivo
  - Estudos In Vitro
  - Estudos In Virtuo
  - Estudos In Silico

O que estudar

- Qual será o objeto de estudo
  - Ex. um processo ou produto
- Qual é a finalidade do estudo?
  - Caracterizar (o quê está acontecendo?)
  - Avaliar (é bom?)
  - Prever (é possível estimar o comportamento futuro?
  - Controlar (é possível manipular eventos e situações?)
  - Melhorar (é possível melhorar eventos e situações?)
- Qual é o foco?
  - Quais aspectos e variáveis do objeto de estudo são de meu interesse?
- Qual é a perspectiva?
  - Quais são os grupos de pessoas interessadas?

Tipos de Estudo Experimentais

- In Vivo
  - Envolve pessoas no seu próprio ambiente de trabalho em condições realistas de trabalho
- In Vitro
  - Realizado em condições controladas tais como em um laboratório ou um grupo fechado
- In Virtuo
  - Realizado em condições controladas nas quais os participantes interagem com modelos computacionais da realidade (simuladores)
- In Silico
  - Participantes e o mundo real são descritos por modelos computacionais (dinâmica de sistemas)

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Tipos de Estudos Experimentais
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

Comece por definir o objetivo

- O que será investigado e por que investigar.
- Exemplo: Avaliar se o método de projeto XYZ produz resultados melhores que o método ABC.
- Também chamado de “hipótese experimental” (K,P&P)
- Tenha em mente a razão ou a finalidade para a qual você pretende fazer essa avaliação e que os dados que serão coletados podem confirmar ou refutar que XYZ é melhor que ABC.

Tipos de pesquisa experimental

Estudo primário

- Levantamento de Campo (Survey)
  - Trabalho de campo de levantamento de opinião de várias pessoas (caracterizando o universo consultado)
- Estudo de caso
  - Aplicação do objeto de estudo em um pequeno número de casos (caracterizando o ambiente de aplicação)
- Pesquisa Ação
  - Aplicação do objeto de estudo em um pequeno número de casos (caracterizando o ambiente de aplicação)
  - Todavia o objeto de estudo também estará sendo desenvolvido, adaptado, ou evoluído durante o estudo
- Experimento Controlado
  - Aplicação do objeto de estudo e do tratamento de controle em vários casos sob condições fortemente controladas
- Revisão Sistemática (meta-análise)
  - Sintetizar a evidência, identificando, avaliando e interpretando todas as pesquisas disponíveis em relação a uma questão específica.

Estudo secundário

Definir os Objetivos em Termos de Variáveis Mensuráveis

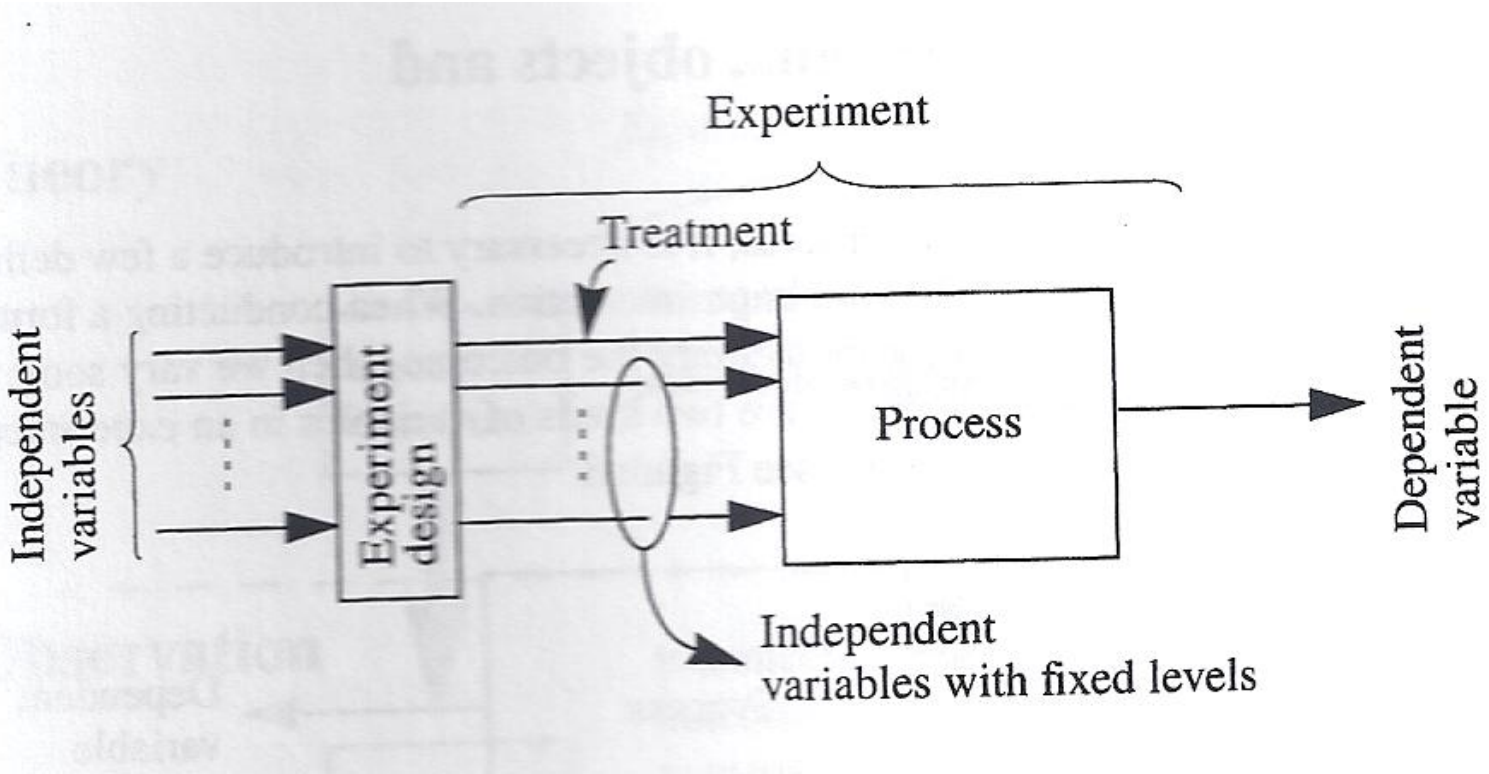
- Defina Objetivos em termos quantitativos
  - Em vez de: Avaliar se o método de projeto XYZ produz resultados melhores que o método ABC
  - Prefira: O código produzido pelo método XYZ possui um menor número de defeitos por milhares de linhas de código fonte que o método ABC
- Defina a relação entre conceitos e medidas
  - No exemplo anterior o objetivo é medir a qualidade e a media “número de defeitos” é usada para isto.
  - A relação entre o que se quer e o que se mede deve ser documentada, e eventualmente explicitada em modelos de relacionamento.

Identificar as variáveis

- Variável dependente: é o resultado, é a variável estudada para ver o efeito das mudanças nas variáveis independentes
  - Ex. número de defeitos por milhares de linhas de código
- Geralmente há apenas uma variável dependente em um experimento.
- Todas as variáveis que são manipuladas e controladas são chamadas de variáveis independentes.

Variáveis independentes

- As variáveis independente que terão seus efeitos de mudança estudados são chamadas de fatores.
  - Um tratamento é um valor particular de um fator.
- As outras variáveis independentes são mantidas com um valor fixo.
- Controle: variável que será controlada no estudo causa-efeito .
  - Tratamento: o método XYZ que será avaliado
  - Controle: o “status quo”, i. é, o método com o qual você quer comparar.
- De estado: medidas que descrevem o sujeito experimental, objetos e condições. Ex. experiência do projetista
- De contexto: variável de estado que assume somente um valor no estudo experimental, ex., o tamanho do código é pequeno, os participantes do experimento são todos estudantes.



Um experimento : (Wohlin et all)

EXEMPLO

- Queremos estudar o efeito de um novo método de desenvolvimento sobre a produtividade do pessoal. Ex. um método oo está sendo introduzido.
- Variável dependente: produtividade
- Variáveis independentes: o método de desenvolvimento, a experiência do pessoal, ferramentas de apoio e o ambiente.

Observações e outras terminologias

- Sujeitos experimentais e objetos experimentais são as pessoas ou coisas envolvidas em um experimento. Ex. pessoas que usam um método ou ferramenta (s.e) e programas, algoritmos e problemas para os métodos ou ferramentas são aplicados.
- Muitas vezes se usa uma variável dependente substituta (*surrogate*) para medir, ao invés de uma medida direta. Isso pode prejudicar a qualidade dos resultados obtidos. Ex. medir confiabilidade contando-se o número de falhas reveladas durante o teste.

Exemplo: medir produtividade : para cada projeto participando do estudo, medir a produtividade em pontos por função

Tipos das variáveis

TABLE 1
COMPARISON OF PRODUCTIVITY MEASURES

Variable	Method A	Method B
Productivity (function points/hour)	0.054	0.237
Size (function points)	118	168
Team experience (years)	1	1
Project management experience (years)	1	1
Duration (months)	10	9
Function point	25	27

Determine o grau de controle sobre as variáveis

- Determinar o grau de controle sobre as variáveis independentes
  - Se a coleta de dados ocorre depois do fato e não se tem nenhum controle , então deve-se fazer um survey;
  - Se você os dados são coletados enquanto o desenvolvimento ou manutenção está acontecendo, mas há controle básico sobre variáveis , então deve-se realizar um estudo de caso;
  - Se o objeto de análise evolui enquanto os dados estão sendo coletados, então deve-se realizar uma pesquisa-ação;
  - Se há controle sobre a maioria das variáveis e controle sobre os participantes – você deve realizar um experimento controlado.

Exemplo

- Suponha que se deseja avaliar o efeito de um método de projeto sobre a qualidade do software resultante
- Se você não tem controle sobre quem está usando qual método, então deve-se realizar um estudo de caso para documentar os resultados;
- Se você pode controlar quem usa cada método, quando e como estes métodos são usados, então deve-se realizar um experimento controlado.

Experimentos *in-vivo x in-vitro*

- Experimentos in-vitro são feitos em laboratórios, simulando a forma como eles aconteceriam no mundo real;
- Experimentos in-vivo são feitos no mundo real e monitorados à medida em que o uso do objeto de estudo realmente ocorre;
- Em engenharia de software, geralmente experimentos controlados são feitos in-vitro e estudos de caso são feitos in-vivo.

Outras considerações

- Em experimentos controlados manipulam-se as amostras sobre as variáveis de estado
  - Se experiência é importante você pode incluir pessoas com experiências bem distribuídas.
- Em estudos de caso manipulam-se as amostras das variáveis de estado
  - A experiência é importante mas ao escolher um certo projeto, você escolhe um conjunto de pessoas com experiência características específicas, preferivelmente a média de sua organização.
- Em experimentos controlados você pode facilmente definir a variável experimental, a de controle, e as variáveis de estado.
  - As variáveis de estados não devem variar ou devem ter variações igualmente distribuídas entre os tratamentos (controle e experimental)

Fatores a se Considerar

Fator	Experimentos	Estudos de Caso	Survey
Nível de Controle	Alto	Baixo	Baixo
Dificuldade de Controle	Alto	Médio	Médio
Facilidade de Replicação	Alto	Baixo	Alta
Custo de Execução	Alto (in-vivo) Médio (in-vitro)	Médio	Baixo Médio
Riscos à validade	Baixo (in-vivo) Médio (in-vitro)	Médio	Baixo Médio

Um framework para experimentos quantitativos (Basili, Selby e Hutchens)

- Estudos de projeto-único: examina objetos tratados por um único time e um único projeto.
- Estudos multiprojetos: examina objetos tratados por um único time e um conjunto de projetos.
- Estudos de projeto replicados: examina objetos tratados por um conjunto de times e um único projeto.
- Estudos de projeto-único agrupados: examina objetos tratados por um conjunto de times e um conjunto de projetos.

Considerando a formalidade dos projetos experimentais e o framework (K,P&P):

- Estudo de caso: se o estudo foca em um único projeto (não é possível ter um experimento formal sem replicação)
- Estudo de caso ou experimento formal: o estudo envolve muitos projetos ou único projeto que é replicado várias vezes.
- Experimento formal ou survey: o estudo analisa muitos time e muitos projetos

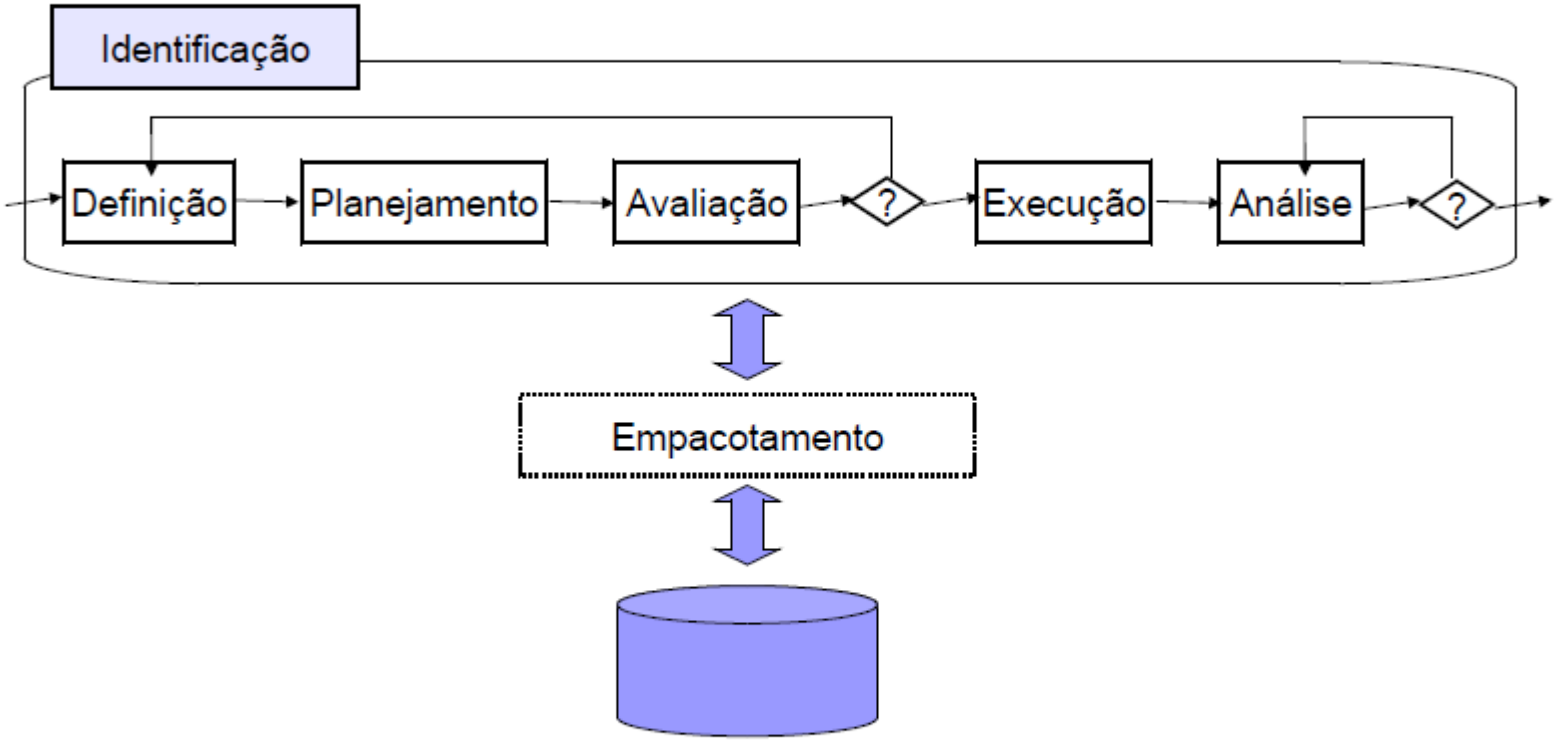
Considerando a formalidade dos projetos experimentais e o framework (K,P&P):

- Os estudos também são diferenciados pela escala
- Estudos formais precisam ser controlados cuidadosamente → pequena escala.
- Estudos de caso investigam o que está acontecendo em um projeto "típico"
- Suveys capturam o que está ocorrendo em grupos grandes.

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Tipos de Estudos Experimentais
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

Processo de Experimentação



Proposta de K,P&P

CHECKLIST FOR CASE-STUDY PLANNING

This checklist, along with the seven steps to design and administer case studies, will help you undertake a valid investigation.

Case study context

1. What are the objectives of your case study?
2. What is the baseline against which you will compare the results of the evaluation?
3. What are your external project constraints?

