

User Story Mapping e Product Backlog Building para criação de Backlogs: um Relato de Experiência

Sara Cristina Monteiro Matos
Instituto de Ciências Exatas e Naturais (ICEN)
Universidade Federal do Pará
Belém, Brasil
sara.matos@icen.ufpa.br

João Pedro Vasconcelos de Lima
Instituto de Ciências Exatas e Naturais (ICEN)
Universidade Federal do Pará
Belém, Brasil
joao.vasconcelos@icen.ufpa.br

Ronivaldo Ferreira
Instituto de Ciências Exatas e Naturais (ICEN)
Universidade Federal do Pará
Belém, Brasil
ronivaldo.junior@icen.ufpa.br

Victor Hugo Santiago C. Pinto
Instituto de Ciências Exatas e Naturais (ICEN)
Universidade Federal do Pará
Belém, Brasil
victor.santiago@ufpa.br

Resumo

A construção inicial do *Product Backlog* a partir de documentos pouco estruturados é uma competência exigida na prática profissional, mas raramente ensinada de forma sistemática em disciplinas de Engenharia de Software. Neste relato de experiência, comparamos duas técnicas de construção de backlog em uma intervenção com 26 estudantes de graduação: *User Story Mapping* (USM) e *Product Backlog Building* (PBB). A intervenção adotou um desenho quase-experimental em que cada um dos seis grupos aplicou as duas técnicas, com contrabalanceamento de domínio entre dois Termos de Referência reais distintos, um sistema de Gestão de Viagens, mais processual, e a plataforma Marajó Resiliente, mais exploratória. Três especialistas avaliaram os artefatos produzidos quanto a completude, clareza, estruturação, priorização e coerência do fluxo, e os estudantes responderam a instrumentos sobre facilidade de uso, utilidade e carga cognitiva. Observamos um contraste relevante: o PBB foi descritivamente superior na qualidade dos artefatos, sobretudo em estruturação e clareza, enquanto o USM foi percebido pelos estudantes como mais claro e fácil de aplicar. A priorização foi o critério mais frágil em ambas as técnicas, o que expõe uma lacuna pedagógica que o método, por si só, não supre. Como contribuições, apresentamos uma intervenção didática replicável baseada em documentos realistas, evidências sobre quando cada técnica apoia melhor a aprendizagem e discutimos a complementaridade entre PBB e USM no ensino de requisitos ágeis.

Palavras-chave

Product Backlog Building, Story Mapping, User Stories, Experience Report

1 Introdução

A construção inicial do *Product Backlog* é uma atividade central no desenvolvimento ágil e uma competência cada vez mais exigida na prática profissional, mas seu ensino sistemático ainda é pouco explorado em disciplinas de Engenharia de Software. Estudantes costumam aprender a estrutura textual de uma história de usuário no formato “Como [papel], quero [objetivo], para [benefício]”, mas enfrentam dificuldades quando precisam transformar documentos extensos e pouco estruturados — como Termos de Referência,

demandas institucionais e descrições de produto — em itens de backlog compreensíveis, priorizáveis e orientados a valor. Essa lacuna é relevante porque a literatura em Engenharia de Requisitos Ágil mostra que a qualidade de histórias de usuário e backlogs depende de aspectos que vão muito além da fórmula textual, como clareza, valor, granularidade, testabilidade e priorização [10, 13, 22].

Para apoiar essa atividade, duas técnicas têm ganhado destaque na prática profissional: o *User Story Mapping* (USM) [15], que organiza histórias de usuário em torno de jornadas de uso e fatias incrementais de entrega, e o *Product Backlog Building* (PBB) [2], que apoia a decomposição do produto a partir de problemas, expectativas, personas e funcionalidades. Embora o USM tenha presença consolidada na literatura, o PBB ainda é pouco investigado em contextos educacionais, especialmente quando comparado diretamente a técnicas mais difundidas. Como consequência, há poucas evidências sobre como cada técnica apoia diferentes etapas da construção de backlog em sala de aula e sobre quando uma é preferível à outra.

Diante dessa lacuna, este estudo tem como objetivo avaliar e comparar o uso das técnicas de User Story Mapping (USM) e Product Backlog Building (PBB) na construção inicial de backlogs de produto e histórias de usuário por estudantes de Engenharia de Software. Para isso, apresentamos um relato de experiência realizado em uma disciplina de graduação com 26 estudantes, organizados em seis grupos. Os grupos aplicaram ambas as técnicas a partir de dois Termos de Referência com naturezas distintas: um sistema de Gestão de Viagens, mais processual [21], e a plataforma Marajó Resiliente, mais exploratória [24]. O estudo adotou um desenho quase-experimental com contrabalanceamento de domínio, permitindo análises pareadas nos níveis de grupo e individual. Os artefatos produzidos foram avaliados por três especialistas, segundo cinco critérios de qualidade, e os estudantes responderam a instrumentos sobre facilidade de uso, utilidade e carga cognitiva.

Os resultados revelam um contraste pedagogicamente relevante entre o que os artefatos mostram e o que os estudantes percebem. Especialistas avaliaram os artefatos produzidos com PBB como superiores em estrutura e clareza, e essa superioridade se manteve em ambos os domínios estudados, com maior vantagem no domínio mais processual. O USM, por sua vez, manteve-se estável em ambos os domínios, sem se destacar em nenhum dos critérios de qualidade do artefato. Na percepção dos estudantes, o padrão se

inverte: o USM foi avaliado como mais claro, mais fácil de aplicar e menos custoso em esforço do que o PBB. A priorização foi o ponto fraco comum, expondo uma lacuna pedagógica independente do método. Tal divergência sugere que as técnicas são complementares e contextuais, não substitutivas.

As contribuições deste relato são três: (1) descrevemos uma intervenção replicável para o ensino de construção de backlog a partir de documentos realistas, com instrumentos, rubrica de avaliação e materiais disponibilizados; (2) oferecemos evidências comparativas entre PBB e USM quanto à qualidade dos artefatos, à percepção dos estudantes e à carga cognitiva, ampliando a discussão sobre o PBB como abordagem pedagógica ainda pouco explorada na literatura; (3) articulamos lições aprendidas sobre quando cada técnica apoia melhor a aprendizagem e sobre a necessidade de tratar a priorização como objeto pedagógico explícito, fornecendo recomendações acionáveis para docentes que ensinam requisitos ágeis.

2 Background

O *Product Backlog* representa uma lista ordenada e evolutiva de itens de valor, servindo como ferramenta de comunicação e decisão [20]. No Scrum, esse artefato é continuamente refinado e ajustado conforme novas informações sobre o produto. Mais que uma lista de tarefas, o backlog é um mecanismo de comunicação e decisão entre o time e os *stakeholders*, e sua qualidade impacta diretamente a evolução incremental do produto.

As histórias de usuário são comumente usadas para representar itens de backlog em projetos ágeis, seguindo a estrutura “Como [papel], quero [objetivo], para [benefício]” [5]. No entanto, evidências da literatura mostram que essa fórmula textual, a eficácia dessa técnica depende menos da sintaxe e mais da granularidade, ausência de ambiguidade e orientação a valor [13]. Atualmente, são vistas como objetos de fronteira entre diferentes papéis envolvidos no desenvolvimento, reforçando que sua produção é um processo colaborativo, e não apenas um exercício de redação [22].

O *User Story Mapping* (USM) [15] é uma técnica que organiza visualmente as histórias em torno da jornada de uso. Em vez de tratar o backlog como uma lista linear, o USM estrutura o produto em três níveis hierárquicos — atividades, tarefas e histórias — dispostos em um mapa bidimensional. O eixo horizontal representa a sequência narrativa do uso, enquanto o eixo vertical organiza as histórias em fatias incrementais de entrega, o que apoia a definição de produtos mínimos viáveis e de *releases*. Essa organização favorece a visualização do produto como um todo, evidenciando as dependências entre as etapas e ajuda a equipe a discutir o que entregar primeiro.

O *Product Backlog Building* (PBB) [2] utiliza um *canvas* para a construção estruturada do backlog através de etapas encadeadas: visão, problemas, expectativas, personas, funcionalidades e itens de backlog (PBIs). Enquanto o USM, foca na jornada de uso, o PBB enfatiza a decomposição conceitual do produto: a partir de problemas e personas, para chegar a funcionalidades e itens prioritizáveis, oferecendo um caminho mais explícito para articular valor, papéis e entregas. Apesar dessa estrutura, o PBB ainda é pouco discutido na literatura de Educação em Engenharia de Software, sobretudo em comparação direta com técnicas mais consolidadas como o USM.

USM e PBB, portanto, respondem a necessidades distintas da construção inicial de backlog: o primeiro apoia a organização narrativa da experiência de uso, enquanto o segundo apoia a estruturação conceitual do produto a partir de problemas e personas. Essa complementaridade motiva este estudo, que compara como cada técnica apoia estudantes na transformação de documentos realistas em backlogs e histórias de usuário.

3 Trabalhos Relacionados

A literatura em Educação em Engenharia de Software investiga estratégias para alinhar o ensino às práticas industriais. Trabalhos baseados em aprendizagem por projetos relatam ganhos em competências técnicas e de gestão ao utilizar problemas realistas [7, 18]. Outros estudos exploram o ensino de Scrum por meio de dinâmicas práticas e com colaboração remota [6, 14]. Esses trabalhos reforçam o valor de estratégias práticas, visuais e colaborativas para o ensino de métodos ágeis. Nosso estudo se diferencia desse conjunto ao focar especificamente na fase inicial de descoberta e organização do backlog, com comparação direta entre duas técnicas estruturadas.

Sobre a qualidade das histórias de usuário, Bennett e Onek [3] aplicaram os critérios INVEST, notando dificuldades em estimabilidade e testabilidade. Péraire [16] propôs o modelo 4C para apoiar a escrita de histórias em sala de aula. Sporse et al. [22] sintetizam discussões recentes sobre histórias como objetos de fronteira, cuja qualidade depende do processo colaborativo. Esses estudos dialogam com nossos resultados, mas evidenciam uma lacuna na literatura: a maioria das abordagens pedagógicas foca na redação e no artefato final, negligenciando o processo cognitivo de abstração e organização para extrair esses itens de documentos não estruturados. Avançamos ao deslocar o foco pedagógico da mera escrita para as técnicas de origem do backlog. Comparar abordagens como PBB e USM permite investigar não apenas se o estudante sabe formatar um requisito, mas como diferentes estruturas conceituais apoiam sua carga cognitiva e a transformação de problemas complexos em incrementos de valor.

Este relato contribui em três frentes ainda pouco exploradas na literatura. Primeiro, comparamos diretamente duas técnicas estruturadas de organização do backlog utilizando Termos de Referência reais como insumo, expondo os estudantes a documentos próximos da prática profissional. Em segundo lugar, garantimos a robustez da avaliação submetendo os artefatos a especialistas com perfis complementares (acadêmico, Product Owner e técnico), com alta confiabilidade ($ICC = 0,978$). Terceiro, triangulamos a qualidade do artefato, a percepção discente e a carga cognitiva, oferecendo um olhar amplo que fundamenta lições aprendidas para o ensino de métodos ágeis.

4 Contexto Educacional e Objetivo do Estudo

4.1 Disciplina e Contexto da Intervenção

A intervenção foi realizada em uma disciplina obrigatória de Engenharia de Software, com carga horária de 60 horas, ofertada no quarto semestre de um curso de Bacharelado em Sistemas de Informação. A atividade ocorreu em formato presencial, em quatro encontros sucessivos de 01h40, distribuídos em duas semanas, com dois encontros por semana. Quando a intervenção foi conduzida, os estudantes já haviam tido contato com conceitos fundamentais

de Engenharia de Requisitos e métodos ágeis, incluindo requisitos funcionais e não funcionais, histórias de usuário, critérios de aceitação, priorização, *Product Backlog* e planejamento incremental. A atividade não teve peso na nota final, sendo conduzida como prática pedagógica integrada ao módulo de requisitos ágeis da disciplina. Essa decisão buscou reduzir a pressão sobre o desempenho individual e favorecer a discussão entre pares nas sessões, sem que os estudantes se sentissem penalizados por dificuldades inerentes à aplicação de técnicas com as quais ainda tinham pouca prática.

4.2 Participantes e Considerações Éticas

A turma era composta por 30 estudantes, dos quais 26 consentiram voluntariamente via Termo de Consentimento Livre e Esclarecido (TCLE) em sala. Os estudantes que não aderiram realizaram as atividades normalmente, sem qualquer prejuízo no acompanhamento da disciplina, apenas seus dados não foram coletados nem analisados. Antes da intervenção, aplicou-se um questionário de caracterização para mapear o conhecimento prévio da turma. De modo geral, os participantes possuíam familiaridade com práticas tradicionais de engenharia de requisitos (como casos de uso e entrevistas) e conhecimento teórico sobre User Stories, mas apresentavam baixa exposição prática à execução contínua de métodos ágeis e autopercepção de habilidade inicial baixa tanto em USM quanto em PBB (detalhamento quantitativo na Seção 6.1).

A equipe da pesquisa, composta pelo docente responsável, um monitor de pós-graduação e uma estudante de iniciação científica, formou manualmente seis grupos de três a quatro integrantes, equilibrando perfis com diferentes níveis de experiência. Essa estratégia visou mitigar vieses de desempenho técnico e simular a escala de equipes ágeis de pequeno porte. Os grupos permaneceram fixos para garantir a continuidade da dinâmica.

Todos os dados foram anonimizados antes da análise, como critério de exclusão, definiu-se previamente que estudantes ausentes nos nivelamentos conceituais, que consistia na apresentação das técnicas e piloto, teriam seus dados descartados. Variações pontuais no quantitativo de respostas ao longo do estudo decorrem desses critérios de exclusão, da adesão voluntária e de faltas ocasionais em sessões experimentais específicas.

4.3 Objetos do Estudo

Para a execução da atividade, adotou-se dois Termos de Referência (TR) reais, obtidos de fontes públicas, ambos contendo demandas concretas de contratação de software. A escolha por TRs, visou aproximar os estudantes da prática profissional, na qual requisitos raramente estão estruturados de forma sistemática e exigem esforço de interpretação e decomposição. A seleção combinou uma busca ativa por documentos contrastantes e uma oportunidade institucional gerada pelo segundo TR, vinculado a um projeto socioambiental regional. Os domínios foram escolhidos por representarem desafios distintos na elaboração de um backlog ágil.

O primeiro domínio foi um sistema de Gestão de Viagens, descrito em um TR publicado por uma instituição pública de apoio a micro e pequenas empresas [21]. O documento, com cerca de 12 páginas, organiza as informações em um fluxo administrativo definido, com etapas de cotação, autorização, aquisição, consulta de status e prestação de contas. O TR identifica explicitamente os

atores envolvidos, apresenta diagramas de processos e detalha um conjunto de oito relatórios gerenciais. O documento também especifica restrições técnicas (performance e plataforma) e requisitos de segurança. Tais características fazem dele um domínio com fluxo operacional dominante, atores bem identificados e granularidade homogênea, permitindo inferir jornadas de uso com clareza.

O segundo domínio foi a plataforma Marajó Resiliente, provém de um TR de uma organização voltada à inovação socioambiental [24]. Com cerca de cinco páginas, o documento descreve uma plataforma digital de Gestão do Conhecimento para disseminação de dados e indicadores climáticos no Arquipélago do Marajó. Diferente do primeiro TR, este não se organiza em um único fluxo linear, mas em capacidades transversais: georreferenciamentos, repositórios colaborativos, painéis de indicadores e suporte a baixa conectividade. Os atores são diversos e não sequenciais (gestores, técnicos, agricultores familiares, quilombolas, dentre outros), e a granularidade dos requisitos é heterogênea, variando de conceitos abstratos como “design centrado no usuário”, a funções concretas de “catálogo multimídia”. Trata-se, portanto, de um domínio exploratório que exige maior abstração para identificar personas e objetivos.

Em conjunto, os dois TRs oferecem um contraste pedagógica-mente relevante: o primeiro favorece o raciocínio em torno de fluxos e jornadas de uso, enquanto o segundo demanda decomposição conceitual a partir de problemas, perfis e capacidades. Esse contraste é central para examinar como os alunos se comportam diante do uso de técnicas e de problemas com diferentes graus de estruturação.

4.4 Objetivo e Questões de Pesquisa

O objetivo do estudo foi avaliar e comparar o uso de USM e PBB na construção de backlogs de produto e histórias de usuário por estudantes. Em particular, buscou-se observar como cada técnica apoia a interpretação de requisitos, a identificação de funcionalidades, a organização do artefato e a escrita de histórias de usuário, bem como avaliar a percepção do discente durante a atividade. O estudo foi orientado pelas seguintes questões de pesquisa:

RQ1. Há diferença na qualidade dos backlogs produzidos quando os estudantes utilizam USM em comparação com PBB?