Setting the hypothesis

4. What is your evaluation hypothesis?
5. How do you define, in measurable terms, what you want to evaluate (that is, what are your response variables and how will you measure them)?

Planning

6. What are the experimental subjects and objects of the case study?
7. When in the development process or life cyle will the method be used?
8. When in the development or life cycle will the response variables be measured?

continuação

Validating the hypothesis.

9. Can you collect the data you need to calculate the selected measures?
10. Can you clearly identify the effects of the treatment you want to evaluate and isolate them from the other influences on the development?
11. Have you taken adequate procedures to ensure that the method or tool is being correctly used?
12. If you intend to integrate the method or tool into your development process, is the method or tool likely to have an effect other than the one you want to investigate?
13. Which state variables or project characteristics are most important to your case study?
14. Do you need to generalize the result to other projects? If so, is your pro posed case study project typical of those projects?
15. Do you need a high level of confidence in your evaluation result? If so, do you need to do a multiproject study?

Analyzing the results

16. How are you going to analyze the case study results?
17. Is the type of case study going to provide the level of confidence you require?

IDENTIFICAÇÃO

INTRODUÇÃO

- Parte inicial da documentação experimental
- Deve ser atualizada à medida que mais informações fiquem disponíveis.
- Deve conter:
  - Informações básicas do experimento
  - Introdução ao problema
  - Caracterização do experimento

Informações básicas

- Título
- Tema e Área Técnica
- Autores e suas afiliações
- Local e data
- Informação para obter acesso ao repositório e aos dados experimentais (pacote experimental)

Introdução ao Problema

- Caracterização do Problema
- Descrição textual sucinta do problema que será estudado
- Resultados anteriores e/ou conhecimento vigente na área
- Contexto do trabalho
- Resultados esperados

Caracterização do estudo experimental

- Tipo de estudo
- Objeto de estudo
- Domínio
- Objetivo
- Linguagem
- Glossário
- Número de execuções e replicações.

1. IDENTIFICATION

Title: PBR Replication – R1

Theme: Verifying a reading technique that aims to detect defects on Requirement Documents

Technical Area: Requirements Document Inspection

Author: José Carlos Maldonado e Sandra Fabbri

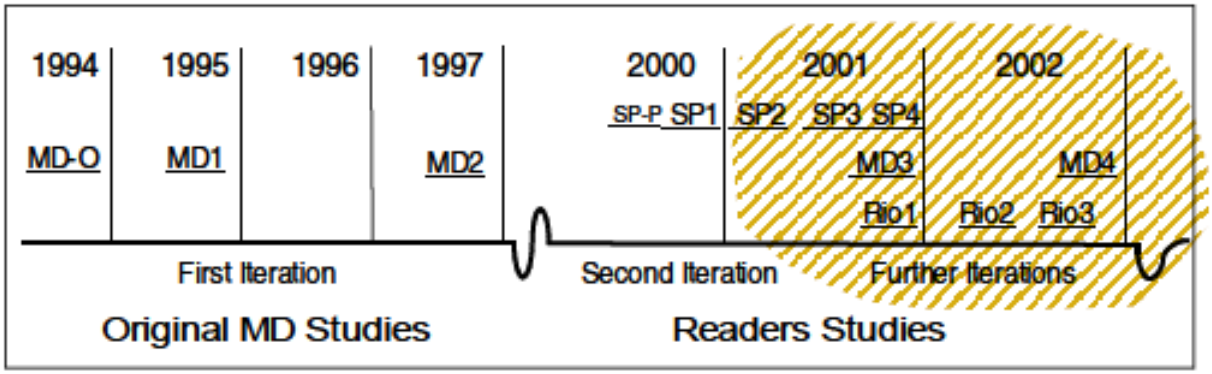
Affiliation: ICMC-USP e DC-UFSCar

Local: ICMC-USP

Data: Dez/2000

2. INTRODUCTION

This example describes an experiment that was based on another one, named Original PBR experiment, conducted at the University of Maryland, that compared the effectiveness of teams of subjects using PBR to the effectiveness of teams of subjects using their normal technique for detecting defects in a requirements document.
In that study the treatments allowed multiple variables to be studied, which provided some solid evidence that PBR was effective for inspection teams. The Original study left some open questions about PBR that were of interest. We were able to refine the experimental design and goals to investigate other related variables



PBR Studies Timeline

3. CHARACTERIZATION

Type: In Vitro

Domain: Information System

Language: Portuguese (explanations)/English (material)

Partners (Institutions, Address, Phone, Fax, e-mail and url):
ICMC-USP e DC-UFSCar

Links: http://www.labes.icmc.usp.br/readers/

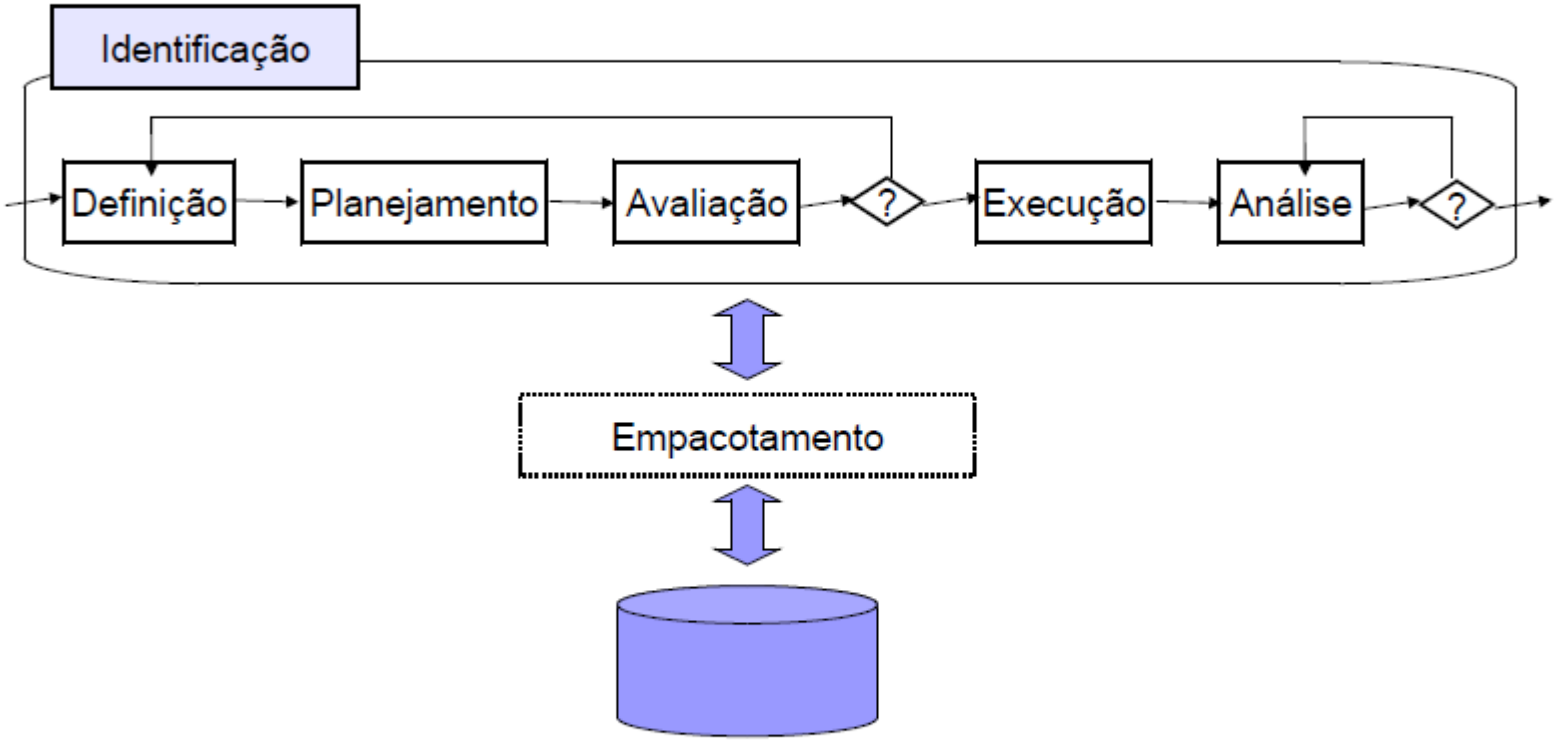
Estimated accomplishing

Estimated replication numbers: 4

Glossary

DEFINIÇÃO

Processo de Experimentação



DOIS PASSOS

- Definir os Objetivos do Estudo
  - Sobre o que se quer aprender, qual o objetivo da análise, quais aspectos de interesse, qual a finalidade do estudo, sob qual ponto de vista e em que contexto o estudo será feito
- Traduzir os Objetivos em Questões, Variáveis e Métricas
  - Identifique que questões você quer responder para atingir seu objetivo
  - Use o objetivo e as questões para identificar as variáveis independentes e dependentes do seu estudo.
  - Determine as métricas pelas quais as variáveis serão medidas.

Objetivos

- O seguinte gabarito ou diretriz, proposto por Basili no modelo GQM, pode ser utilizado para definição do objetivo específico do experimento

Analise	Objeto da análise
Com a finalidade de	Caracterizar ou avaliar
Com respeito a	Variável foco da análise
Do ponto de vista de	Grupo interessado no resultado
No contexto de	Definição do contexto de estudo

- Caracterizar ~ estudo observacional
- Avaliar ~ estudo experimental controlado e quantitativo

Derive Questões e Variáveis

- Dada uma meta como:

  Analisar *o método de projeto XYZ* **com a finalidade de** *avaliá-lo* **com respeito à** *eficiência e eficácia* **do ponto de vista do** *gerente de projeto* **no contexto de** *projetos típicos de minha empresa***.**

- Deve-se responder as seguintes perguntas:
  - Comparado com indicadores/referenciais XYZ será avaliado?
    - Método ABC
  - Que aspectos de eficácia e eficiência serão utilizados para comparar XYZ com ABC?
    - Nível de entrelaçamento de interesses, nível de conformidade com padrões, complexidade dos artefatos, tempo de projeto,etc.
  - Que outras variáveis independentes podem afetar estas variáveis?
    - Experiência do projetista, conformidade de uso do método

Derive métricas para as variáveis

- Toda variável relevante necessita ser medida.
- A medição só pode ser materializada por meio de uma métrica.
- Uma métrica relaciona um conceito a uma medida (um símbolo que quantifica este conceito para o determinado objeto de estudo).
- A discussão sobre a qualidade/adequação de uma métrica usada como uma medida para uma variável deve ser feita em separado da discussão sobre a qualidade das medições executadas.

Variáveis e seus Valores

- As variáveis de um estudo podem ser:
  - Categóricas: os valores representam tipos, formas eprocedimentos
  - Numéricas: os valores representam doses ou níveis de aplicação da variável
- Os valores das variáveis são coletados em escalas:
  - Existem diversas escalas para coleta e representação destes valores: nominal, ordinal, intervalar e razão
  - As escalas determinam as operações que podem ser aplicadas sobre os valores das variáveis

Escala Nominal

- Os valores de uma escala nominal representam diferentes tipos de um elemento, sem interpretação numérica e de ordenação entre eles
- • Exemplos em software incluem:
  - Diferentes medidas de tipos de métodos (XYZ e ABC)
  - Diferentes linguagens de programação (Java, C++, C#, Pascal, ...)
- A escala não nos permite dizer, por exemplo, que Java é menor que C#

Escala Ordinal

- Os valores de uma escala ordinal representam diferentes tipos de um elemento que podem ser ordenados, ainda que sem qualquer interpretação numérica
- Exemplos em software incluem:
  - Diferentes níveis no CMMI (Nível 1, ..., Nível 5) ou MPS.BR (Nível G, ..., Nível A)
  - Diferentes graus de coesão (funcional, procedimental, temporal, sequencial, ...)
- A escala permite dizer que, no CMMI, “Nível 2” é menor do que “Nível 3”, mas não permite dizer que a diferença de qualidade entre empresas do “Nível 2” e empresas do “Nível 3” é a mesma que entre empresas do “Nível 3” e“Nível 4”

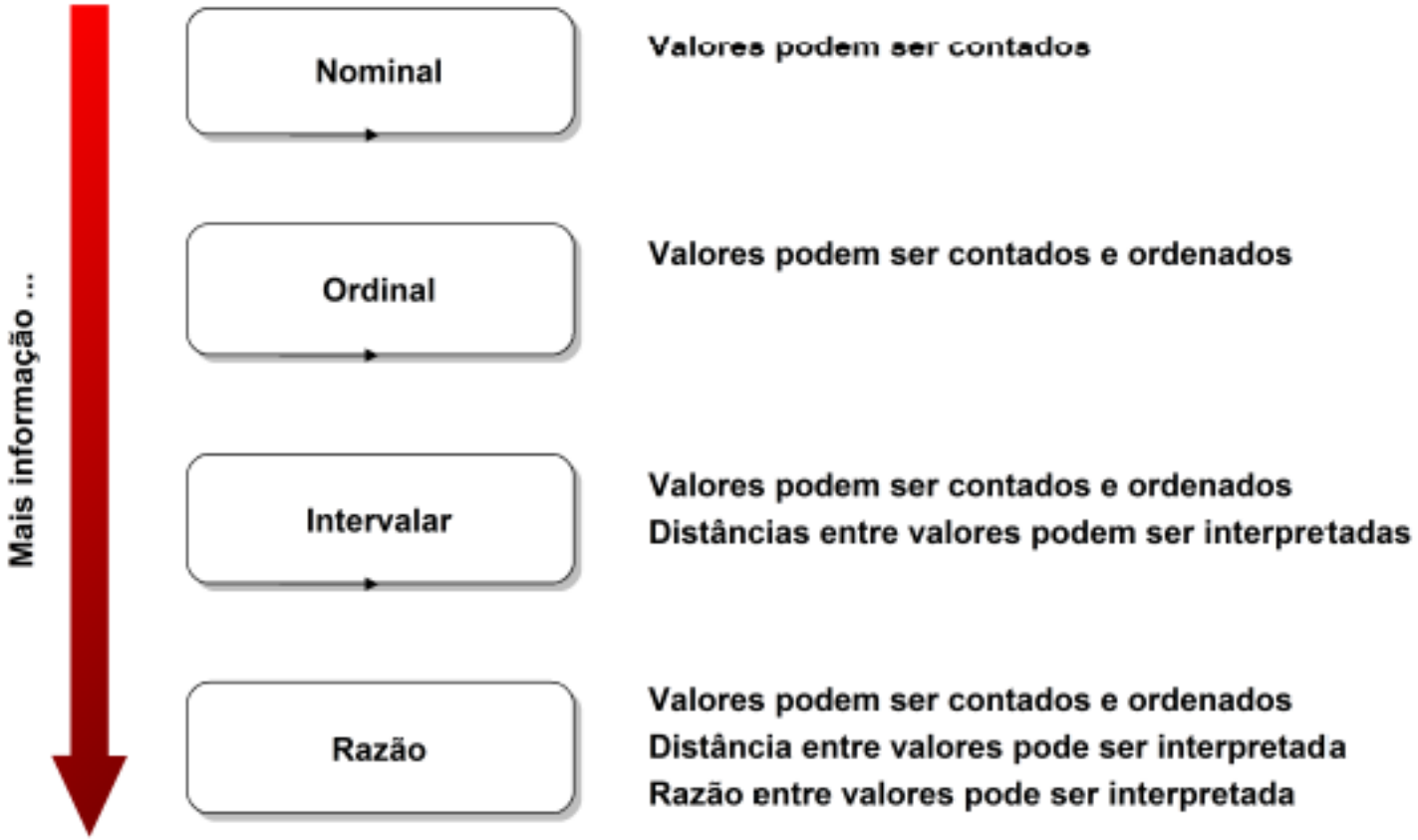
Escala Intervalar

- Os valores de uma escala intervalar podem ser ordenados e distâncias entre valores consecutivos possuem a mesma interpretação, porém a razão entre estes valores não tem significado
- Por exemplo: embora possamos dizer que 2006 é um ano após 2005 e um ano antes de 2007, não faz sentido calcular a razão entre 2006 e 2007.
- Isto é possível porque toda escala intervalar possui um zero arbitrário (no caso das datas, o ano zero)

Escala Razão

- Os valores de uma escala razão podem ser ordenados, distâncias entre valores consecutivos possuem o mesmo significado e a razão entre valores pode ser interpretada
- Exemplos em software incluem o tamanho de um sistema, o esforço necessário para a sua construção e o tempo de realização do projeto que resultou no sistema
- A escala permite dizer, por exemplo, que um software com X linhas de código é duas vezes menor que um software de 2X linhas de código

Informação nas Escalas



Exemplo:
Definição de um Estudo Experimental

4. EXPERIMENTAL STUDY DEFINITION

Object of Study: PBR technique

Global Objective: To evaluate PBR effectiveness for detecting defects in relation to Checklist

Specific Aims:

Analyze.....	the PBR and Checklist techniques
For the purpose of.....	evaluation
With respect to.....	effectiveness and efficiency
From the point of view of..	the researcher
In the context of.....	undergraduate students

Quality Focus: Requirements Document Quality

Context: The study will be conducted in the academic environment and undergraduate students will be the subjects of the experiment

4. EXPERIMENTAL STUDY DEFINITION

Questions and Metrics:

Questions:

- O1') Do PBR teams detect a more defects than Checklist teams?
- O2') Do individual PBR or Checklist reviewers find more defects?
- O3') Does the reviewer's experience affect his or her effectiveness?
- R1) Do individual reviewers using PBR and Checklist find different defects?
- R2) Do the PBR perspectives have the same effectiveness and efficiency?
- R3) Do the PBR perspectives find different defects?