RQ2. Como os estudantes percebem a facilidade de uso e a utilidade de cada técnica para a elaboração de histórias de usuário?

RQ3. Há diferença na carga cognitiva percebida pelos estudantes ao utilizar USM e PBB?

Essas questões visam analisar tanto a qualidade técnica dos artefatos produzidos quanto a experiência dos estudantes, considerando a facilidade de uso, a utilidade e a carga cognitiva.

5 Metodologia

Esta seção descreve o desenho do estudo, as variáveis e métricas, o procedimento de execução, a avaliação dos artefatos e os métodos de análise dos dados utilizados para responder à Questão Principal (QP) do estudo: De que forma as técnicas PBB e USM impactam a construção de backlogs por estudantes de graduação, considerando a qualidade dos artefatos gerados, a utilidade percebida e o esforço cognitivo exigido? A investigação foi conduzida como um relato de experiência com desenho quase-experimental em ambiente de sala de aula, seguindo recomendações gerais para estudos empíricos em Engenharia de Software [17, 23].

5.1 Desenho do Estudo

Os estudantes foram organizados em seis grupos, distribuídos em dois conjuntos. Optou-se por padronizar a técnica aplicada em cada sessão experimental (Sessão 1: PBB, Sessão 2: USM). Essa ausência de contrabalanceamento na ordem das técnicas foi uma decisão intencional para viabilizar a execução dentro da carga horária da disciplina. O contrabalanceamento ocorreu exclusivamente nos domínios: o Conjunto A (Grupos 1, 2 e 3) aplicou PBB em Marajó Resiliente e USM em Gestão de Viagens, enquanto o Conjunto B (Grupos 4, 5 e 6) realizou a ordem inversa dos domínios. A Tabela 1 apresenta o desenho geral do estudo.

Table 1: Desenho quase-experimental do estudo.

Conjunto	Sessão 1 (PBB)	Sessão 2 (USM)
A (1, 2 e 3)	PBB × Marajó Resiliente	USM × Gestão de Viagens
B (4, 5 e 6)	PBB × Gestão de Viagens	USM × Marajó Resiliente

Esse arranjo permitiu que ambas as técnicas fossem aplicadas aos dois domínios por diferentes conjuntos de grupos, viabilizando análises pareadas no nível do grupo (cada grupo aplicou as duas técnicas) e a comparação entre técnicas e domínios. Reconhecemos, no entanto, duas limitações relevantes do desenho. Primeiro, a ordem das técnicas não foi contrabalanceada: todos os grupos aplicaram PBB antes de USM. Essa decisão decorreu da opção logística de padronizar a técnica em cada sessão, mas introduz a possibilidade de que as diferenças observadas entre PBB e USM possam, em parte, refletir efeitos de ordem. Segundo, a ordem dos domínios também não foi totalmente isolada do contrabalanceamento, uma vez que cada conjunto sempre encontrou um domínio na primeira sessão e o outro na segunda. Ambas as limitações são tratadas nas Ameaças à Validade e foram consideradas na interpretação dos resultados.

5.2 Variáveis e Métricas

A variável independente do estudo foi a técnica utilizada para construção do backlog, com dois níveis: *User Story Mapping* (USM) e *Product Backlog Building* (PBB). As variáveis dependentes foram a qualidade dos artefatos produzidos, a percepção dos estudantes sobre as técnicas e a carga cognitiva percebida.

A qualidade dos artefatos foi avaliada com base em cinco critérios: completude, clareza, estruturação, priorização e coerência do fluxo. A definição operacional desses critérios é apresentada na Tabela 2, na Subseção 5.4. Cada critério foi pontuado em uma escala ordinal de 1 a 5, na qual 1 representa desempenho muito baixo e 5, desempenho muito alto. Como métrica complementar, foi registrado o número de itens de backlog ou de histórias de usuário produzidos por grupo em cada sessão, utilizada apenas para caracterizar o porte dos artefatos. Como todas as sessões experimentais tiveram a mesma duração, o tempo foi controlado pelo desenho da intervenção e não foi utilizado como métrica comparativa entre técnicas.

A percepção dos estudantes foi avaliada por meio de um questionário pós-tarefa elaborado pelos pesquisadores, com itens em escala Likert de 5 pontos [12]. Os itens cobriram as seguintes dimensões: facilidade de uso, utilidade percebida, clareza das etapas, organização do backlog, compreensão do domínio, identificação de funcionalidades, satisfação com o resultado e percepção de controle

sobre a construção do produto. O instrumento foi aplicado individualmente ao final de cada sessão experimental, permitindo captar a percepção do estudante sobre a técnica que foi utilizada.

A carga cognitiva percebida foi avaliada por meio de uma versão reduzida do NASA-TLX, instrumento amplamente utilizado para mensurar carga de trabalho subjetiva [9]. Adotou-se o formato *Raw TLX* [8], no qual cada dimensão é pontuada diretamente, sem o procedimento de ponderação par a par entre as dimensões da versão original. A versão adotada considerou cinco dimensões, demanda mental, demanda temporal, esforço, frustração e desempenho percebido, aplicadas em escala Likert de cinco pontos para manter coerência com os demais instrumentos do estudo. A demanda física, presente na versão original, foi removida por não ser relevante para a tarefa cognitiva e colaborativa em questão.

5.3 Procedimento

A intervenção ocorreu em quatro encontros presenciais sucessivos, com duração de 01h40, distribuídos ao longo de duas semanas. Os artefatos produzidos estão disponíveis no repositório.

Encontro 1 — Apresentação e treinamento. Foram apresentados os objetivos da atividade, o termo de consentimento e as duas técnicas (USM e PBB), com aulas para a turma toda. A instrução incluiu slides com os elementos centrais de cada técnica, exemplos práticos e os *canvas* utilizados. O material das aulas foi disponibilizado previamente aos estudantes para leitura. Ao final, os estudantes responderam ao questionário de caracterização.

Encontro 2 — Piloto. Realizou-se um piloto com um domínio familiar, onde cada grupo aplicou ambas as técnicas (USM e PBB) para familiarização. Os artefatos foram coletados e os facilitadores forneceram *feedback* imediato para sanar dúvidas conceituais. A experiência foi fundamental para o refinamento do procedimento experimental: simplificamos a terminologia de formulários para mitigar ambiguidades, otimizamos o cronograma (reduzindo a explicação inicial em cinco minutos para maximizar o tempo de prática) e reforçamos as instruções das etapas que geraram maior confusão. Estas adaptações foram cruciais para garantir a clareza das instruções e a consistência da coleta de dados no experimento principal.

Encontros 3 e 4 — Sessões experimentais. Ambas sessões seguiram a mesma estrutura temporal: 20 minutos para leitura do Termo de Referência, 60 minutos para construção do backlog via técnica, 20 minutos para criação das histórias de usuário e fechamento. Foram materiais físicos (cartolinas, post-its) e templates de apoio. Facilitadores auxiliaram nos 30 minutos iniciais, reservando o período para trabalho autônomo. No Encontro 3 (Sessão 1), todos aplicaram PBB (Conjunto A com Marajó; Conjunto B com Viagens); no Encontro 4 (Sessão 2), aplicou-se USM com domínios invertidos. Após cada sessão, os artefatos, estudantes responderam ao questionário e o ciclo encerrou-se com uma discussão sobre as técnicas.

5.4 Avaliação dos Artefatos

Os artefatos (mapas de histórias, quadros PBB, funcionalidades, itens de backlog e histórias de usuário) foram avaliados quanto à forma de organizar informações em uma estrutura de produto compreensível, priorizável e útil para o desenvolvimento incremental.

A avaliação foi conduzida por três especialistas com perfis complementares e sem envolvimento na intervenção: R1 (acadêmico,

Table 2: Rubrica para avaliação dos backlogs e histórias de usuário.

Critério	Definição	Orientação para atribuição do escore
Compleitude	Avalia em que medida o artefato atende às principais necessidades, atores, funcionalidades, restrições e requisitos relevantes do Termo de Referência.	Escore 1: indica cobertura muito limitada, com omissão de funcionalidades centrais. Escore 3: indica cobertura parcial dos principais elementos do domínio. Escore 5: indica cobertura ampla e consistente das funcionalidades, dos atores, das restrições e das necessidades relevantes.
Clareza	Avalia a compreensibilidade dos itens de backlog e, quando existentes, das histórias de usuário, considerando a redação, a ausência de ambiguidade, a identificação de papéis, objetivos, benefícios e critérios de aceitação.	Escore 1: indica itens vagos, ambíguos ou difíceis de interpretar. Escore 3: indica itens compreensíveis, mas com inconsistências, lacunas ou histórias pouco orientadas a valor. Escore 5: indica itens claros, bem formulados, com valor explícito e, quando aplicável, histórias de usuário adequadas à estrutura “Como [persona], quero [ação], para [benefício]”.
Estruturação	Avalia a organização do artefato em níveis coerentes, como personas, atividades, funcionalidades, épicos, temas, tarefas, histórias ou itens de backlog, bem como a granularidade dos itens produzidos.	Escore 1: indica organização frágil ou apenas uma lista desordenada de itens. Escore 3: indica alguma estrutura, mas apresenta problemas de agrupamento, de granularidade ou de dependências. Escore 5: indica organização consistente, hierárquica ou visualmente compreensível, alinhada à técnica utilizada e com itens em nível adequado para evolução do backlog.
Priorização	Avalia se o artefato apresenta critérios ou lógica de priorização, considerando valor, urgência, dependências, riscos, MVP ou incrementos de entrega.	Escore 1: indica ausência de priorização ou ordenação arbitrária. Escore 3: indica priorização parcial, mas pouco justificada. Escore 5: indica priorização coerente, justificada e útil para orientar o planejamento incremental.
Coerência do fluxo	Avalia se o conjunto de itens representa uma narrativa ou uma sequência lógica de uso, especialmente em relação às jornadas, às interações entre atores, às dependências entre funcionalidades e à evolução do produto.	Escore 1: indica itens desconectados, sem relação clara entre si. Escore 3: indica um fluxo parcialmente compreensível, mas com lacunas ou inconsistências. Escore 5: indica fluxo claro, consistente e alinhado às jornadas, processos ou capacidades principais do produto.

experiência em Engenharia de Software), R2 (*Product Owner*, experiência em gestão de backlog) e R3 (técnico, desenvolvedor e analista de produto). A ausência de vínculo dos avaliadores com a facilitação ou com os estudantes mitigou riscos de viés interpessoal.

Os critérios da rubrica baseiam-se em princípios de qualidade de requisitos e agilidade. Compleitude, clareza e estruturação estão associadas a requisitos [1], enquanto a qualidade das histórias dialoga com critérios de valor, granularidade e testabilidade [13]. A priorização foi incluída por ser central na orientação de decisões e no planejamento incremental [10, 19].

Os artefatos foram digitalizados e disponibilizados aos avaliadores em formato digital, e foram analisados após uma reunião de calibração, na qual os avaliadores aplicaram a rubrica aos itens do piloto para alinhar interpretações. Após isso, cada avaliador analisou os artefatos do experimento individualmente, sem acesso aos escores atribuídos pelos demais avaliadores. Os avaliadores tinham conhecimento, durante a pontuação, da identificação do grupo (sem nomes dos estudantes), da técnica aplicada (PBB ou USM) e do Termo de Referência utilizado. Optou-se por essa avaliação aberta, e não cega, por duas razões: (i) os artefatos PBB e USM são visualmente reconhecíveis pela disposição característica de cada técnica, o que tornaria inviável ocultar essa informação dos avaliadores; e (ii) a avaliação de *compleitude* requer comparar o artefato com o conteúdo do TR, o que exige saber qual TR está sendo avaliado. Para mitigar vieses, adotaram-se rubricas com âncoras comportamentais, calibração prévia e verificação estatística de concordância.

Os escores finais foram calculados pela média das notas dos três avaliadores e foram disponibilizados em uma planilha. Divergências superiores a dois pontos foram resolvidas via consenso, apoiado em comentários qualitativos. A confiabilidade entre avaliadores foi analisada por meio da Correlação Intraclass (ICC), calculada no R com o pacote *irr*. Como os escores finais foram consolidados pela média dos avaliadores, reportou-se o ICC para essa média, com acordo absoluto [11]. A confiabilidade geral foi excelente, com ICC = 0,978 para a média geral dos escores. Por critério, os valores foram 0,972 para compleitude, 0,867 para clareza, 0,897 para estruturação,

1,000 para priorização e 0,977 para coerência do fluxo, garantindo alta consistência na interpretação dos resultados.

5.5 Análise dos Dados

A análise dos dados foi organizada em etapas quantitativa e qualitativa. Na quantitativa, estatísticas descritivas (média, mediana, desvio padrão e intervalo interquartil) foram calculadas para escores dos artefatos, questionários, carga cognitiva e produtividade. Para a análise dos artefatos, as notas dos três avaliadores foram consolidadas pela média. Como cada grupo produziu um artefato com PBB e outro com USM, a comparação foi realizada por pareamento entre grupos ($n = 6$). Dada a natureza ordinal da escala, utilizou-se o teste de Wilcoxon para amostras pareadas, reportando o valor de p e o tamanho de efeito por correlação bisserial (r_{rb}) para apoiar a interpretação da magnitude das diferenças. A produtividade foi medida pela quantidade de itens de backlog (histórias de usuário ou tarefas) gerados pelos grupos durante o tempo estipulado em cada sessão.

As respostas aos questionários pós-tarefa foram analisadas por estatísticas descritivas e, quando possível, comparadas via pareamento entre as técnicas (USM e PBB) para interpretar percepções de utilidade e carga cognitiva. Devido ao tamanho da amostra e ao caráter educacional do estudo, os testes estatísticos serviram como apoio interpretativo, e não como base para inferências causais.

A etapa qualitativa fundamentou-se na Análise Temática de Braun e Clarke [4]. O processo seguiu um ciclo iterativo de familiarização, codificação indutiva, categorização, revisão e síntese final. Essa abordagem permitiu identificar padrões sistemáticos sobre vantagens, dificuldades e adequação das técnicas, assegurando o rigor na interpretação e permitindo a triangulação das percepções qualitativas com os resultados quantitativos observados.

6 Resultados

6.1 Perfil dos Participantes e Fluxo de Respostas

O fluxo de participantes ao longo da intervenção está detalhado na Tabela 3. Dos 30 estudantes matriculados, 26 consentiram com a

pesquisa e realizaram o questionário de pré-teste de caracterização. Após as sessões experimentais, obteve-se 36 respostas individuais ao questionário pós-tarefa: 19 referentes ao PBB (Sessão 1) e 17 referentes ao USM (Sessão 2). O cruzamento dos dados anonimizados identificou 21 participantes distintos com ao menos uma resposta pós-tarefa, sendo que 15 responderam após ambas as técnicas, permitindo análises pareadas em nível individual. As variações decorrem de faltas pontuais nas sessões experimentais e do critério de exclusão definido na Seção 4. Em nível de grupo, todos os seis grupos produziram artefatos com PBB e com USM, garantindo cobertura completa do desenho experimental.

Table 3: Conciliação dos quantitativos de participantes.

Etapa	n
Estudantes matriculados na disciplina	30
Consentiram em participar (TCLE)	26
Responderam ao pré-teste de caracterização	26
Respostas ao pós-tarefa após PBB (Sessão 1)	19
Respostas ao pós-tarefa após USM (Sessão 2)	17
Total de respostas pós-tarefa	36
Indivíduos distintos com pelo menos um pós-tarefa	21
responderam após ambas as técnicas (pareados)	15
responderam apenas após PBB	4
responderam apenas após USM	2

Table 4: Práticas previamente utilizadas pelos estudantes ($n = 26$).

Prática	n	%
Casos de uso	17	65,4
Entrevistas com stakeholders	17	65,4
User Stories	15	57,7
Prototipagem	12	46,2
Definição de Product Backlog	8	30,8
Dailys	3	11,5
Workshop de requisitos	3	11,5
Sprints	2	7,7
Reuniões Scrum	1	3,8
Teste	1	3,8

As experiências prévias dos estudantes ($n = 26$) constam na Tabela 4. As mais frequentes foram casos de uso e entrevistas com *stakeholders*, ambas mencionadas por 17 estudantes (65,4%), seguidas por User Stories, indicada por 15 estudantes (57,7%), e prototipagem, por 12 (46,2%). A definição de Product Backlog apareceu em 8 respostas (30,8%). Práticas associadas à execução contínua de métodos ágeis, tiveram menor ocorrência, com valores entre 1 e 3 respostas. Esse perfil indica uma turma familiarizada com o levantamento e representação de requisitos, mas com exposição limitada à execução de processos ágeis estruturados.