Questions O1' - O3' came from the Original study and were modified to compare PBR to checklist rather than to ad hoc.
Questions R1 - R3 were open questions about PBR that were not specifically studied in the original experiment.

Metrics:

- **Defects Found**: The number of unique defects found by one or more subjects (i.e. each defect is counted only one time regardless of how many subjects find the defect);

- **Occurrences of a Defect**: This metric represents the number of times the defect is found, (assuming each subject has the chance to find the defect). The maximum number of occurrences for a defect is the number of inspectors in a group.

$$\mathbf{TotalOc} = \Sigma\,(x_i)\,,\ i = 1..n$$ where x_i is the number of defects found by subject *i*.

4. EXPERIMENTAL STUDY DEFINITION

Questions and Metrics:

Metrics:

- **Effectiveness**: The average percentage of defects found by a group of subjects. It is calculated as:

  $(\Sigma\ (x_i/y)*100)/n$, i = 1..n where *xi* is the number of defects found by the subject *i*, *y* is the total number of defects in the document and *n* is the number of subjects in the group.

- **Efficiency**: The average defects found by each subject per hour. It is calculated as:

  $(\Sigma\ (x_i/k_i))/n$, i = 1..n where *xi* is the number of defects found by the subject *i*, *ki* is the effort (in hours) used by subject *i* and *n* is the number of subjects in the group.

Questions that can not be answered by the experimental study:

Open Questions:

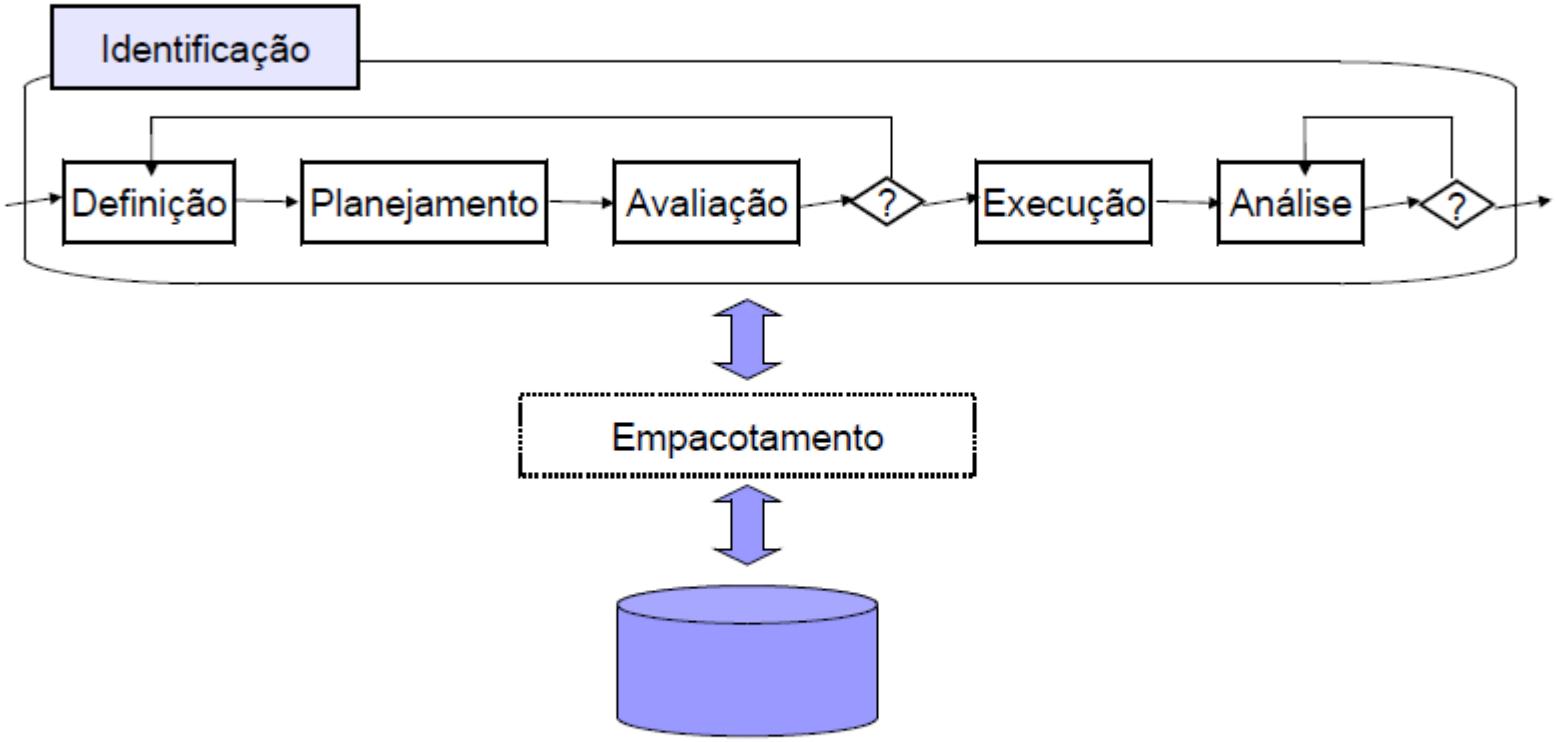
- How effective is the training?
- Are the subjects really following the techniques?

Recapitulando: Definição do Estudo

- Defina os Objetivos do seu estudo
  - Sobre o quê você quer aprender, qual o objeto de análise, quais são os aspectos de interesse, qual a finalidade do estudo, sob qual ponto de vista e em que contexto o estudo será feito?
- Traduza seu Objetivo em Questões, Variáveis e Métricas
  - Identifique que questões você quer responder para atingir seu objetivo.
  - Use o objetivo e as questões para identificar as variáveis independentes e dependentes de seu estudo.
  - Determine as métricas pelas quais estas variáveis serão medidas.

Planejamento

Processo de Experimentação



Organização

- Formulação das hipóteses
- Identificação das variáveis dependentes (resposta)
- Identificação das variáveis independentes (fatores)
- Seleção dos participantes
- Projeto do estudo
- Definição dos instrumentos
- Análise das ameaças à validade do estudo

HIPÓTESES

- Uma hipótese é uma teoria ou suposição que pode explicar um determinado comportamento de interesse da pesquisa.
- Um estudo experimental tem como objetivo colher dados, em um ambiente controlado, para confirmar ou negar a hipótese.

Testes de Hipóteses

- Hipóteses avaliadas por testes estatísticos definidos pelos pesquisadores da estatística inferencial
- Normalmente são definidas duas hipóteses
  - **Hipótese nula (H0):** indica que as diferenças observadas no estudo são coincidentais, ou seja, é a hipótese que o analista deseja rejeitar com a maior significância possível
  - **Hipótese alternativa (H1**): é a hipótese inversa à hipótese nula, que será aceita caso a hipótese nula seja rejeitada
- Os testes estatísticos verificam se é possível rejeitar a hipótese nula, de acordo com um conjunto de dados observados e suas propriedades estatísticas

Hipóteses e Variáveis

- Hipóteses levam à definição de variáveis
- Variáveis independentes (ou fatores, quando controladas)
  - Referem-se à entrada do processo de experimentação, podendo ser controladas durante este processo
  - Representam a causa que afeta o resultado do processo de experimentação. Quando é possível seu controle, os valores são chamados de "tratamentos"
- Variáveis dependentes
  - Referem-se à saída do processo de experimentação, sendo afetadas durante o processo de experimentação
  - Representam o efeito da combinação dos valores das variáveis independentes (incluindo os fatores). Seus possíveis valores são chamados de "resultados"

Exemplo:
Definição de Hipóteses e Seleção de Variáveis

5. PLANNING (detailed)

Hypothesis Formulation:

Example 1

- **O1': Do PBR teams detect more defects than Checklist teams?**

  H0: There is no difference in the defect detection rates of teams applying PBR compared to teams applying the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer has only random effects on team scores.

  Ha: The defect detection rates of teams applying PBR are higher compared to teams using the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer decreases the effectiveness of the team

5. PLANNING (detailed)

Hypothesis Formulation:

Example 2

- **O2': Do individual PBR or Checklist reviewers find more defects?**

  Group effect (RT X DOC interaction)

  H0: There is no difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency.

  Ha: There is a difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency

  Main effect RT

  H0: There is no difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.

  Ha: There is a difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.

  Main effect DOC

  H0: There is no difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.

  Ha: There is a difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.

5. PLANNING (detailed)

Variables Selection:

- **Independent Variables**
  - **Reading techniques**: we have two alternatives: the PBR technique and an usual technique like Checklist.
  - **Perspectives**: Within PBR, a subject uses a technique based on one of the review perspectives. For this experiment we used the three perspectives previously described: Designer, Tester and User.
  - **Requirements documents** (Problem Domain)
  - **Subjects Experience**

- **Dependent Variables**
  - **Effectiveness** in defect detection
  - **Efficiency** to apply the techniques

PROJETO EXPERIMENTAL

- Define o formato experimental, os fatores a serem Investigados e a organização dos tratamentos.
- Define ainda os critérios de seleção de participantes, critérios de agrupamento de participantes e as técnicas de amostragem a serem utilizadas

Experimento com um Fator e Projeto Aleatório Simples

- Divida aleatoriamente as alternativas do Fator sobre as unidades experimentais.
- Exemplo: queremos aplicar a Técnica A e Técnica X com 100 programadores:
  - Aloque aleatoriamente a técnica A a 50 programadores e a técnica B a 50 programadores a Técnica X

Experimento com um Fator e comparação em pares

- Aplique as alternativas do Fator a cada unidade experimental
- Exemplo: queremos aplicar a Técnica A e a Técnica X com 100 programadores:
  - Cada programador aplica a técnica A e a técnica X em uma porção diferente do objeto experimental
  - A ordem de aplicação das técnicas é escolhida aleatoriamente para cada programador
    - 50 aplicam a técnica A primeiro e 50 aplicam a técnica X primeiro

T1-A	T1-X
T2-A	T2-X

Experimento com um Fator e uma variável indesejável

- Use um projeto em blocos em uma matriz com o número de alternativas do fator vezes o número de alternativas da variável
- Este tipo de projeto é dito ser completo
- Exemplo: quer-se aplicar a Técnica A e a Técnica X com 100 programadores, mas o tamanho do objeto usado na especificação pode ser T1 e T2.
- Rode quatro tratamentos com os programadores atribuídos aos pares A-T1; A-T2; X-T1; X-T2

A-T1	A-T2
X-T1	X-T2

SELEÇÃO DE PARTICIPANTES

- O Formato Experimental ajuda a definir a seleção de participantes do experimento
- Deve-se definir
  - o número de participantes do experimento
  - o agrupamento dos participantes
  - As técnicas de amostragem a serem utilizadas
- Em engenharia de software a seleção é frequentemente feita por conveniência, i.e., não aleatória, mas baseada em um grupo de participantes disponíveis
- Isto caracteriza quase-experimentos

Exemplo
Projeto de Experimento e Seleção de Participantes

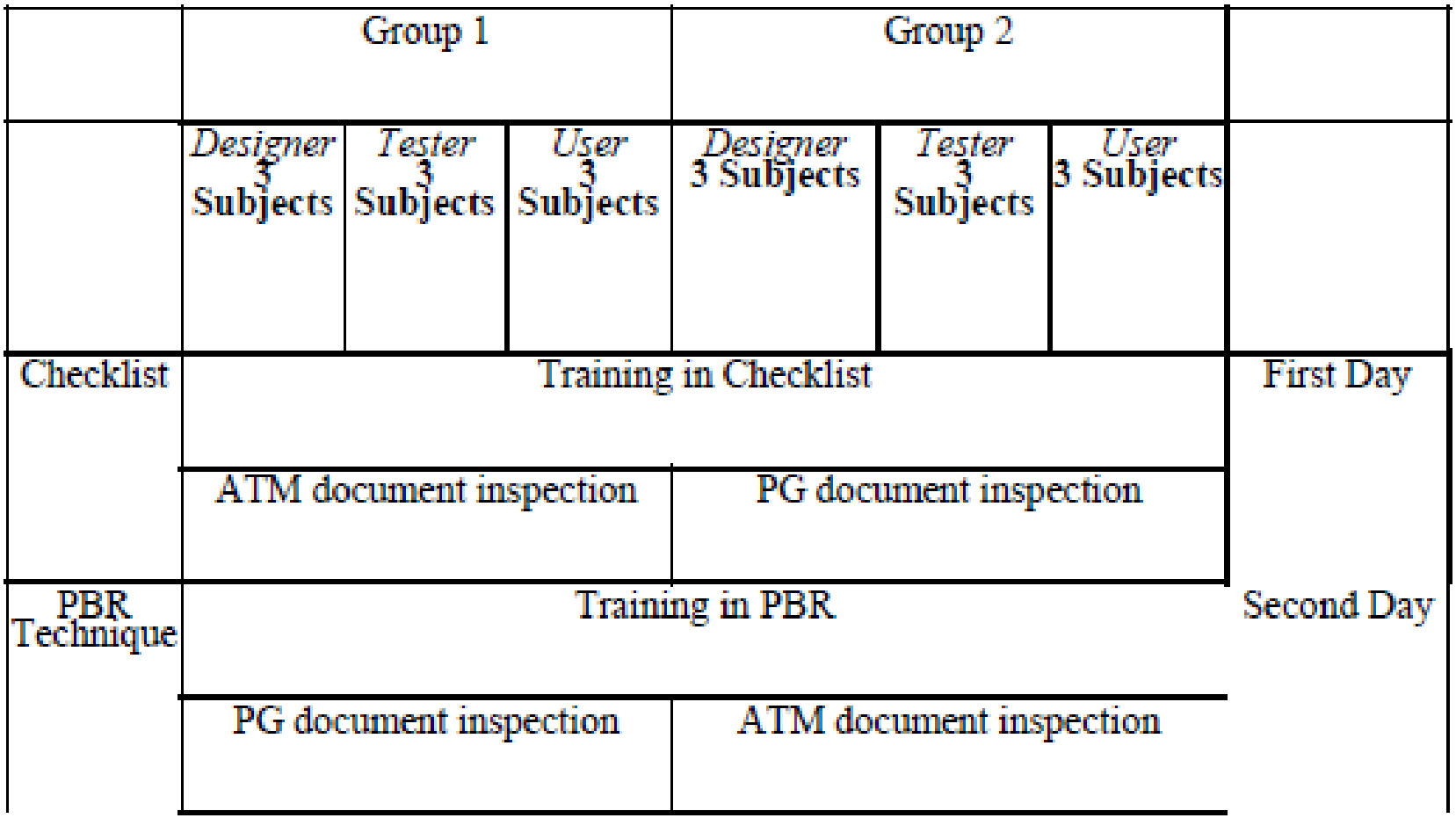
5. PLANNING (detailed)

Experiment Design:

- follows the principles of blocking and balancing
- two factors and two treatments for each factor
- 2*2 factorial design

- **Factor A:** Technique
- **Treatments:** PBR and Checklist

- **Factor B:** Requirements Document
- **Treatments:** ATM and PG documents


Group 1
Group 2
Designer 3 Subjects
Tester 3 Subjects
User 3 Subjects
Designer 3 Subjects
Tester 3 Subjects
User 3 Subjects
Checklist
Training in Checklist
First Day
ATM document inspection
PG document inspection
PBR Technique
Training in PBR
Second Day
PG document inspection
ATM document inspection

5. PLANNING (detailed)

Subjects Selection:

- 18 undergraduate students with slightly more than one year of classroom experience on average, from the Software Engineering course at University of São Paulo at São Carlos, randomly divided into two groups of nine

INSTRUMENTAÇÃO

- Um dos passos mais trabalhosos de um experimento é o desenvolvimento de toda a instrumentação necessária à sua realização
- Isso envolve:
  - Mecanismos de coleta de dados (formulários e questionários em papel ou automatizados)
    - *Documento de consentimento*
    - *Formulários de caracterização de participantes*
    - *Formulários de coleta de dados*
    - *Formulários de feedback*
- Ferramentas e artefatos necessários à execução do experimento

5. PLANNING (detailed)

Instrumentation:

- Artifacts:
  - Consent Form
  - Analyst Survey
  - List of Participants
  - Training Material
  - Objects: ATM and PG Requirement Documents
  - Defect Form
  - Defect List
  - Feedback Questionnaire

Formulário de Participação

Project title and purpose
The experiment is commonly called "The Reading Experiment" and is intended to compare the effectiveness, in terms of defect detection, of various techniques for reading requirements specifications. For this particular experiment, a technique called "Perspective-based Reading" will be compared to the standard reading technique currently applied in the university and industry.