A autopercepção inicial sobre USM e PBB (Tabela 5) revelou médias de conhecimento próximas para ambas as técnicas (2,88 e 2,85, respectivamente). Contudo, as médias de habilidade de aplicação foram menores: 2,62 para USM e 2,42 para PBB. Com as medianas situadas abaixo do ponto central da escala nos itens de habilidade, evidencia-se que a maioria dos estudantes não se sentia apta a aplicar as técnicas antes da intervenção, o que reforça a relevância do treinamento e do piloto descritos na Seção 5.3.

Table 5: Autopercepção prévia sobre USM e PBB ($n = 26$, escala 1–5).

Item	Média	Mediana
Conhecimento em USM	2,88	3
Habilidade em aplicar USM	2,62	2
Conhecimento em PBB	2,85	3
Habilidade em aplicar PBB	2,42	2

6.2 RQ1: Qualidade dos Backlogs Produzidos

A RQ1 investigou se houve diferença na qualidade dos backlogs produzidos com USM e PBB. A análise baseou-se nos escores médios dos três avaliadores para cada um dos 12 artefatos gerados (seis grupos $\times$ duas técnicas). Como cada grupo aplicou ambos os métodos, a comparação foi tratada como pareada. Dado o tamanho da amostra ($n = 6$) e a natureza ordinal da escala, utilizou-se o teste de Wilcoxon para amostras pareadas, complementado pelo tamanho de efeito via correlação bisserial por postos (r_{rb}), servindo como apoio exploratório à interpretação, discutido na Seção 5.5.

A Tabela 6 detalha a comparação por critério. Em termos descritivos, o PBB obteve média geral de 3,00, enquanto o USM obteve média geral de 2,66. Esse padrão foi consistente: em cinco dos seis grupos o PBB obteve média geral superior ao USM, com diferenças entre +0,20 e +0,80 pontos. Apenas o Grupo 1 apresentou uma leve vantagem do USM (−0,13), sugerindo uma tendência descritiva favorável ao PBB, consistente na maior parte da amostra.

A disparidade mais acentuada ocorreu na **estruturação**, com média 3,61 para PBB e 3,06 para USM. O tamanho de efeito foi alto ($r_{rb} = 1,00$), indicando que, em todos os pares onde houve diferença, o PBB apresentou escores mais altos. Esse resultado alinha-se à natureza do método, cuja sequência de descoberta (visão, problemas, personas, funcionalidades, itens de backlog) parece facilitar a transposição de requisitos brutos para uma estrutura mais organizada.

Quanto à **clareza**, o PBB (3,39) superou o USM (3,11). Os avaliadores observaram que o PBB mostra melhor a relação entre problemas e funcionalidades. Já no USM, apesar dos fluxos serem claros, houve maior uso de papéis genéricos e de itens escritos como tarefas técnicas, em vez de histórias orientadas a valor, refletindo seu foco na narrativa de uso em vez da estruturação dos elementos.

A **priorização** foi o critério de menor desempenho em ambas as técnicas, com PBB (1,67) levemente acima de USM (1,00). Esse resultado indica que a priorização permaneceu uma dificuldade transversal, mesmo quando os grupos identificaram funcionalidades, personas e fluxos. A diferença observada sugere que o PBB ofereceu algum apoio adicional à organização dos itens, possivelmente pela presença explícita de etapas que demandam ordenação, mas isso não foi suficiente para garantir uma priorização consistente. Esse achado é discutido como lição pedagógica na Seção 7.

Em **completude**, as diferenças foram pequenas (PBB 3,22; USM 3,17), indicando que ambas as técnicas contribuíram para a identificação de parte relevante dos elementos presentes nos Termos de Referência. A análise qualitativa dos artefatos, no entanto, indicou que vários grupos deixaram de contemplar requisitos transversais, como autenticação por perfis, integração com sistemas externos,

Table 6: Comparação pareada entre PBB e USM por critério de qualidade ($n = 6$ grupos).

Critério	Média PBB	Mediana PBB	Média USM	Mediana USM	p Wilcoxon	r_{rb}
Compleitude	3,22	3,17	3,17	3,00	1,0000	0,17
Clareza	3,39	3,34	3,11	3,00	0,2500	1,00
Estruturação	3,61	4,00	3,06	2,84	0,0625	1,00
Priorização	1,67	2,00	1,00	1,00	0,1250	1,00
Coerência do fluxo	3,06	3,00	2,83	3,00	0,6250	0,40
Média geral	3,00	3,13	2,66	2,57	0,0625	0,90

requisitos não funcionais, relatórios específicos, acessibilidade, segurança e suporte a baixa conectividade. Essa lacuna é especialmente relevante no domínio Marajó Resiliente, no qual requisitos transversais têm peso significativo no Termo de Referência.

Em **coerência do fluxo**, a diferença foi moderada (PBB 3,06; USM 2,83). Embora estruturado por jornadas de uso, o USM aproximou-se do PBB no domínio mais processual, apresentada na Subseção 6.5, mas teve mais dificuldade no domínio exploratório, onde a definição de uma jornada única é menos evidente.

A resposta à RQ1 aponta para uma diferença descritiva favorável ao PBB na qualidade geral dos backlogs produzidos, especialmente em estruturação e clareza. O teste de Wilcoxon, utilizado como apoio exploratório, não indicou diferença estatisticamente significativa ($p = 0,0625$ na média geral; $r_{rb} = 0,90$). O PBB apresentou indícios de apoiar melhor a decomposição conceitual e a organização inicial, enquanto a baixa pontuação em priorização nas duas técnicas aponta uma necessidade de atenção pedagógica explícita.

6.3 RQ2: Percepção sobre Facilidade de Uso e Utilidade

A RQ2 investigou como os estudantes perceberam a facilidade de uso e a utilidade de USM e PBB para a elaboração de histórias de usuário e organização do backlog. A análise seguiu duas perspectivas complementares: (i) médias por técnica, considerando todas as respostas pós-tarefa coletadas, refletindo a percepção média de cada técnica; e (ii) análise pareada, considerando apenas os 15 estudantes que responderam às duas sessões, isolando variações individuais.

A Figura 1 compara as médias por técnica, considerando todas as 36 respostas pós-tarefa (19 PBB e 17 USM). Observa-se uma preferência declarada pelo USM na maioria dos itens. As disparidades mais acentuadas ocorreram em **clareza das etapas** (PBB 2,74; USM 3,88), **tempo suficiente** (PBB 2,00; USM 2,82) e **pouco esforço de uso** (PBB 2,68; USM 3,41). O USM também superou o PBB em itens como organização lógica, identificação de funcionalidades, definição de MVP e visão sistêmica *Big Picture*. Relatos qualitativos reforçam que o USM auxiliou na estruturação do projeto e na distinção dos objetivos centrais, tendo o suporte de materiais didáticos e exemplos práticos como facilitadores.

Em contrapartida, a avaliação menos favorável do PBB associa-se à curva de aprendizado da técnica e à pressão de tempo. Os dados qualitativos indicam que o método exigiu mais tempo para a leitura, a organização física dos itens e o preenchimento das estruturas do *canvas*. A falta de familiaridade gerou frustração em algumas equipes que não finalizaram os artefatos integralmente. Houve também críticas à ambiguidade dos requisitos nos documentos-base,

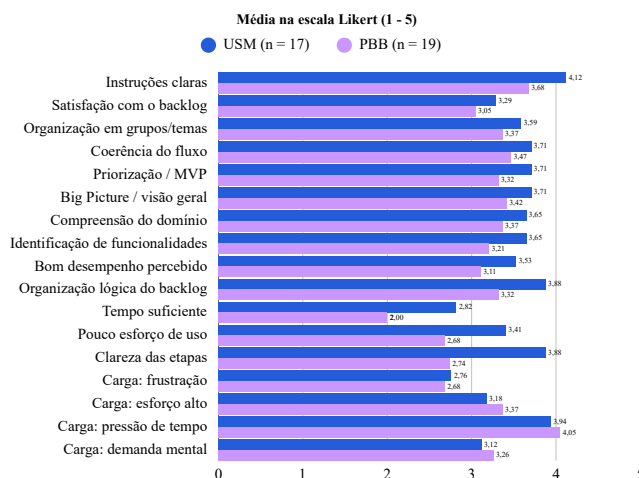


Figure 1: Médias dos itens de percepção e carga cognitiva por técnica, considerando todas as respostas pós-tarefa ($n = 19$ PBB; $n = 17$ USM).

que, somada à estrutura mais rígida do PBB, dificultou a identificação de funcionalidades e experiências.

A Figura 2 detalha as diferenças pareadas (USM – PBB) entre os 15 estudantes que experienciaram ambas as técnicas. As maiores vantagens do USM ocorreram em *clareza das etapas* (+1,00), *pouco esforço de uso* (+0,73), *tempo suficiente* (+0,67), *organização lógica do backlog* (+0,47) e *bom desempenho percebido* (+0,47). Em contraste, dimensões como compreensão do domínio, identificação de funcionalidades, *Big Picture* e priorização apresentaram diferenças próximas de zero, sugerindo que ambas as técnicas foram percebidas como equivalentes nesses aspectos conceituais. A satisfação com o backlog apresentou diferença próxima de zero (−0,07), reforçando que, mesmo percebendo o USM como mais fácil, os estudantes não se mostraram mais satisfeitos com o produto final.

A análise por domínio também revela nuances. No projeto Marajó Resiliente, o USM enfrentou críticas por uma documentação percebida como “confusa”, dificultando a criação de personas e jornadas. No PBB, apesar da limitação de tempo, os alunos reconheceram que a técnica ajudou a “entender o objetivo da plataforma” e a organizar as necessidades. Esse contraste é coerente com os achados sobre a qualidade dos artefatos discutidos na Subseção 6.5.

Desse modo, a resposta à RQ2 indica que a percepção de “pouco esforço” associada ao USM deve ser interpretada como relativa ao

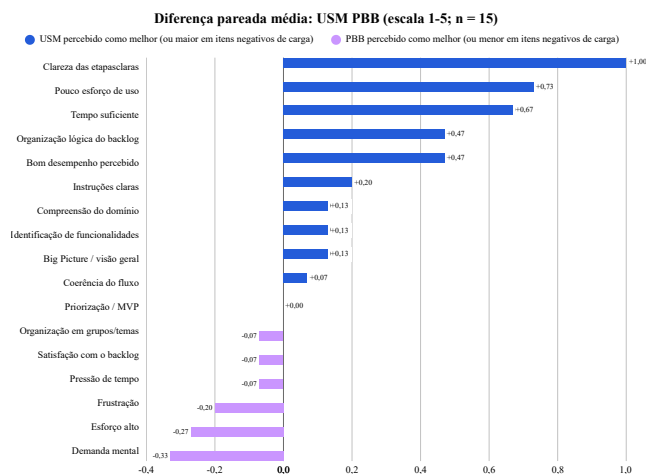


Figure 2: Diferença pareada (USM – PBB) por item, (n = 15). Valores negativos favorecem o PBB.

PBB. Em termos absolutos, os relatos qualitativos indicam que ambas as técnicas exigiram esforço alto. Essa divergência reflete a distinção entre a facilidade na estruturação do método (onde o USM se destacou) e o esforço operacional elevado (trabalho manual e pressão de tempo), que foi comum a ambos. Assim, o USM foi percebido como menos custoso por oferecer um suporte que facilitou a articulação dos requisitos, e não por ser uma tarefa simples.

6.4 RQ3: Carga Cognitiva Percebida

A RQ3 investigou se houve diferença na carga cognitiva percebida pelos estudantes ao utilizarem o USM e o PBB. A análise considerou os itens do questionário associados a demanda mental, pressão de tempo, esforço, desempenho percebido e frustração, conforme a versão *Raw TLX* adaptada para este estudo. As diferenças entre as técnicas foram menores nos itens de carga cognitiva do que nos itens de percepção de uso, sugerindo que a carga cognitiva foi influenciada não apenas pela técnica, mas também pelas características da tarefa e pelo tempo disponível.

Considerando todas as respostas pós-tarefa, a alta demanda mental apresentou média 3,26 para PBB e 3,12 para USM, indicando que ambas as técnicas foram percebidas como cognitivamente exigentes. A pressão de tempo apresentou médias elevadas em ambas as técnicas (PBB 4,05; USM 3,94), o que sugere que foi uma dificuldade comum. Esse resultado é coerente com a natureza da tarefa, que exigiu interpretar documentos extensos, discutir em grupo, identificar atores ou personas, decompor funcionalidades e produzir artefatos de backlog em um tempo limitado.

Na análise pareada com os 15 estudantes que responderam ambas as sessões, observa-se diferença em favor do USM em *pouco esforço de uso* (+0,73) e em *bom desempenho percebido* (+0,47), mostrando que o USM era menos custoso de aplicar e tiveram maior sensação de desempenho. Por outro lado, as diferenças em demanda mental (-0,33, ligeiramente menor para o USM) e pressão de tempo (-0,07)

foram menos expressivas. Frustração e esforço também foram percebidos como ligeiramente menores no USM. No conjunto, esses resultados indicam que, embora o USM tenha sido percebido como menos custoso, ambas as técnicas impuseram uma carga cognitiva relevante em um contexto de tempo limitado.

A análise qualitativa dos relatos reforça que a carga cognitiva não decorreu apenas das técnicas, mas também da combinação entre o tempo limitado e a complexidade intrínseca dos Termos de Referência. Os relatos descrevem o tempo como curto ou limitado, com afirmações de que “a falta de tempo e calma me limitou” e que o período foi insuficiente para o “trabalho manual” de alocação física dos itens. Em alguns casos, os relatos indicam que a pressão temporal dificultou a conclusão do artefato.

Quanto à demanda mental, os dados qualitativos reforçam que a “falta de clareza” e as “funcionalidades ambíguas” exigiram alto esforço de interpretação. No domínio Marajó Resiliente, essa carga foi acentuada pela natureza “confusa” percebida no material, tornando a criação de personas “extremamente difícil”. A diferença em *pouco esforço de uso* a favor do USM é compatível com relatos que destacam sua simplicidade de montagem. Em contrapartida, o PBB parece ter imposto uma carga de aprendizado maior, com estudantes relatando frustração por “não saber como funciona” ou pela necessidade de seguir o método de forma mecânica (“passo a passo”), por falta de prática. Limitações físicas, como a “falta de espaço” nas cartolinas e o tempo gasto para “ler e alocar dentro das estruturas”, elevaram o esforço percebido no PBB.

Em síntese, a resposta à RQ3 não indica diferença significativa na carga cognitiva entre USM e PBB. A demanda mental foi ligeiramente menor para o USM, e a pressão de tempo foi elevada em ambas as técnicas. Os estudantes perceberam menor esforço de uso no USM, sugerindo que o PBB exigiu maior articulação conceitual durante a atividade. Essa exigência ajuda a explicar o maior esforço percebido pelos estudantes e, ao mesmo tempo, é compatível com a melhor estruturação dos artefatos PBB observada na RQ1.

6.5 Comparação entre Domínios

Para investigar como a natureza do domínio interage com a técnica utilizada, os escores foram decompostos pelos quatro pares possíveis: técnica × domínio. A Tabela 7 consolida as médias por critério, onde cada célula representa o desempenho de três grupos sob a avaliação de três especialistas.

Três padrões merecem atenção. Primeiro, o **PBB foi descritivamente superior ao USM em ambos os domínios**, embora sua vantagem tenha sido mais expressiva em *Gestão de Viagens* (+0,47) do que em *Marajó Resiliente* (+0,20). Esse resultado indica que o PBB se beneficia de domínios processuais, onde a clareza de atores e fluxos no Termo de Referência (TR) potencializa a decomposição estruturada do *canvas*. Segundo, o **USM apresentou desempenho praticamente estável entre os dois domínios** (2,64 em Marajó; 2,69 em Viagens). Essa estabilidade contradiz a expectativa de que o USM teria um desempenho superior em domínios com fluxo operacional predominante. Em coerência do fluxo, especificamente, USM × Viagens (3,33) empatou com PBB × Viagens (3,33), indicando que o USM conseguiu representar bem o fluxo, mas não obteve vantagem clara sobre o PBB nesse aspecto. Terceiro, o **domínio Marajó foi mais desafiador para ambas as técnicas**, o que se

Table 7: Médias dos escores dos avaliadores por técnica × domínio (3 grupos por célula).

Técnica	Domínio	Compleitude	Clareza	Estruturação	Priorização	Coerência	Média geral
PBB	Marajó Resiliente	2,67	3,45	3,56	1,67	2,78	2,84
PBB	Gestão de Viagens	3,78	3,33	3,67	1,67	3,33	3,16
USM	Marajó Resiliente	3,33	3,33	3,22	1,00	2,33	2,64
USM	Gestão de Viagens	3,00	2,89	2,89	1,00	3,33	2,69

manifestou principalmente em coerência do fluxo (PBB 2,78; USM 2,33). Os grupos que aplicaram USM em Marajó precisaram inferir múltiplas jornadas paralelas (coleta de dados, indicadores, gestão de conteúdos, repositórios) o que fragmentou a narrativa dos mapas e dificultou a organização conceitual.