Statement of age
I state that I am over 18 years of age and wish to participate in an experiment conducted by Dr. José Carlos Maldonado at the Universidade de São Paulo and Dr. Sandra Camargo Pinto Ferraz Fabbri at the Universidade Federal de São Carlos, under the scope of the **NSF-CNPq Readers Project.**

Procedures
The experiment involves two sessions, each of about 6 hours. The first session is held on ___/___/____ and the second is held on ___/___/____. Each session will include various presentations, tutorials, training, and sample documents that are to be read and reviewed.

Confidentiality
All information collected in the experiment is confidential, and my name will not be identified at any time.

Survey do Analista

Reviewer Background Questionnaire – E1

Reviewer ID: _____ **Date:** ___/___/___

This form asks you a few questions about your background and experience.

General Questions

1. How many years/months of experience have you had in each function?

Time	Years	Months
Manager		
Developer		
Tester		
Analyst		
Other_____		

2. How comfortable are you with reading or reviewing requirements documents? [Please mark with an x in the scale]

Value	0	1	2	3
Comfort level	Not at all	Low	Moderate	High

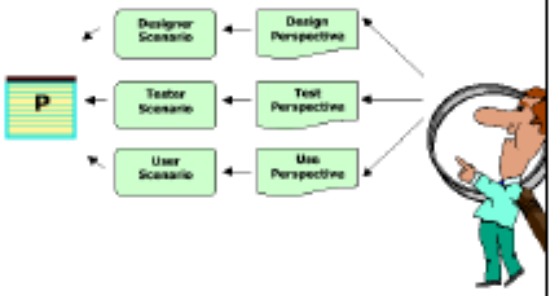
Lista de Participantes

ID #	Name (Print)	Signature
ID 1		
ID 2		
ID 3		
ID 4		
ID 5		
ID 6		
ID 7		
ID 8		

Material de Treinamento

Perspective Based Reading

(Basili et al., 96)



Test-based Reading

For each Requirement/functional specification, generate a test or set tests that allow you to ensure that an implementation of the system satisfies the requirement/functional specification. Use your standard test approach and technique, and incorporate test criteria in the test suite. In doing so, ask yourself the following questions for each test.

1. Do you have all the information necessary to identify the item being tested and the test criteria? Can you generate a reasonable test case for each item based upon the criteria?
Page 11: Functional requirement 10: What it means to 'update' the different files never is specified.

2. Can you be sure that the tests generated will yield the correct values in the correct units?
Page 9: Functional requirement 3: "information" is not specified. There are

equirement that the implementer might
al/functional specification is defined?

Fault Taxonomy

Omission

- **Missing Functionality**: Information describing internal operational behavior of the system has b from the SRS.
- **Missing Performance**: Information describing performance specifications has either been described in a way that is unacceptable for testing.
- **Missing Interface**: Information describing how th system will interface and communicate with othe

DEFECT REPORT – FORM E5

Reviewer ID: _____ Date: ___/___/___

Page 1 of 1

Document name: **ABC Video System**

Defects			
# Defect	**Page**	**Class**	**Description**
1	11	O	FR10: What it means to "update" the different files is never specified.
2	9	O	FR8: "Information" is not specified. There are several interpretations for this!

Formulário de Defeitos

DEFECT REPORT – FORM E5

Reviewer ID: _____ Date: ___/___/___

Page **1** of **1**

Document name: **ABC Video System**

Defects			
# Defect	**Page**	**Class**	**Description**
1	11	O	FR10: What it means to "update" the different files is never specified.
2	10	O	FR6: It is not specified that the account number must be entered before the number of a tape can be entered.
3	9	O	FR3: What is a "transaction record"?

Características dos documentos de requisitos

- Documento ATM (caixa eletrônico):
  - 17 páginas
  - 39 requisitos
  - 26 funcionais
  - 13 não funcionais
  - 30 defeitos semeados
- Documento PG (garagem automatizada):
  - 17 páginas
  - 37 requisitos
  - 21 funcionais
  - 16 não funcionais
  - 28 defeitos semeados

Lista de defeitos

1. Defect list for the generic documents

The defects lists are written according to the format defined by the tables below. Please note that page number refers to the page numbering of the separately printed documents, i.e. not as they appear in this manual. If the defect is pertinent to one requirement, the number of the requirements as it appears in the generic document should also be indicated here.

The following classification is used for analysis:

A mbiguous information
E – Extraneous information
II – Inconsistent Information
IF – Incorrect Fact
MD – Miscellaneous Defect
MI – Missing Information

Noting the type of effort required t

O – Error of Omission
It is necessary to add information t

C – Error of Commission
It is necessary to edit/delete inform

1.1 ABC Video System

Def. #	Page	Req. #	O/C	Class	Description
1	8	FR1	O	MI	Only clerk functions are listed. How manager functions are accessed is not clear.
2	9	FR2	O	MI	Definition omitted: "current status".
3	9	FR3	O	MI	Information required for the transaction record is not specified

Questionário de feedback

Feedback Questionnaire – Form E7

Reviewer ID: _____ Date: ___/___/___

This form asks you a few questions about the experiment itself after it is conducted.

General Questions

1. Please assess the explanations before and throughout the training session. [Please mark with an x in the scale]

	0	1	2
Training, Explanation	Not Enough	Enough	Too Much

2. Did you need more time than allocated for reviewing the documents?

Yes	
No	

Análise de Riscos à Validade do Experimento

- Todo experimento tem riscos à validade de seus resultados. Quão válidos são eles?
- O projetista do experimento tem a obrigação de identificar e discutir estes riscos.
- As principais categorias de riscos são:
  - Riscos à conclusão
  - Riscos à validade interna
  - Riscos à validade externa
  - Riscos à validade da construção

Validade de conclusão

- Trata-se da habilidade de chegar a uma conclusão correta a respeito dos relacionamentos entre o tratamento e o resultado do experimento.
- É preciso considerar:
  - O teste estatístico a utilizar
  - Escola do tamanho do conjunto de participantes
  - Confiabilidade das medidas
  - Confiabilidade da implementação dos tratamentos.

Validade interna

- É a verdade aproximada sobre relações de causa e efeito (relações causais).
- Validade interna só é relevante em estudos que tentam estabelecer uma relação causal.
- A questão-chave na validade interna é se as mudanças observadas podem ser atribuídas ao tratamento ou à intervenção e **não** a outras causas possíveis (às vezes chamadas de explicações alternativas)

Validade externa

- Trata da validade da generalização de inferências (causais) em estudos científicos, geralmente baseadas em estudos experimentais.
- Problemas comuns com estudos envolvendo seres humanos: baixa amostra obtidas de uma única localização geográfica ou com características idiossincráticas (ex. Voluntários)

Validade de Construção

- Considera os relacionamentos entre a teoria e a observação, isto é, se o tratamento reflete bem a causa e o resultado reflete bem o efeito.
- Os problemas podem surgir por falhas do experimentador ou dos participantes:
  - Os participantes podem basear seu comportamento em suposições sobre a hipótese.
  - O ser humano geralmente tenta parecer melhor do que é quando está sendo avaliado.
  - Os pesquisadores podem projetar o experimento pensando nos resultados que esperam (viés).

Ordem de importância das validades

- Verificação de uma teoria:
  - Interna, Construção, Conclusão, e Externa
- Experimentos aplicados (em engenharia de software, por exemplo)
  - Interna, externa, construção e conclusão

Exemplos
de Análise de Riscos à Validade

5. PLANNING (detailed)

Results Validity:

- **Internal Validity**
  - Language: The class lecture notes, assignment instructions, techniques and artifacts are written in English, so the lack of proficiency in English can affect the results of the study.
  - Learning: PBR is a more procedurally defined technique than a checklist. If subjects are trained in the use of the more procedural PBR techniques prior to using the less-procedural checklist, there is a danger that they will perform the tasks on the checklist in a more ordered fashion.
  - Conformance to the Original study: There are some changes made to the experimental procedures by the replicators before running the study. There are two main issues that must be considered for this threat:
    - The replicators made some adjustments to the training time but are keeping it equal for both techniques and they provide different levels of detail and require different levels of detail and background.
    - The techniques will be applied just after training, without giving the subjects time to mature and assimilate the underlying concepts.
  - Process conformance of the subjects: We do not have mechanism to observe the subjects while they are working nor will collect any intermediate artifacts. Thus, we can not be certain that the subject will follow the technique.

5. PLANNING (detailed)

Results Validity:

- **External Validity**
  - This study will run in the classroom at a university and the subjects are not as experienced as industrial professionals. The subjects are not experienced in PBR perspectives. Thus, the conclusions of this study may not be directly transferable to industrial inspectors.
  - The small number of subjects who will participate in the replication. It is possible that any result of the study be a function of this small sample size

Planejamento Detalhado e sua Avaliação

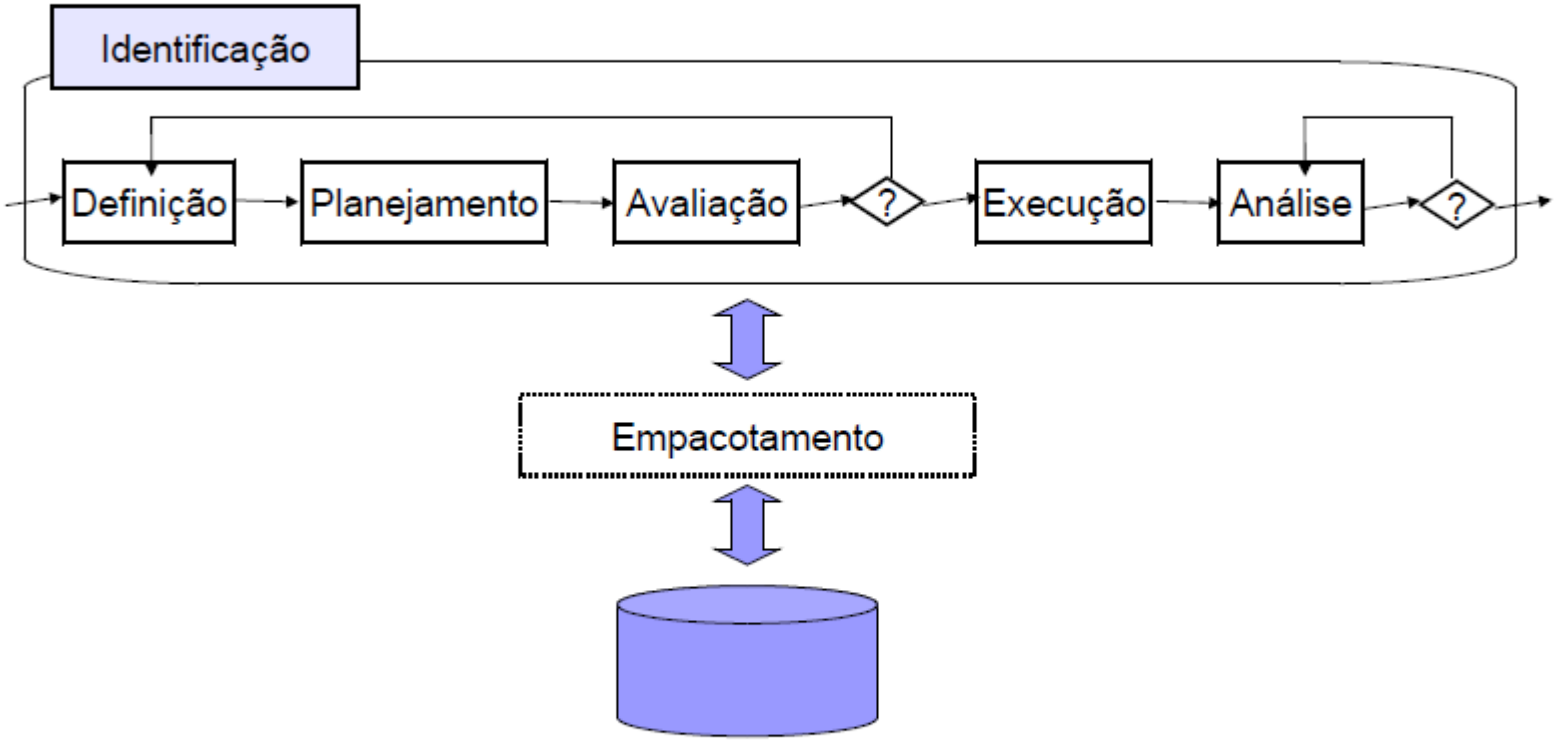
- Treinamento
  - Aplicadores, participantes (processos, artefatos e técnicas)
- Definição do Processo Experimental
- Procedimentos de Execução
  - Objetivos
  - Participantes
  - Processo Experimental
  - Artefatos Usados
  - Resultados Esperados e Artefatos Resultantes (lições aprendidas e sugestões de modificação)
- Custos experimentais
  - Tempo por tipo de participante, custos de aplicação, custos de análise, custos de empacotamento e divulgação dos resultados

Resumo

- Definição das Hipóteses
- Definição de variáveis independentes e dependentes
- Projeto Experimental
  - Objetos, medidas, instruções, técnicas, formato experimental e tratamentos.
- Critérios de seleção de participantes, critérios de agrupamento de participantes, técnicas de amostragem a serem utilizadas
- Instrumentação e Recursos necessários: software, hardware, questionários, formulários
- Mecanismos de Análise
- Análise de Validade
  - Interna, externa, construção, instrumentação e conclusão.

Avaliação do Planejamento Detalhado

Processo de Experimentação



Avaliação

- A avaliação tem o objetivo de verificar se o projeto experimental é consistente e pode realmente ser executado
- Esta fase deve avaliar o projeto experimental e seus artefatos
- Ela deve ser baseada em opinião de especialistas e estudos pilotos
- Os pontos mais críticos a serem avaliados em um experimento são o projeto experimental e o processo subjacente:
  - Tempo destinado a cada atividade
  - Treinamento nas técnicas e artefatos a serem utilizados

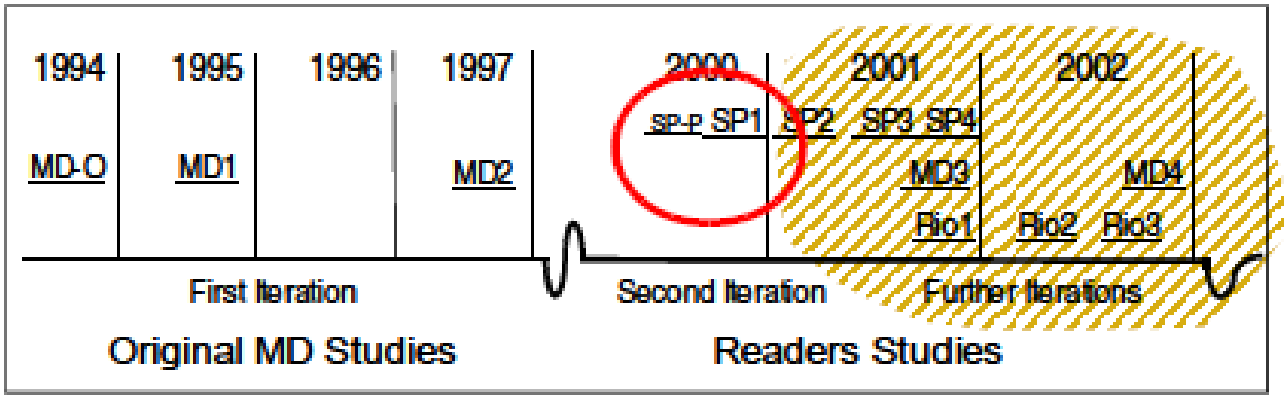
Avaliação do Planejamento

- Consulte especialistas
  - Consulte especialistas no domínio para avaliar seus artefatos experimentais
  - Consulte especialistas em ES para avaliar objetivos e projeto experimental
  - Consulte especialistas em ESE e estatísticos para avaliar formato experimental e planos de coleta e análise de dados
- Se possível, rode um estudo piloto
  - Rode uma versão simplificada do experimento para avaliar seu projeto e artefatos.
- Redesenhe o experimento e melhore os artefatos conforme necessário

ESTUDOS-PILOTO

- Versões simplificadas do estudo experimental
- Simplifica o estudo em dimensões como
  - Número de participantes
  - Volume de tarefas a serem executadas
  - Tamanho dos artefatos utilizados
- Resultados do estudo piloto não devem ser usados na análise de dados e teste de hipóteses
- O estudo piloto deve ser usado para avaliar a razoabilidade do projeto experimental, seus artefatos e processos.