A leitura conjunta da Tabela 7 sugere uma interação entre técnica e domínio: **o PBB ampliou sua vantagem em domínios mais processuais, enquanto o USM mostrou-se relativamente robusto, embora não ultrapassasse o PBB**. Em particular, em Gestão de Viagens o PBB destacou-se em completude (3,78), evidenciando que suas etapas de descoberta favorecem a captura sistemática de requisitos. Em Marajó Resiliente, o USM teve completude superior à do PBB (3,33 vs. 2,67). Essa incompletude do PBB evidencia uma limitação da técnica: sua rigidez hierárquica atua como um funil restritivo em problemas exploratórios, penalizando a captura de requisitos quando o domínio carece de fluxos processuais definidos. Em contraste, a natureza horizontal do USM acomoda melhor requisitos transversais em contextos de incerteza.

Esses achados reforçam que a escolha entre PBB e USM deve considerar a estruturação prévia do problema. Quando o documento de origem apresenta atores e fluxos explícitos (domínio processual), as restrições do canvas do PBB funcionam como um andaime cognitivo eficiente, guiando o aluno a categorizar informações já disponíveis e elevando a completude do backlog. Por outro lado, quando o problema é mal estruturado ou exploratório, a rigidez do PBB pode se tornar um obstáculo de abstração, enquanto a flexibilidade narrativa do USM permite conexões mais fluidas. Pedagogicamente, isso indica que a introdução ao levantamento ágil de requisitos deve ser intencional: problemas vagos demandam abordagens focadas em jornada (USM) para ancorar o entendimento, enquanto documentos delineados se beneficiam de abordagens de decomposição (PBB) para refinar a estrutura.

7 Lições Aprendidas

Os resultados deste estudo evidenciam um contraste pedagógica-mente relevante: o PBB produziu artefatos tecnicamente superiores em estruturação e clareza, enquanto o USM foi avaliado pelos estudantes como mais claro e mais fácil de aplicar. Esse contraste, somado à fragilidade na priorização e à influência do domínio, sustenta cinco lições aprendidas. Cada lição articula o achado, uma sugestão prática e uma armadilha a evitar em futuras replicações.

Lição 1. Percepção de facilidade não é sinônimo de qualidade do artefato. A técnica percebida como mais fácil pelos estudantes (USM) não foi a que produziu artefatos mais bem avaliados (PBB). *Sugestão:* pode-se considerar as duas dimensões — experiência de aprendizagem e qualidade do produto — ao escolher técnicas para uma intervenção, e apresentar esse contraste aos estudantes

como objeto de reflexão. *Armadilha:* basear a escolha apenas no *feedback* subjetivo dos estudantes pode privilegiar técnicas mais palatáveis em detrimento das que efetivamente desenvolvem competências de organização e estruturação.

Lição 2. Documentos realistas são insumos formativos, mas exigem mediação pedagógica explícita. Os Termos de Referência reais aproximaram os estudantes da prática profissional, mas impuseram um esforço de interpretação significativo, refletido em relatos sobre “ambiguidade” e “falta de clareza no documento”. A leitura crítica de TRs mostrou-se uma competência distinta da aplicação das técnicas. *Sugestão:* pode-se reservar tempo prévio à leitura do documento, com atividades que exercitem a identificação de atores, a separação entre necessidades e funcionalidades, e a detecção de requisitos transversais. *Armadilha:* sem essa mediação, estudantes podem confundir dificuldades do TR com dificuldades da técnica, prejudicando a avaliação e a aprendizagem.

Lição 3. PBB e USM oferecem apoios pedagógicos distintos e complementares. O PBB favoreceu a decomposição conceitual a partir de problemas, expectativas e personas, traduzindo-se em maior estruturação. O USM mostrou-se mais acessível, mas com desempenho relativamente estável entre os domínios, enquanto o PBB se beneficiou particularmente do domínio mais processual. *Sugestão:* pode-se apresentar as duas técnicas explicitamente como complementares — PBB quando o desafio é construir entendimento conceitual; USM quando é organizar uma jornada já compreendida em incrementos. *Armadilha:* tratar uma como substituta da outra perde o aprendizado pedagógico mais rico, que é entender quando cada uma agrega mais valor.

Lição 4. A priorização deve ser tratada como um objeto pedagógico próprio. Este critério obteve o pior desempenho em ambos os métodos (1,67 para PBB e 1,00 para USM), mesmo quando os grupos identificaram funcionalidades, personas e fluxos com qualidade razoável. *Sugestão:* pode-se introduzir atividades específicas — definição de MVP, classificação valor versus esforço, identificação de dependências, justificativa de *releases* — antes ou em paralelo à construção do backlog. *Armadilha:* assumir que a priorização emergirá “naturalmente” da aplicação de uma técnica.

Lição 5. PBB e USM podem ser ensinados em sequência. A natureza das técnicas sugere uma sequência viável: iniciar pelo PBB e evoluir para o USM, para organizar jornadas e fatias incrementais sobre o mesmo domínio. *Sugestão:* quando o tempo da disciplina permitir, pode-se planejar a sequência integrada e propor que os estudantes reflitam sobre como cada técnica contribuiu para um aspecto distinto do produto. *Armadilha:* aplicar as duas técnicas em encontros curtos, gera frustração e prejudica especialmente a técnica que demanda mais articulação conceitual.

8 Ameaças à Validade

Como em qualquer estudo empírico em ambiente educacional, os resultados devem ser interpretados considerando as ameaças à validade [23].

Validade interna. Por restrições logísticas, todos os grupos aplicaram PBB na primeira sessão e USM na segunda, sem contrabalanceamento de ordem, o que introduz possível efeito de aprendizado ou cansaço. Vale notar que o PBB foi descritivamente superior mesmo aplicado primeiro, o que fortalece em parte o achado sem eliminar a limitação. Para mitigar a ameaça da experiência prévia heterogênea, aplicamos um questionário de caracterização e formamos grupos balanceados. O tempo limitado de 1h40 por sessão pode ter afetado a profundidade da análise dos TRs e foi parcialmente mitigado pelo treinamento prévio e pelo encontro de piloto.

Validade de construção. Os cinco critérios da rubrica não capturam todos os aspectos possíveis de qualidade (por exemplo, rastreabilidade detalhada e validação de valor com *stakeholders* reais). A avaliação dos artefatos não foi cega: os avaliadores conheciam grupo, técnica e Termo de Referência. Optamos por avaliação aberta porque os artefatos PBB e USM são visualmente reconhecíveis e a avaliação de completude exige conhecer o TR. Mitigamos esse risco com rubrica explícita, calibração prévia entre avaliadores sobre artefatos do piloto, independência da equipe que conduziu a intervenção e ICC reportado (0,978). Os instrumentos de percepção foram elaborados pelos pesquisadores, sem validação psicométrica formal; a triangulação com dados qualitativos atenua, em parte, essa limitação.

Validade externa. O estudo foi conduzido em uma única turma de Bacharelado em Sistemas de Informação, com 26 participantes consentidos e dois domínios, o que limita a generalização. Outros perfis de estudantes, contextos institucionais ou domínios podem produzir resultados diferentes, assim como decisões pedagógicas específicas (canvas físico, duração das sessões, nivelamento prévio).

Validade de conclusão. O número reduzido de grupos ($n = 6$) e a natureza ordinal da escala fazem com que os testes estatísticos tenham baixo poder; por isso, foram tratados como apoio exploratório, complementados por tamanhos de efeito, análise por critério e por domínio, e observações qualitativas. Outra ameaça é a diferença de unidade de análise: artefatos coletivos versus questionários individuais. Mantivemos análises separadas em cada nível, mas a percepção individual pode não corresponder diretamente à qualidade do artefato coletivo.

9 Considerações Finais

Este artigo apresentou um relato de experiência, de desenho quase experimental, sobre o uso de USM e PBB no ensino de Engenharia de Software. A intervenção envolveu uma turma na qual 26 consentiram em participar, organizados em seis grupos. Os grupos aplicaram as duas técnicas à construção de backlogs e de histórias de usuário a partir de dois Termos de Referência: um sistema de Gestão de Viagens, mais processual, e a plataforma Marajó Resiliente, mais exploratória. O desenho contrabalanceou os domínios entre os conjuntos de grupos, permitindo análises pareadas tanto no nível do grupo (qualidade dos artefatos) quanto no nível do indivíduo (percepção dos estudantes).