Exemplo:
Avaliação e Melhoria do Projeto Experimental



PBR Studies Timeline

- Um estudo piloto foi realizado para entender as tarefas associadas ao estudo experimental PBR
- Este estudo ajudou a:
  - Determinar os tempos associados às tarefas listadas no projeto
  - Avaliar formulários e questionários
  - Refinar e avaliar o material de treinamento

6. Training

- Definition and procedure:
  - The training will be done in two 2-hour sessions using another artifact, ABC Video Store.
  - The sessions will consist of 30 minutes of theoretical presentation and 90 minutes of practice with the techniques.
  - At the end of the training, the researchers will give the subjects feedback on their performance and the full list of defects for the ABC Video Store document.
  - This feedback will allow the subjects to see the types of defects that they would not uncover and use this information in future applications of the technique.

- Instructor: Emerson Sillas Dória

- Participants: 18 undergraduate students

- Artifacts:
  - ABC Video Store Requirements Document
  - Defect Form
  - Defect List

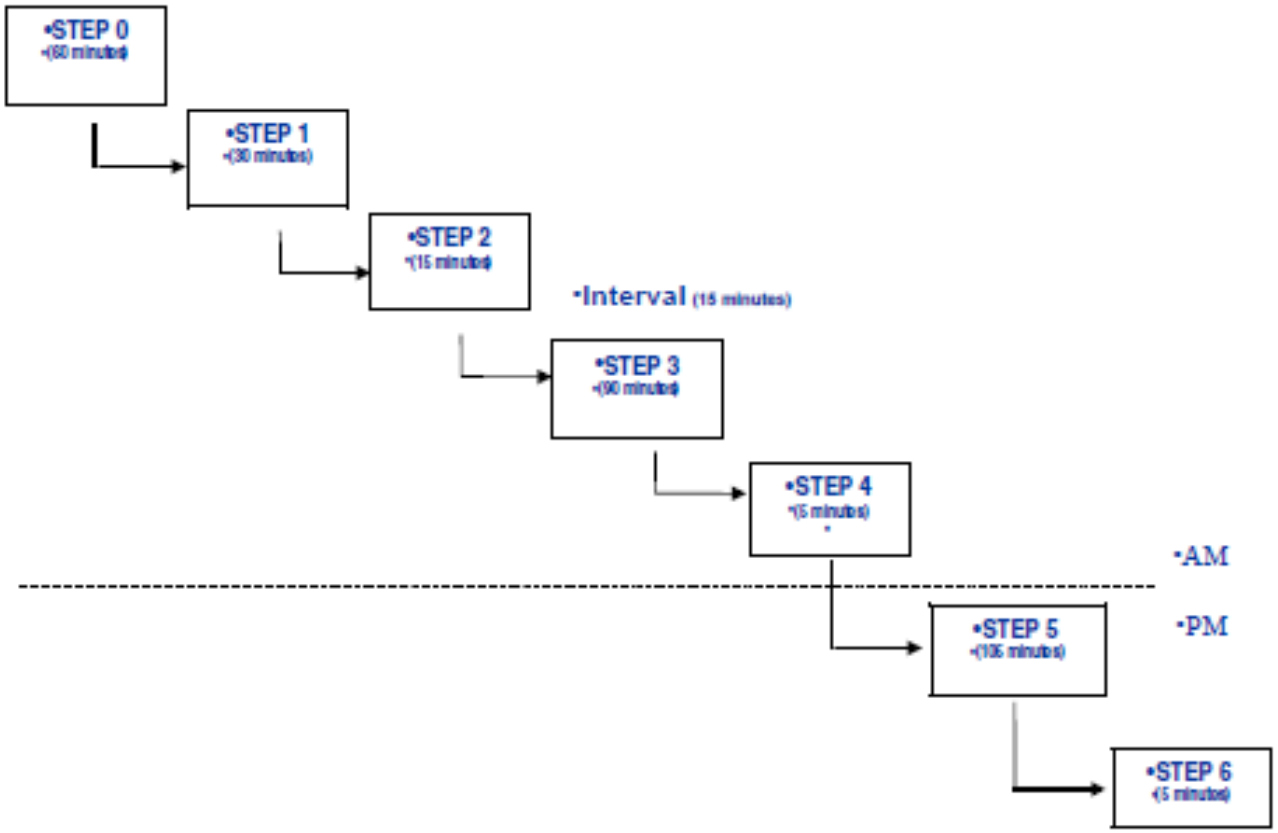
7. EXECUTION PROCEDURE

- Definition:
  - The subjects will apply the Checklist and PBR techniques in sessions of 1 hour and 45 minutes each.
  - First Day:
    - after receiving training in the Checklist method, the subjects from Group 1 will review the ATM document and subjects from Group 2 will review the PG document.
    - Each subject will be assigned to one of three subgroups for PBR (one for each perspective).
  - Second Day:
    - after receiving training in the assigned PBR perspective, the subjects reviewed the other requirements document.
  - The subjects will perform the inspections in a classroom while the experimenters will be present.
  - During the inspection of the documents the subjects should record any defects they find along with a classification for the defect.

- Instructor: Emerson Sillas Dória

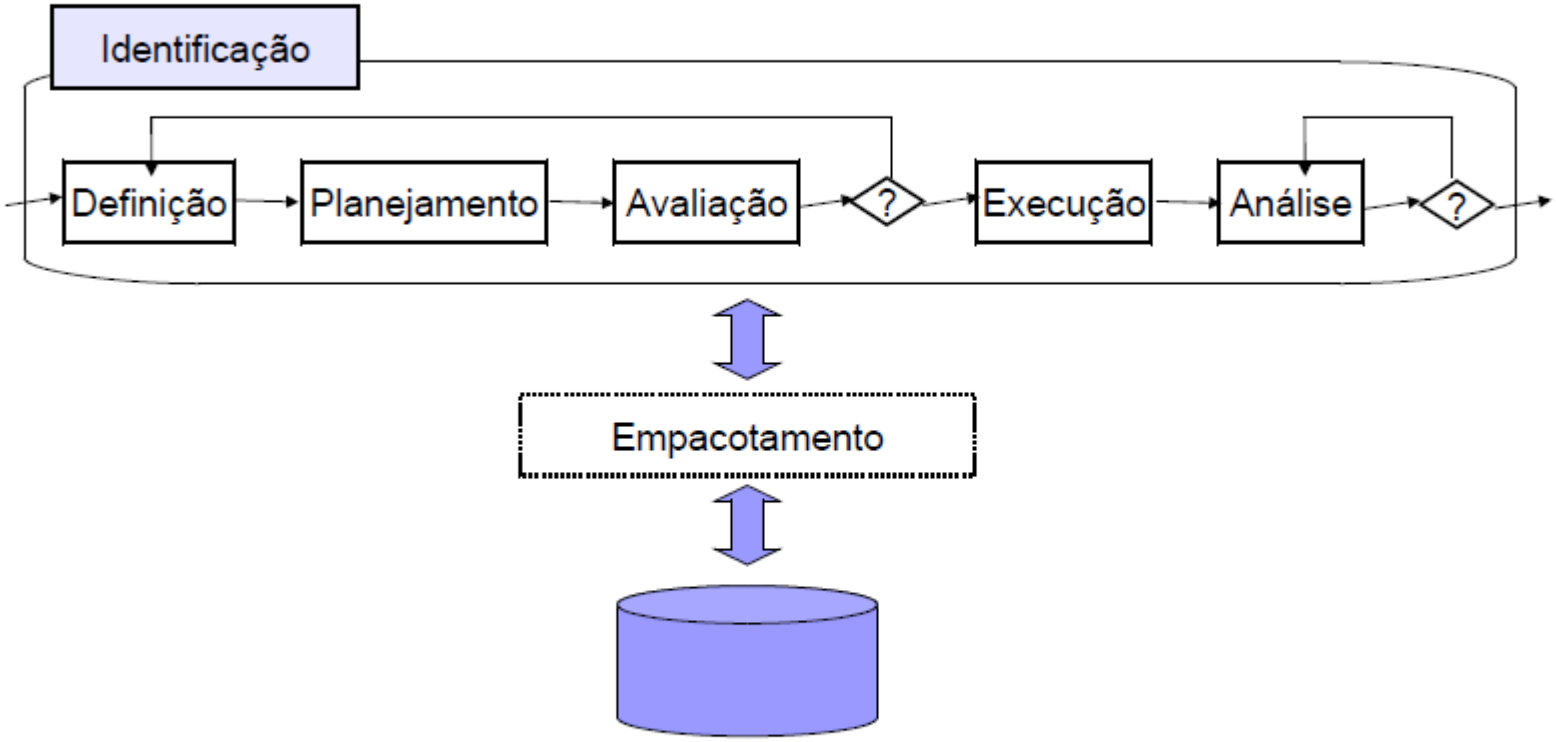
7. EXECUTION PROCEDURE

- First Day Detailed Definition: Example



EXECUÇÃO

Processo de Experimentação



Execução

- Nesta parte os experimentadores “se encontram” efetivamente com os sujeitos.
- Os tratamentos são aplicados aos sujeitos
  - Siga os passos especificados no plano e não se desvie dele;
  - Desvios devem cancelar o experimento
  - Não se deve “consertar” um experimento durante sua execução (ex. corrigir artefatos).

Execução

- Ocorre em três passos, tendo como entrada o projeto do experimento e como saída os dados do experimento:
  - Preparação
    - Os sujeitos são escolhidos, os formulários são preparados, etc.
  - Operação
    - Os sujeitos executam suas tarefas de acordo com os diferentes tratamentos e os dados são coletados
  - Validação dos dados

PREPARAÇÃO

- Obter o comprometimento dos participantes
  - Obter consentimento;
  - Resultados sensíveis para os participantes?
  - Estímulo/pagamento;
  - Enganar.
- Preparar a instrumentação

Operação

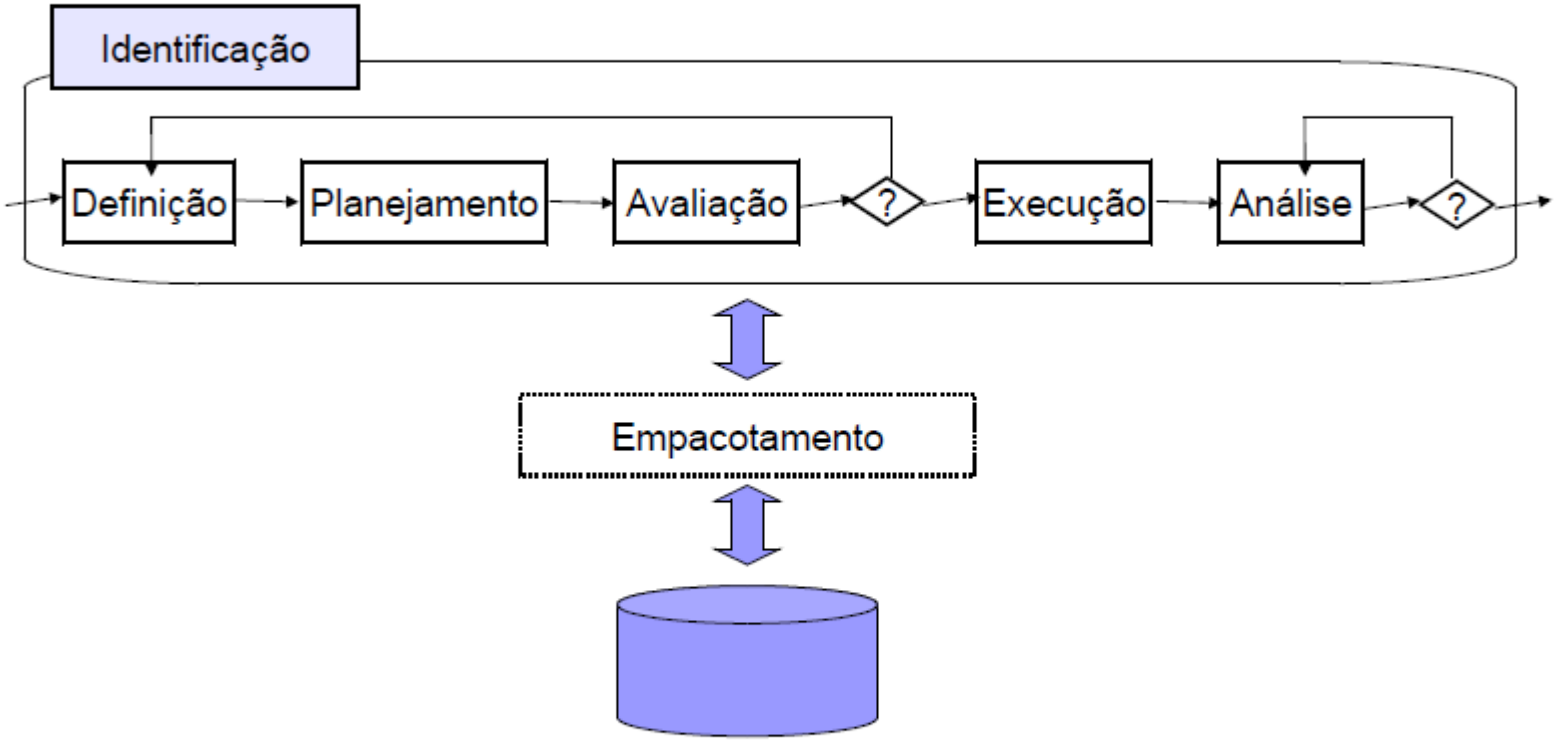
- Há várias formas de operacionalizar o experimento: Ex.
  - Em uma reunião
  - Ao longo de uma disciplina
  - Quando o experimento é muito longo, o experimentador pode não conseguir participar de todos os detalhes do experimento e da coleta de dados.
- Os dados podem ser coletados de diferentes formas
- Experimentos podem ser realizados como parte de um projeto de software regular.
  - O experimento não deve afetar o projeto mais do que o necessário (se ambiente do projeto mudar, pode afetar os resultados do projeto)

Validação dos Dados

- Checar que os dados são razoáveis e que foram coletados corretamente.
  - Os participantes entenderam o formulários?
  - Os formulários foram corretamente preenchidos?
  - Ocorreram erros na aplicação do tratamento ou em sua ordem?
- Um seminário para apresentar e discutir os formulários pode ser feito.

Análise

Processo de Experimentação



Análise dos Resultados

- Se possível, entreviste os participantes para obter feedback:
  - Sobre os artefatos
  - Sobre o processo experimental
  - Para capturar sua impressão sobre os resultados
- Revise os dados coletados para verificar se eles são úteis e válidos
- Organize os dados em conjuntos para análise de validade, exploração e teste das hipóteses
- Analise os dados com base em princípios estatísticos válidos
- Verifique se as hipóteses são aceitas ou rejeitadas
- O processo de análise pode ser iterativo.

Análise dos resultados

- Análise descritiva dos dados
- Testes de hipóteses

Análise Descritiva

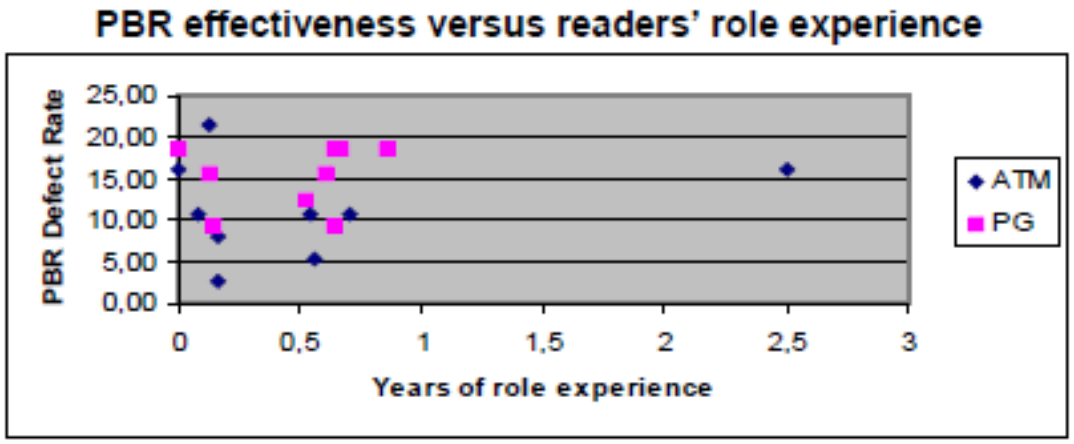
- Estatística Descritiva
  - Medidas de Tendência Central (média, mediana, moda)
  - Medidas de dispersão (desvio padrão, variância)
  - Correlações (Pearson, Spearman)
- Análise Gráfica
  - Diagramas de dispersões
  - Histogramas e Gráficos de Pizza
  - Box Plots

Metas da Análise Descritiva

- Identificar tendências centrais de variáveis e seus tratamentos
- Identificar o grau de dispersão
- Identificar pontos fora da curva (outliers)
- Identificar Correlações

Exemplo de Análise Descritiva

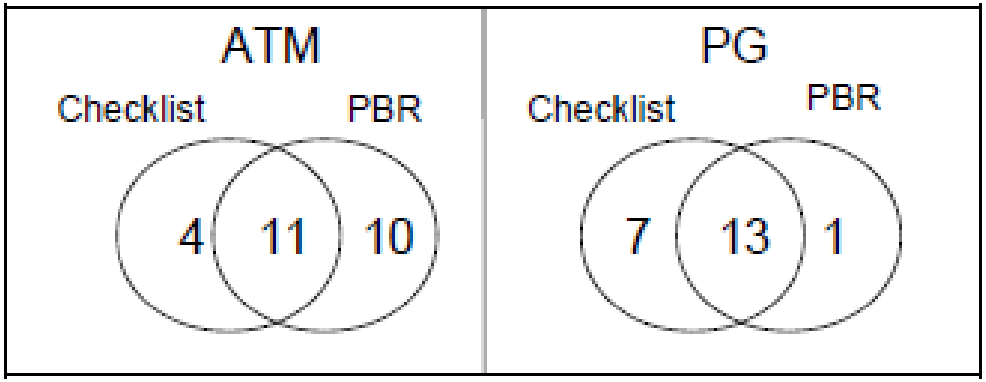
- **O3') Does the reviewer's experience affect his or her effectiveness?**



Analysis

- We used a questionnaire to measure the subject's experience in their assigned perspective. The relationship between experience and effectiveness is weak
- Reviewers with more experience do not perform better than reviewers with less experience
- This conclusion is supported by the results of the Spearman's and Pearson's correlation tests that showed numbers smaller than 14%, far from indicating a high degree of correlation

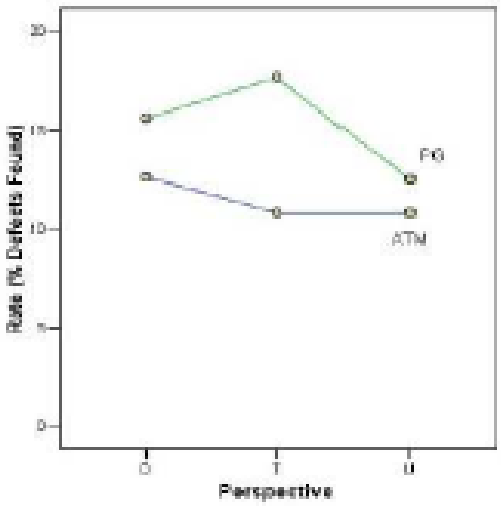
- **R1) Do individual reviewers using PBR and Checklist find different defects?**

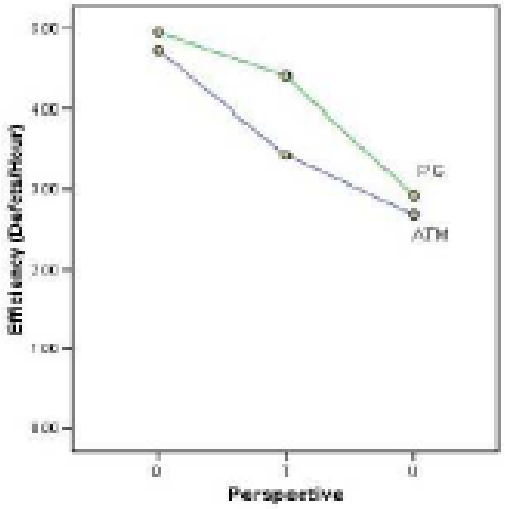


Analysis

- **ATM**: the two techniques appear to be complementary in that users of each technique found defects that were not found by the other technique
- **PG**: the techniques do not appear to be complementary, because the PBR users only found 1 defect not found by the checklist users

- **R2) Do the PBR perspectives have the same effectiveness and efficiency?**



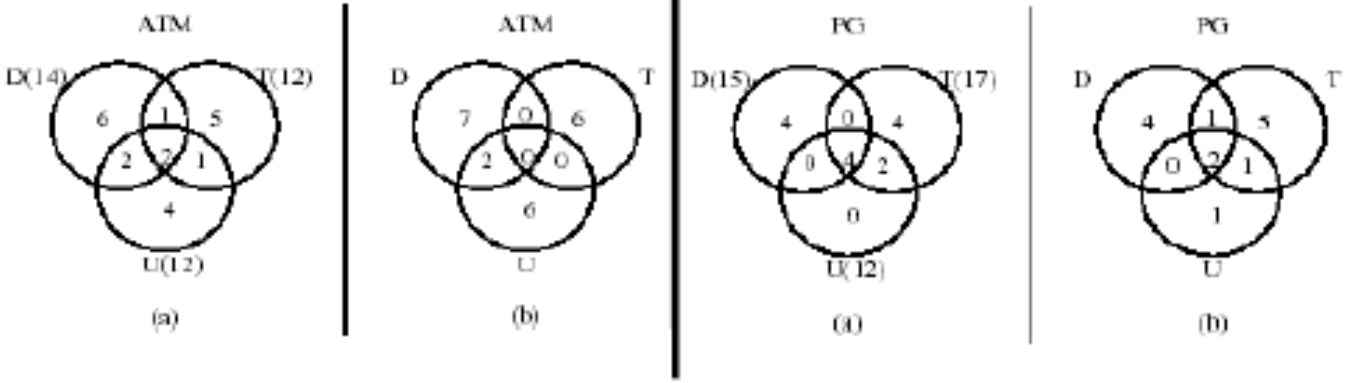


Rate and Efficiency by Perspective

Analysis

- Each point represents the mean of the 3 reviewers composing the group.
- ATM: Designer were the most effective and efficient
- PG: Tester were the most effective and Designer were the most efficient
- The perspectives had no significant effect on either effectiveness (p=.654) or efficiency (p=.182)

R3) Do the PBR perspectives find different defects?



Analysis

- **ATM:**
  - each perspective identified unique defects with little overlap
  - the three perspectives were more likely to find different defects
  - The perspectives identified a similar number of occurrences overall
- **PG:**
  - the Designer and Tester perspectives appear to be complementary, but the User perspective does not provide much added benefit
  - The perspectives identified a similar number of occurrences

Estatísticas relevantes por tipo de escala

Tipo escala	Med. Tend. Central	Dispersão	Dependência
Nomimal	Moda	Frequência	
Ordinal	Média, percentil	Intervalo de variação	Coefs. de correção de Spearman e de Kendall
Intervalar	Média	Desvi padrão, variância, range	Coef. De correção de Pearson
Quociente	Média geométrica	Coeficiente de variação	

Medidas de Tendência Central

- Média

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

  - Ex. para o conjunto de dados (1,1,2, 4), m=2
- Mediana
  - Representa o valor médio do conjunto de dados
  - Se n é impar, pega-se a amostra média do conjunto ordenado
  - Se n é par, pode-se calcular a média das duas amostras centrais.
  - Ex. a mediana de (1,1,2,4) é 1,5

Medidas de Dispersão

- A variância é definida como: $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$
- O desvio padrão s é definido

  como a raiz quadrada da variância.
- Ele é geralmente preferido em relação à variância porque tem a mesma unidade de medida que os valores da amostra.
- O range de um conjunto de dados é a distância entre os valores máximos e mínimos do conjunto de dados.
  - Range = Xmax - Xmin

Medidas Dispersão (Cont.)

- Intervalo de variação (Xmin,Xmax)
- Coeficiente de variação: é expresso como uma porcentagem da média: 100 . s/x (x traço)
- Uma visão geral da dispersão é dada pelo frequência de cada valor.
- A frequência relativa é calculada dividindo-se cad frequência pelo número total da amostra.
  - Ex (1,1,1,2,2,3,4,4,4,5,6,6,7), com tamanho 13. A frequência relativa do valor 1 é 23%, do 2 é 15% etc.

Regressão linear

- Se o conjunto de dados contém variáveis estocásticas X e Y em pares (Xi,Yi) e suspeitamos que há uma funçao y= f(x) que relaciona os pares x e y.
- Se y=c1+c2.x então dizemos que a regressão é linear.

Regressão Linear

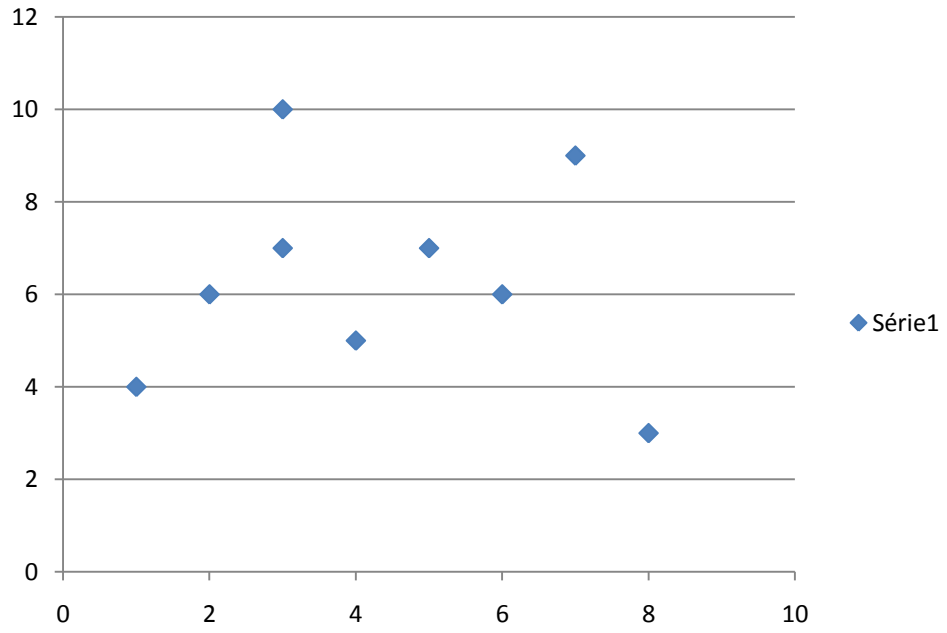
$$r = \frac{c_{xy}}{s_x \cdot s_y} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{\left(n \cdot \sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

- r= Coeficiente de correlação (de Pearson)
- Se r igual a zero não há correlação
- O valor de r fica entre -1 e +1
- Pode haver uma correlação não linear mesmo se r=0.

Regressão Linear

- Se a escalara é ordinal ou se o conjunto de dados não está distribuído normalmente, o coeficiente de correlação de Spearman pode ser usado (Rs)
- O cálculo é feito da mesma forma mas os ranks (isto é, os números ordinais em que a amostra é ordenada) é que são usados, ao invés dos valores da amostra

Visualização gráfica (Dispersão)



Visualização gráfica (Histrograma)

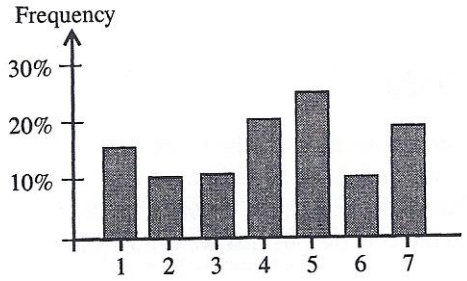
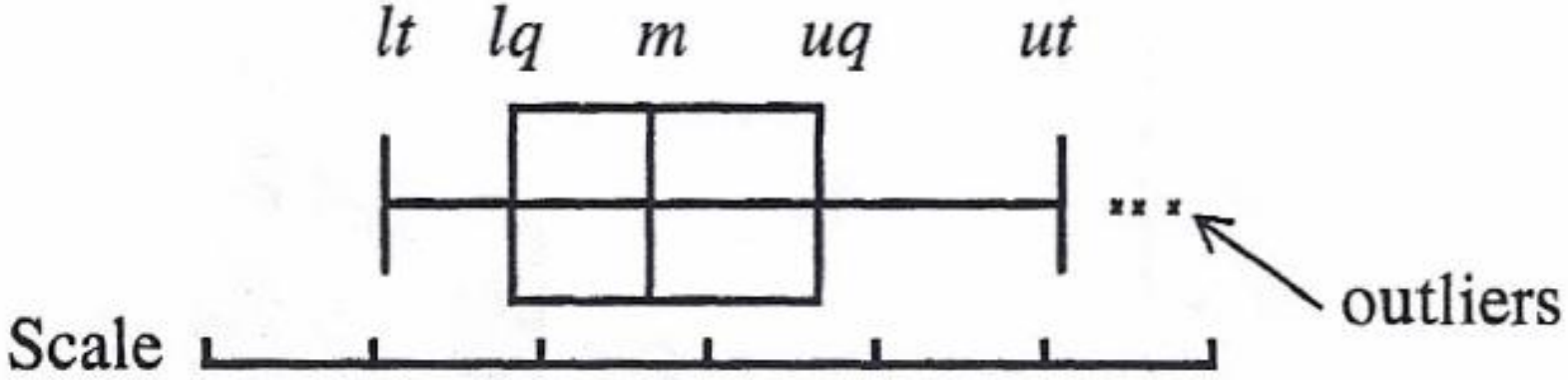


Figure 19. *A histogram.*

Box Plot



m = mediana
lq= 25%
uq = 75%
O tamanho da caixa é d = lq – uq
ut= uq+1.5d
lt= lq – 1.5d

Redução do conjunto de dados

- Um critério é baseado no resultado do experimento. Ex. sujeitos que não participaram seriamente do experimento.
- Os erros podem ser sistemáticos ou ocorrer como outliers.
- Ao encontrar outliers, é necessário decidir se eles serão retirados ou não. Ex. surgiram porque pessoas inexperientes participaram do experimento.
- Olhar também dados redundantes.