Em resposta à RQ1, observamos uma tendência descritiva favorável ao PBB na qualidade geral dos artefatos, com vantagens mais expressivas nos critérios de estruturação e clareza, conforme detalhado na Seção 6.2. Em resposta à RQ2, os estudantes perceberam o USM como mais claro e mais fácil de aplicar, com maiores diferenças quanto à clareza das etapas, ao esforço de uso e ao tempo necessário. Em resposta à RQ3, não observamos diferença significativa na carga cognitiva entre as técnicas, sugerindo que a complexidade da tarefa e o tempo limitado afetaram ambas as técnicas. Adicionalmente, a comparação por domínio revelou que o PBB se beneficiou particularmente do domínio mais processual, enquanto o USM apresentou desempenho relativamente estável em ambos.

O contraste entre qualidade do artefato (favorável ao PBB) e percepção dos estudantes (favorável ao USM) emergiu como o achado mais marcante deste estudo, sugerindo que a técnica mais palatável não é necessariamente a mais formativa. A priorização foi o critério mais frágil em ambas as técnicas, o que indica uma lacuna pedagógica que não se resolve apenas pela escolha do método. Esses achados sustentam cinco lições aprendidas, articuladas na Seção 7, com recomendações práticas e armadilhas a evitar para docentes que ensinam requisitos ágeis.

Como contribuição, este trabalho oferece uma intervenção didática replicável baseada em documentos realistas, evidências comparativas entre PBB e USM em três dimensões (qualidade do artefato, percepção dos estudantes e carga cognitiva), e uma discussão sobre o PBB como abordagem pedagógica ainda pouco explorada na literatura de Educação em Engenharia de Software. Para apoiar futuras replicações, organizamos um pacote de artefatos contendo os materiais utilizados na intervenção, os instrumentos de coleta, a rubrica de avaliação, os Termos de Referência e exemplos dos artefatos produzidos.

Como trabalhos futuros, pretendemos aplicar a intervenção a outras turmas, reforçando a robustez da comparação. Também planejamos incluir novos domínios de aplicação, aprofundar a análise da qualidade das histórias de usuário (em particular, critérios de valor, granularidade e testabilidade), e investigar empiricamente o uso combinado de PBB e USM em uma sequência pedagógica integrada. Estudos futuros também podem examinar o impacto de instrução dedicada à priorização da qualidade dos backlogs produzidos, atendendo à lacuna mais persistente observada neste estudo.

Disponibilidade de Artefatos

Os materiais utilizados na intervenção, os instrumentos de coleta, a rubrica de avaliação, os Termos de Referência, os dados anonimizados e exemplos dos artefatos produzidos pelos estudantes estão disponíveis em: Artefatos.

Agradecimentos

O Grammarly e o Claude foram utilizados como ferramentas de apoio à revisão textual, especialmente para aprimorar a ortografia, a gramática, a clareza e a fluidez do texto. O ChatGPT foi utilizado como auxílio na geração e formatação de tabelas em LaTeX. Todo o conteúdo produzido com apoio dessas ferramentas foi cuidadosamente revisado e adaptado pelos autores, que assumem integral responsabilidade pelo conteúdo final do artigo.

REFERÊNCIAS

- [1] 2018. ISO/IEC/IEEE 29148:2018 Systems and software engineering – Life cycle processes – Requirements engineering.
- [2] Fábio Aguiar and Paulo Caroli. 2021. *Product Backlog Building: Um guia prático para criação e refinamento de backlog para produtos de sucesso*. Editora Caroli. 144 pages.
- [3] Brian T. Bennett and Tristan Onek. 2019. Categorizing User Stories in the Software Engineering Classroom. *Journal of Computing Sciences in Colleges* 35, 4 (2019), 52–59.
- [4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp0630a
- [5] Mike Cohn. 2004. *User Stories Applied: For Agile Software Development*. Addison-Wesley Professional, Boston, MA.
- [6] Kiev Gama and Higor Oliveira. 2022. An Experience Report on Teaching Scrum Principles in a Playful Way through Distant Collaboration with Online Whiteboards. In *Proceedings of the XXXVI Brazilian Symposium on Software Engineering (SBES '22)*. Association for Computing Machinery, New York, NY, USA, 143–152. doi:10.1145/3555228.3555256
- [7] Lina Garcés and Brauner R. N. Oliveira. 2024. Teaching Software Engineering with Project-Based Learning: A Four Years Experience Report. In *Anais do XXXVIII Simpósio Brasileiro de Engenharia de Software*. SBC, 476–486. doi:10.5753/sbes.2024.3530
- [8] Sandra G. Hart. 2006. NASA-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. 904–908. doi:10.1177/154193120605000909
- [9] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX: Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183. doi:10.1016/S0166-4115(08)62386-9
- [10] Irum Inayat, Siti Salwah Salim, Sabrina Marczyk, Maya Daneva, and Shahabodin Shamshirband. 2015. A systematic literature review on agile requirements engineering practices and challenges. *Computers in Human Behavior* 51 (2015), 915–929. doi:10.1016/j.chb.2014.10.046
- [11] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163. doi:10.1016/j.jcm.2016.02.012
- [12] Rensis Likert. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology* 22, 140 (1932), 1–55.
- [13] Garm Lucassen, Fabiano Dalpiaz, Jan Martijn E. M. van der Werf, and Sjaak Brinkkemper. 2016. Improving agile requirements: the Quality User Story framework and tool. *Requirements Engineering* 21, 3 (2016), 383–403. doi:10.1007/s00766-016-0250-x
- [14] Eloisa Toffano Seidel Masson, Angélica Toffano Seidel Calazans, Ian Nery Bandeira, Geovana Ramos Sousa Silva, and Edna Dias Canedo. 2023. Scrum in Practice: City Reconstruction as a Pedagogical Game Challenge. In *Proceedings of the XXII Brazilian Symposium on Software Quality (SBQS '23)*. Association for Computing Machinery, New York, NY, USA, 321–331. doi:10.1145/3629479.3629480
- [15] Jeff Patton. 2014. *User Story Mapping: Discover the Whole Story, Build the Right Product*. O'Reilly Media, Sebastopol, CA.
- [16] Cécile Péraire. 2023. Learning to Write User Stories with the 4C Model: Context, Card, Conversation, and Confirmation. In *2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG)*. IEEE, 33–36. doi:10.1109/SEENG59157.2023.00011
- [17] Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering* 14, 2 (2009), 131–164. doi:10.1007/s10664-008-9102-8
- [18] Mickael Fonseca Santos, Ricardo Angelo Filipe, and João Carlos Cunha. 2024. From Traditional to Agile Methodologies in Software Project Management Education: A Case Study. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers (ICETC '24)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3702163.3702460
- [19] Eva-Maria Schön, Jörg Thomaschewski, and María José Escalona. 2017. Agile Requirements Engineering: A systematic literature review. *Computer Standards & Interfaces* 49 (2017), 79–91. doi:10.1016/j.csi.2016.08.011
- [20] Ken Schwaber and Jeff Sutherland. 2020. The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game. <https://scrumguides.org/scrum-guide.html>. Acesso em: 24 abr. 2026.
- [21] SEBRAE/RO. 2014. *Termo de Referência: Contratação de empresa especializada para desenvolver software de gestão de viagens*. Termo de Referência 49-06/11/2014. Serviço de Apoio às Micro e Pequenas Empresas do Estado de Rondônia. Documento utilizado como objeto de estudo na atividade educacional.
- [22] Tor Sporssem, Torgeir Dingsøyr, and Klaas-Jan Stol. 2026. User stories as boundary objects in agile requirements engineering: A theoretical literature review. *Journal of Systems and Software* 233 (2026), 112693. doi:10.1016/j.jss.2025.112693
- [23] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-29044-2
- [24] World-Transforming Technologies. 2025. *Termo de Referência: Contratação de serviço para desenvolvimento, implementação e manutenção de uma plataforma digital para Gestão do Conhecimento*. Termo de Referência. World-Transforming Technologies. https://wtventures.net/wp-content/uploads/2025/08/TdR-_Servico_Des_Plataforma-GC-rev.pdf Termo de Referência do Projeto Marajó Resiliente, utilizado como objeto de estudo na atividade educacional. Acesso em: 29 abr. 2026.