Testes de Hipóteses

- Hipóteses avaliadas por testes estatísticos definidos por pesquisadores da estatística inferencial
- Normalmente são definidas duas hipóteses
  - **Hipótese nula (H0): indica que as diferenças observadas no estudo** são coincidentais, ou seja, é a hipótese que o analista deseja rejeitar com a maior significância possível
  - **Hipótese alternativa (H1): é a hipótese inversa à hipótese nula, que** será aceita caso a hipótese nula seja rejeitada
- Os testes estatísticos verificam se é possível rejeitar a hipótese nula, de acordo com um conjunto de dados observados e suas propriedades estatísticas

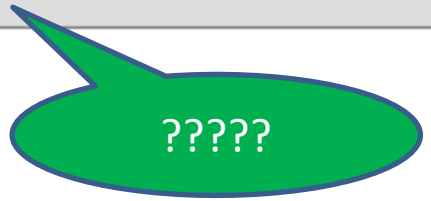
Testes de Hipótese

- Os testes comparam médias entre grupos de participantes realizando tratamentos diferentes

"Utilizando a técnica XYZ, os desenvolvedores concluem a atividade de projeto em menos tempo do que utilizando a técnica ABC"

Hipótese Nula: $\mu\ (Tempo_{XYZ}) = \mu\ (Tempo_{ABC})$

Hipótese Alternativa: $\mu\ (Tempo_{XYZ}) \neq \mu\ (Tempo_{ABC})$



Teste Estatístico

- Calculados fundamentalmente a partir de uma função de teste que considera três valores:
  - Diferença entre os valores “médios” das estatísticas para os tratamentos
  - “Dispersão” dos valores da estatística
  - Número de amostras
- A função de teste, $F(m,\sigma,N)$, depende do:
  - tipo de distribuição dos dados, e.x., normalidade e homocedasticidade.
  - Número de fatores e tratamentos

Homogeneidade da variância

Exemplo

- Quer-se testar a Hipótese “homens são mais altos que mulheres”
  - Determina-se uma amostra da população utilizando um fator e dois tratamentos
  - A certeza depende de:
    - Número de pessoas amostradas
    - Diferença entre a altura média nos tratamentos
    - Dispersão da altura nos tratamentos

Tipos de Erros

- A verificação das hipóteses sempre lida com o risco de um erro de análise acontecer
  - O erro do tipo I (a) acontece quando o teste estatístico indica um relacionamento entre causa e efeito e o relacionamento real não existe
  - O erro do tipo II (b) acontece quando o teste estatístico não indica o relacionamento entre causa e efeito, mas existe este relacionamento

α = P (erro-tipo-I) = P (H_{NULA} é rejeitada | H_{NULA} é verdadeira)

β = P (erro-tipo-II) = P (H_{NULA} não é rejeitada | H_{NULA} é falsa)

Nível de Significância

- Indica a probabilidade de se cometer um erro do tipo I:
  - Os níveis de significância (α) mais comumente utilizados são 10%, 5%, 1% e 0.1%
  - Chama-se de *p-value o menor nível de significância com que se* pode rejeitar a hipótese nula
  - Dizemos que há significância estatística quando o *p-value é menor* que o nível de significância adotado

Procedimento para o Teste de Hipótese

- Fixar o nível de significância do teste
- Obter uma estatística (estimador do parâmetro que se está testando) que tenha distribuição conhecida sob H0
- A estatística de teste e o nível de significância constroem a região crítica pela qual o teste passa
- Usando as informações amostrais, obter o valor da estatística (estimativa do parâmetro)
- Se valor da estatística pertencer à região crítica, rejeita-se a hipótese nula, aceitando-se a hipótese alternativa
- Caso contrário, não se rejeita a hipótese nula e nada se pode dizer a respeito da hipótese alternativa

Teste de Hipótese na Prática

- Na prática:
  - escolhe-se a estatística de teste
  - escolhe-se o valor P (significância)
  - Usa-se uma ferramenta estatística para aplicar o teste e verificar o valor de P
- A escolha do teste depende da determinação do tipo de distribuição dos dados e de quantos fatores e tratamentos vão ser analisados no teste
  - Testes paramétricos: assumem uma distribuição e são mais poderosos
  - Testes não paramétricos: não assumem uma distribuição . Têm uma aplicação mais abrangente

Alguns Tipos de Teste

Projeto	Teste paramétrico	Teste não-paramétrico
Um fator, um tratamento	-	Binomial Chi-2
Um fator, dois tratamentos aleatórios	Teste T Teste F	Mann-Whitney Chi-2
Um fator, dois tratamentos pareados	Teste T pareado	Wilcoxon
Um fator, mais de dois tratamentos	ANOVA	Kruskal-Wallis Chi-2

ANOVA (ANalysis Of VARiance)

- Usado para avaliar experimentos com várias quantidades de projetos.
- Baseado na análise da variabilidade total dos dados e da variabilidade de uma partição de acordo com diferentes componentes.
- Na sua forma mais simples compara a variabilidade devida ao tratamento com a variabilidade devida a erros randômicos.

Forma mais simples: compara se as amostras têm o mesmo valor médio, isto é, o projeto tem um fator e dois tratamentos.

ANOVA, one factor, more than two treatments	
Input	a samples: $x_{11}, x_{12}, \ldots x_{1n_1}$; $x_{21}, x_{22}, \ldots x_{2n_2}$;...; $x_{a1}, x_{a2}, \ldots x_{an_a}$
H_0	$\mu_{x_1} = \mu_{x_2} = \ldots = \mu_{x_a}$, i.e. all expected means are equal
Calculations	Calculate: $SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{x_{..}^2}{N}$ $SS_{Treatment} = \sum_{i=1}^{a} \frac{x_{i.}^2}{n_i} - \frac{x_{..}^2}{N}$ $SS_{Error} = SS_T - SS_{Treatment}$ $MS_{Treatment} = SS_{Treatment}/(a-1)$ $MS_{Error} = SS_{Error}/(N-a)$ $F_0 = MS_{Treatment}/MS_{Error}$ where N is the total number of measurements and a dot index denotes a summation over the dotted index, e.g. $x_{i.} = \sum_j x_{ij}$
Criterion	Reject H_0 if $F_0 > F_{\alpha,a-1,N-a}$. Here, F_{α,f_1,f_2} is the upper α percentage point of the F distribution with f_1 and f_2 degrees of freedom, which is tabulated in, for example, Table A5.1, Table A5.2 and [Montgomery97, Marascuilo88].

Teste t (Student)

- Usado para comparar duas amostras independentes (um fator com dois níveis).
- Pode ser usado com diferentes suposições.

t-test	
Input	Two independent samples: $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_m$.
H_0	$\mu_x = \mu_y$, i.e. the expected mean values are the same.
Calculations	Calculate $t_0 = \dfrac{\bar{x} - \bar{y}}{S_p \sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$, where $S_p = \sqrt{\dfrac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}$, and, S_x^2 and S_y^2 are the individual sample variances.
Criterion	Two sided (H_1: $\mu_x \neq \mu_y$): reject H_0 if $\lvert t_0 \rvert > t_{\alpha/2,\, n+m-2}$. Here, $t_{\alpha,\, f}$ is the upper α percentage point of the t distribution with f degrees of freedom, which is equal to $n+m-2$. The distribution is tabulated in, for example, Table A1 and [Montgomery97, Marascuilo88]. One sided (H_1: $\mu_x > \mu_y$): reject H_0 if $t_0 > t_{\alpha,\, n+m-2}$.

Example of t-test. The defect densities in different programs have been compared in two projects. In one of the projects the result is $x = \{3.42, 2.71, 2.84, 1.85, 3.22, 3.48, 2.68, 4.30, 2.49, 1.54\}$ and in the other project the result is $y = \{3.44, 4.97, 4.76, 4.96, 4.10, 3.05, 4.09, 3.69, 4.21, 4.40, 3.49\}$. The null hypothesis is that the defect density is the same in both projects, and the alternative hypothesis that it is not. Based on the data it can be seen that $n = 10$ and $m = 11$. The mean values are $\bar{x} = 2.853$ and $\bar{y} = 4.1055$.

It can be found that $S_x^2 = 0.6506$, $S_y^2 = 0.4112$, $S_p = 0.7243$ and $t_0 = -3.96$.

The number of degrees of freedom is $f = n+m-2 = 10+11-2 = 19$. In Table A1, it can be seen that $t_{0.025,\,19} = 2.093$. Since $|t_0| > t_{0.025,\,19}$ it is possible to reject the null hypothesis with a two tailed test at the 0.05 level.

Exemplo
Teste de Hipóteses

- **O1') Do PBR teams detect a more defects than Checklist teams?**
  - **H0**: There is no difference in the defect detection rates of teams applying PBR compared to teams applying the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer has only random effects on team scores.
  - **Ha**: The defect detection rates of teams applying PBR are higher compared to teams using the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer decreases the effectiveness of the team.

Analysis:

- Doing a permutation test as done in the original experiment, there were 48620 distinct ways to assign the reviewers into groups of 9.
- The group with no dilution (all PBR reviewers) had the 24769th highest test statistic, corresponding to a p-value of 0.51.
- Therefore, unlike the original study, we cannot reject the hypothesis H0.

O2') Do individual PBR or Checklist reviewers find more defects?

- **Group effect (RT X DOC interaction)**
- **H0**: There is no difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency.
- **Ha**: There is a difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency

- **Main effect RT**
- **H0**: There is no difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.
- **Ha**: There is a difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.

- **Main effect DOC**
- **H0**: There is no difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.
- **Ha**: There is a difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.

Because the experimental groups had the same number of subjects, the ANOVA for balanced design was used. This analysis involved two different factors, or treatments: the reading technique (RT), the requirement document (DOC).

- **O2') Do individual PBR or Checklist reviewers find more defects?**

ANOVA summary table with respect to the individual effectiveness

Independent Variables	Effectiveness (average percentage MINITAB)	P
RT X DOC	-	0.275
RT	Checklist= 11.417; PBR= 13.346	0.404
DOC	ATM= 9.310; PG= 15.453	0.005✓

ANOVA summary table with relation to the individual efficiency

Independent Variables	Efficiency (average)	P
RT X DOC	-	0.417
RT	Checklist= 2.775; PBR= 3.856	0.101
DOC	ATM= 2.817; PG= 3.814	0.131

- **O2') Do individual PBR or Checklist reviewers find more defects?**

Document	ATM		PG	
Technique	**Checklist**	**PBR**	**Checklist**	**PBR**
Defects Found/Total defects	15/37 (40.5%)	21/37 (56.8%)	20/32 (60.5%)	14/32 (43.75%)
Occurrences of Defects/Total occurrences	24/333	38/333	45/288	44/288
Effectiveness	7.21	11.41	15.63	15.28
Efficiency	2.00	3.62	3.53	4.10

Analysis

- **ATM**: PBR found a higher percentage of the defects than checklist.

p-value = 0.143 (not statistically significant at the .05 level)

- **PG**: Checklist found a higher percentage of the defects than PBR.

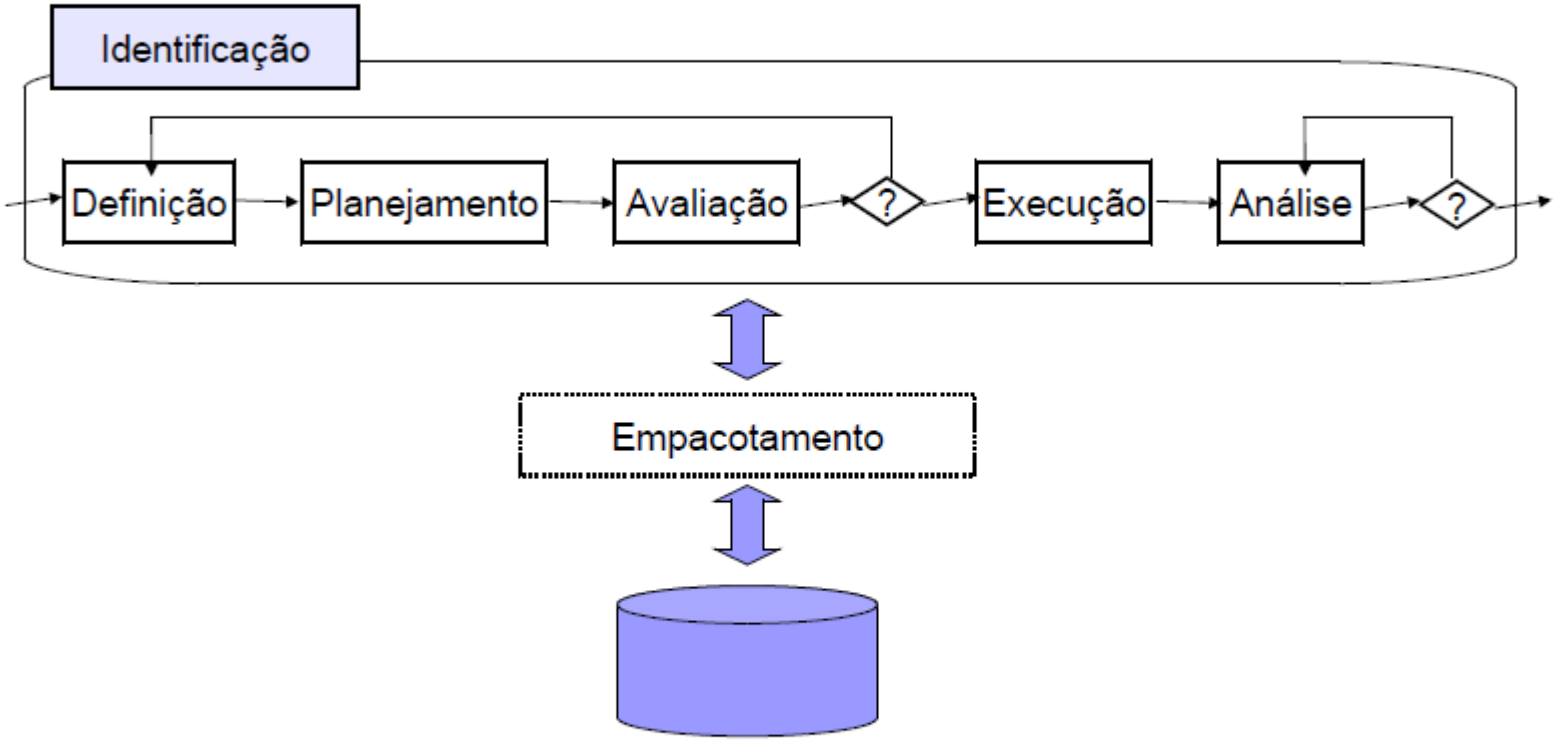
p-value = 0.911 (not statistically significant at the .05 level)

- **Efficiency** (errors/hour): PBR were more efficient for both documents.

ATM p-value=.107, PG pvalue=.51 (not statistically significant at the .05 level)

EMPACOTAMENTO

Processo de Experimentação



EMPACOTAMENTO

- Documente os resultados e conclusões para que eles possam ser usados pelos seus colegas
- Documente todos os aspectos importantes do experimento para facilitar a sua replicação
  - Objetivos, hipóteses, variáveis, tratamentos e dados obtidos.
  - Instruções, descrição do processo experimental, ferramentas e artefatos, manuais e material de treinamento, etc.
- Descreva problemas encontrados, lições aprendidas, e sugestões de evolução do experimento e do material utilizado.

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Tipos de Estudos Experimentais
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

Investigar o desempenho usando PSP (personal software process)

- PSP é um processo para uso individual
- O processo inclui as seguintes atividades: medição, estimação, planejamento e rastreamento, reúso de dados e de experiência.
- Autor: Watts Humphrey

- Objeto de estudo: participantes do curso de PSP e suas habilidades em termos de formação e experiência.
- Propósito: avaliar o desempenho com base na formação (CSE ou EE)
- Perspectiva: pesquisadores e professores.
- Foco qualitativo: Produtividade (Kloc/tempo de desenvolvimento) e Índice de defeitos (falhas/kloc)
- Contexto: Curso do Depto CS, Un. Lunden, Uso de C, 65 estudantes, quase-experiment, por que não há randomização de estudantes.

Propósito

Analisar *os resultados de PSP*
para o propósito de *avaliação*
com respeito à *formação dos indivíduos*
do ponto de vista de *pesquisadores e professores*
no contexto de um *curso de PSP.*

Planejamento

Contexto

- Curso na universidade
- Off-line (não é um ambiente industrial)
- Experimento específico, focado em PSP.
- Alunos de 4º ano da graduação
- O experimento trata de um problema real: diferenças nos desempenhos individuais e o entendimento dessas diferenças.

Hipóteses

- Estudantes de CSE e EE assistem o curso.
- CSE tem mais matérias de ciências de computação e engenharia de software e portanto, deveriam ter melhor desempenho.
- Na primeira aula eles responderão a um questionário (conhecimento de C). Estudantes com maior experiência em C deveriam cometer menos erros.

Hipóteses formalizadas

- Produtividade
  - H0: Prod(CSE) = Prod (EE)
  - H1: Prod(CSE) != Prod (EE)

- Experiência
  - H0: Número de falhas por Kloc é independente da experiência em C.
  - H1: Número de falhas por Kloc muda com experiência em C.

Dados que precisam ser coletados

- Tipo do estudante: CSE ou EE (escala nominal)
- Produtividade: Kloc/tempo de desenvolvimento (escala quociente)
  - Loc será contada usando um programa, serão contadas as linhas novas e modificadas.
  - Desenvolvimento será medido em minutos
- Experiência será contada usando uma classificação (escala ordinal)
  - 1. Sem experiência anterior
  - 2. Leu um livro ou fez um curso
  - 3. Alguma experiência profissional (Menos que 6 meses)
  - 4. Experiência profissional (Mais que 6 meses)
- Falhas/Kloc -- medida de produtividade

- Randomização: não há. PSP não é avaliado contra outra técnica. As medidas serão tomadas em relação ao desenvolvimento de 10 programas.
- Blocagem: não há. Serão usados os dados dos 10 programas juntos
- Balanceamento do conjunto de dados: não é possível, pois não há como escolher os estudantes de cada tipo de curso.
- Seleção de variáveis
  - Independentes: curso do estudante e experiência em C
  - Dependentes: produtividade e falhas/kloc
- Seleção dos sujeitos
  - Por conveniência, sem randomização
  - São os estudantes que fazem o curso

Tipos de Projeto

- 1. um fator e dois tratamentos: o programa e os dois cursos
  - A variável dependente é media na escala quociente e o teste paramétrico é adequado. T-test será usado.
- 2. um fator com mais de dois tratamentos: o fator é a experiência em C e os tratamentos são as quatro classificações.
  - A variável dependente é medida em uma escala quociente e ANOVA e adequada para avaliação.

Avaliação da validade

- Validade interna: o grande número de testes assegura uma boa validade interna.
- Validade externa: deve valer provavelemente para outras edições do mesmo curso. Pode não valer para outros cursos e estudantes (podem não estar interessados em desenvolvimento de software). Pode valer para outros cursos de PSP.

Avaliação da validade (cont.)

- Validade de construcão: a validade dos dados pode ser um problema. Os estudantes devem coletar muitos dados e podem errar ou inventar. Eventuais inconsistências, não devem estar relacionadas com o tipo de curso do estudante. Não parece ser um problema crítico.
- Validade da conclusão: 1) as medidas são apropriadas? 2) os alunos receberão notas(os alunos foram avisados no início do curso que as notas não dependem dos dados, mas sim de entregar no prazo e de forma correta.)

Operação

- Preparação
  - Os estudantes não foram informados do que se pretendia estudar.
  - Anonimato foi garantido
- Execução
  - Durante 14 semanas.
  - Dados coletados por meio de formulários
  - Entrevista no final.

Operação (cont.)

- Validação dos dados
  - Dados coletados de 65 estudantes
  - Dados de 6 estudantes foram desconsiderados
    - Dois não preencheram corretamente os formulários
    - Um estudante terminou os trabalhos muito tempo depois.
    - Dois entregaram os dados sempre atrasados e precisaram de muito apoio.
    - Um estudante tinha formação completamente diferente.

Análise e interpretação

Estatística Descritiva

Produtividade

- A população foi dividida em classes com base na produtividade. Oito classes: 5-10 loc/h, ...,40-45 loc/h.
- Estudantes de EE parecem ter menor produtividade
- Estudantes de CSE claramente tem maior variação.
- 32 estudantes de CSE com média 23 e dp 8.7
- 27 estudantes de EE com média 16.4 e dp 6.3

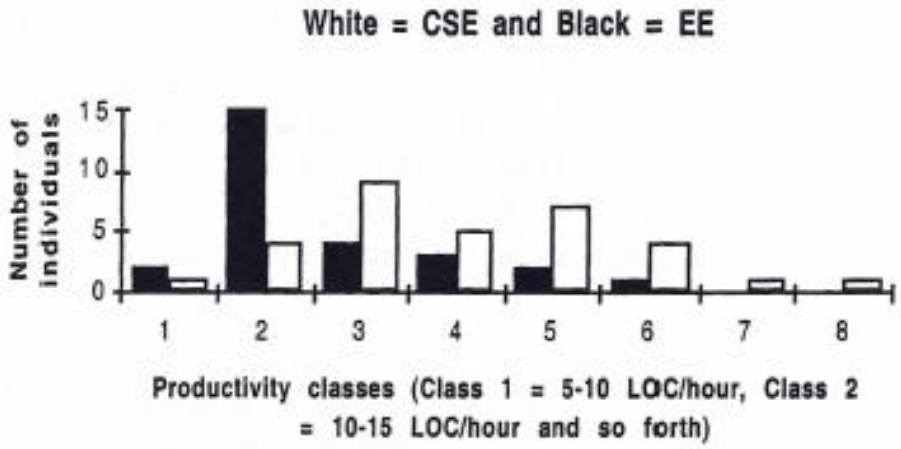


Figure 26. *Frequency distribution for the productivity (in classes).*

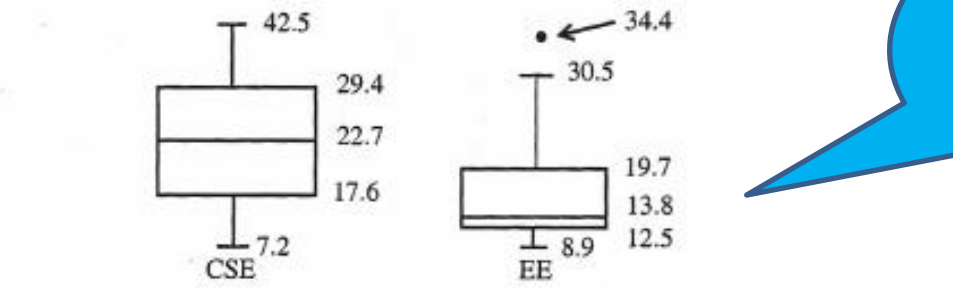


Figure 27. *Box plot of productivity for the two study programs.*

É possível provar essa análise com teste de hipótese?

Tamanho da caixa= 29.4-17.6=11.8

ut= 29.4+1.5*11.8=47.1 →42.5 (maior valor da amostra)

Experiência em C x no. de falhas.

Table 28. *Faults/KLOC for the different C experience classes.*

Class[a]	Number of students	Median value of faults/ KLOC	Mean value of faults/KLOC	Standard deviation of faults/KLOC
1	32	66.8	82.9	64.2
2	19	69.7	68.0	22.9
3	6	63.6	67.6	20.6
4	2	63	63.0	17.3

a. The different classes are explained in Section 11.2.2.

A distribuição é ligeiramente inclinada para pouca ou nenhuma experiência.
Uma pequena tendência, pelas médias, de menos falta com mais experiência.
DP é alto
O DP da primeira classe é muito alto

Box plot para a primeira classe

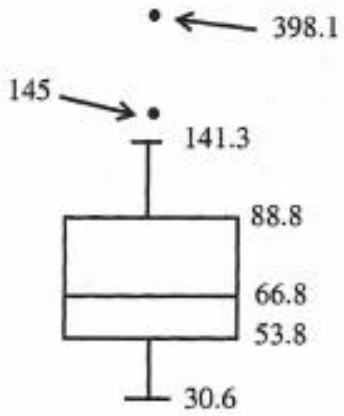


Figure 28. *Box plot for faults/KLOC for class 1.*

As bp das classes 2-4 são pouco interessantes
Na classe 1, o alto dp pode ser explicado pelo outlier

Redução do conjunto de dados.

- É sempre controvertido reduzir o conjunto de dados, pois há o risco de estarmos favorecendo um certo resultado.
- Há métodos estatísticos para redução: análise do componente principal e análise de fatores [Kaching86, Manly94]
- É sempre melhor ser restritivo.

Novo resultado retirando-se o outlier da classe 1.

Table 29. *Faults/KLOC for C experience class 1.*

Class	Number of students	Median value of faults/ KLOC	Mean value of faults/KLOC	Standard deviation of faults/KLOC
1	31	66	72.7	29.0

Teste de Hipótese

T-test (não pareado, dois-tailed)
Pela tabela 30, H0 é rejeitada : há uma diferença significativa em produtividade para os estudantes de diferentes cursos.
O p-value é muito baixo e portanto os resultados são significativos.

Table 30. *Results from the t-test.*

Factor	Mean diff.	Degrees of freedom (DF)	t-value	p-value
CSE vs. EE	6.1617	57	3.283	0.0018

As classes 2, 3 e 4 foram agrupadas, pois não diferenças significativas entre elas
Nenhuma resultado significativo resulta da comparação da classe 1 com as classe agrupada
H0 não pode ser rejeitada.

Table 31. *Results from the ANOVA-test.*

Factor: C vs. Faults/KLOC	Degrees of freedom (DF)	Sum of squares	Mean square	F-value	p-value
Between treatments	3	3483	1160.9	0.442	0.7236
Error	55	144304	2623.7		

Agenda

- 1. Motivação
- 2. Conceitos Básicos
- 3. Engenharia de Software Experimental
- 4. Tipos de Estudos Experimentais
- 5. Como executar um Estudo Experimental
- 6. Um exemplo: Processo
- 7. Conclusões e Bibliografia

Concluindo ...

- Decisões sobre tecnologias devem ser baseadas em raciocínio científico e evidência empírica.
- Este material contém apenas uma introdução aos conceitos de experimentação em ES
- Para fazer um experimento é necessário ler a bibliografia sobre o assunto e deve-se procurar um perito na área para apoiar o trabalho.
- Projetos experimentais devem ser cuidadosamente planejados:
  - São trabalhosos,
  - Não dão resultado imediato,
  - Precisam de apoio da gerência da empresa.

Concluindo ... (cont.)

- Por ser intensivamente humana, a Eng. de Software tem características de experimentos em ciências sociais:
  - Difícil de controlar
  - Difícil de coletar dados
  - Envolvem dados subjetivos e qualitativos
- Por usar fortemente ferramentas, pode ter partes automatizadas
  - Existe pouco suporte automatizado para experimentação
- Quanto mais cuidadoso for o projeto e o controle...
  - Mais confiança nos resultados,
  - Melhor compreensão do problema,
  - Mais efetivas são as nossas ações e os resultados obtidos.

Endereços úteis

- ESELAW:
  - http://lens.cos.ufrj.br:8080/eselaw
- ESEM:
  - http://www.esem.org
- International Software Engineering Research Network (ISERN):
  - http://www.iese.fhg.de/ISERN/
- Experimental Software Engineering Latin-American Network (ESELAN) discussion list :
  - http://listas.cos.ufrj.br/mailman/listinfo/eselan-l

Bibliografia Básica

- Wohlin, C. Experimentation in Software Engineering, Kluwer Academic Publishers, 2000.
- Juristo, N. And Moreno, A. Basics of Software Engineering Experimentation, Kluwer Academic Publishers, 2000.
- MALDONADO, J. C.; CARVER, J.; SHULL, F.; FABBRI, S. C. P. F.; DÓRIA, E. S.; MARTIMIANO, L. A. Fondazzi; MENDONÇA, M. G. de; BASILI, V.. Perspective-Based Reading: A Replicated Experiment Focused on Individual Reviewer Effectiveness. Empirical Software Engineering, v. 11, n. 1, p. 119-142, 2006.

Bibliografia Complementar

- Tichy, W. Should Computer Scientists Experiment More?, IEEE Computer, May 1998.
- Zelkowitz, M. and Wallace, D. Experimental Models for Validating Technology?, IEEE Computer, May 1998.
- Marcus Ciolkowski, Oliver Laitenberger, Sira Vegas, and Stefan Biffl. Practical Experiences in the Design and Conduct of Surveys in Empirical Software Engineering, In. R. Conradi and A.I. Wang (Eds.): ESERNET 2001- 2003, LNCS 2765, pp. 104–128, 2003.
- Barbara A. Kitchenham, Tore Dybå, and Magne Jørgensen. Evidence-based Software Engineering. Proceedings of the 26th International Conference on Software Engineering (ICSE'04), 2004.
- Basili, V., "Evolving and Packaging Reading Technologies." Journal of Systems and Software, 1997. 38(1): p. 3-12.
- Basili, V., R. Selby, D. Hutchens, "Experimentation in Software Engineering." IEEE Transactions in Software Engineering, 1986. 12 (7), 733–743.
- Basili, V., F. Shull, and F. Lanubile, "Building Knowledge through Families of Experiments." IEEE Transactions on Software Engineering, 1999. 25(4): p. 456-473.
- Biolchini, J., P. Mian, A. Natali, and G. Travassos, "Systematic Review in Software Engineering." Technical Report ES 679/05, PESC, Federal University of Rio de Janeiro, 2005. Available at http://cronos.cos.ufrj.br/publicacoes/reltec/es67905.pdf
- Brooks, A., M. Roper, M. Wood, J. Daly and J. Miller, "Replication of Software Engineering Experiments." Empirical Foundation of Computer Science Technical Report, EFoCS-51-2003, Department of Computer and Information Sciences, University of Strathclyde University, 2003. Available at http://www.cis.strath.ac.uk/~efocs/home/Research-Reports/EFoCS-51-2003.pdf
- Curtis, B., "Measurement and experimentation in software engineering." Proceedings of the IEEE, 1990 68(9) 1144–1157.

Bibliografia Complementar

- Miller, J. "Replicating Software Engineering Experiments: A Poisoned Chalice or the Holy Grail." Information and Software Technology. 47(4), 2005, pp. 233-244.
- Shull, F., V. Basili, J. Carver, J. Maldonado, G. Travassos, M. Mendonca, and S. Fabbri. "Replicating Software Engineering Experiments: Addressing the Tacit Knowledge Problem." Proceedings of International Symposium on Empirical Software Engineering (ISESE'02). 2002. Nara, Japan, 7-16.
- Shull, F., J. Carver, G. Travassos, J. Maldonado, R. Conradi, and V. Basili, Replicated Studies: Building a Body of Knowledge about Software Reading Techniques, in Lecture Notes on Empirical Software Engineering, N. Juristo and A. Moreno, Editors. 2003, World Scientific.
- Shull F., Cruzes D., Basili V. R., and Mendonca M., "Simulating Families of Studies to Build Confidence in Defect Hypotheses," Information and Software Technology, 47(15), pp. 1019-1032, 2005.
- Shull, F., M. Mendonca, V. Basili, J. Carver, J. Maldonado, S. Fabbri, G. Travassos, and M. Ferreira, "Knowledge-sharing Issues in Experimental Software Engineering." Empirical Software Engineering – An International Journal, 2004. 9(1): p. 111-137.
- Sjøberg, D., J. Hannay, O. Hansen, V. Kampenes, A. Karahasanović, N. Liborg, A. Rekdal, "A Survey of Controlled Experiments in Software Engineering". IEEE Transactions on Software Engineering, 2006. 31(9): p. 733-753.
- Tichy, W., P. Lukowicz, L. Prechelt, and E.A. Heinz, “Experimental Evaluation in Computer Science: A Quantitative Study,” J. Systems and Software, vol. 28, no. 1, pp. 9-18, Jan. 1995.
- Wood, M., J. Daly, J. Miller, M. Roper, "Multi-method research: an empirical investigation of objec oriented technology." Journal of Systems and Software 48(1) 13–26, 1999.
- Zendler, A. “A Preliminary Software Engineering Theory as Investigated by Published Experiments,” Empirical Software Eng., vol. 6, no. 2, pp. 161-180, 2001.

Bibliografia Complementar

- J. Daly, "Replication and a Multi-Method Approach to Empirical Software Engineering research." PhD Thesis, Department of Computer Science, University of Strathclyde, 1996.
- Glass, R., I. Vessey, and V. Ramesh, “Research in Software Engineering: An Analysis of the Literature,” J. Information and Software Technology, vol. 44, no. 8, pp. 491-506, June 2002.
- Jedlitschka, A. and M. Ciolkowski, "Towards Evidence in Software Engineering." Proceedings of the 2004 International Symposium on Empirical Software Engineering (ISESE’04), Redondo Beach, California, pp. 261- 270, 2004.
- Jedlitschka, A. and D. Pfahl, "Reporting Guidelines for Controlled Experiments in Software Engineering.“ International Software Engineering Network Technical Report, ISERN-55-01, 2005.
- Kamsties, E., C. Lott, "An empirical evaluation of three defect detection techniques." International Software Engineering Network Technical Report, ISERN-95-02, 1995.
- Kitchenham, B. "Procedures for Performing Systematic Reviews." Technical Report TR/SE-0401, Keele University, and Technical Report 0400011T.1, NICTA, 2004.
- Kitchenham, B., S. Pfleeger, L. Pikard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary Guidelines for Empirical Research in Software Engineering." IEEE Transactions on Software Engineering, 2002. 28(8): p. 721-734.
- Lott, C., H. Rombach, "Repeatable Software Engineering Experiments for Comparing Defect-detection Techniques." Journal of Empirical Software Engineering, 1(3) 1997, 241–277.
- MALDONADO, J. C.; CARVER, J.; SHULL, F.; FABBRI, S. C. P. F.; DÓRIA, E. S.; MARTIMIANO, L. A. Fondazzi; MENDONÇA, M. G. de; BASILI, V.. Perspective-Based Reading: A Replicated Experiment Focused on Individual Reviewer Effectiveness. Empirical Software Engineering, v. 11, n. 1, p. 119-142, 2006.
- Miller, J. "Applying meta-analytical procedures to software engineering experiments." Journal of Systems and Software. 54(1), 2004, pp. 29-39.