

Medição Automatizada de QoE/UX com Playwright: Um Instrumento de Auditoria Passiva em Aplicações Web

Mariana de Oliveira Ribeiro¹, Vinícius Schineider Januário Viana¹,
Simone Azevedo Bandeira de Melo Aquino^{1,2}

¹Instituto Federal de Educação, Ciência e Tecnologia do Maranhão – Campus
Imperatriz (IFMA)

²Departamento de Ensino Superior e Tecnologia – Instituto Federal do
Maranhão, IFMA
CEP 65.906-335 – Imperatriz – MA – Brasil

{mariana.r, viniciusschneider}@acad.ifma.edu.br,
{simonebandeira}@ifma.edu.br

Abstract. *This paper proposes a Playwright-based passive auditing instrument for QoE/UX in web applications. A listen-only script captures dead clicks, long hesitations (≥ 10 s), and flow duration without intrusive instrumentation in production code, consolidating signals into a composite FrictionScore. In a controlled mixed-methods study ($N = 17$) on DevShare, a purpose-built web platform, logs were triangulated with UEQ-S scales and open-ended responses. 53% of participants reported missing visual feedback, consistent with recorded dead click events. Results demonstrate that cross-reading automated behavioral signals discriminates genuine friction from cognitive engagement, enabling diagnostics that isolated metrics fail to capture.*

Resumo. *Este artigo propõe um instrumento baseado em Playwright para auditoria passiva de QoE/UX em aplicações Web. Um script listen-only captura dead clicks, hesitações longas (≥ 10 s) e duração de fluxos sem instrumentação intrusiva no código de produção, consolidando os sinais em um FrictionScore composto. Em estudo controlado de métodos mistos ($N = 17$) na plataforma DevShare, logs foram triangulados com UEQ-S e respostas abertas. 53% dos participantes relataram ausência de feedback visual, em consonância com os dead clicks registrados. Os resultados demonstram que a leitura cruzada de sinais comportamentais automatizados discrimina fricção real de engajamento cognitivo, viabilizando diagnósticos que métricas isoladas não capturam.*

1. Introdução

A avaliação da Qualidade de Experiência (QoE) e da Usabilidade (UX) em aplicações Web tem sido, historicamente, conduzida por meio de métodos manuais e subjetivos. Os métodos atuais dependem majoritariamente de intervenções aplicadas a posteriori ou de práticas de *Quality Assurance* (QA) focadas na validação sintética de regressões funcionais [Leppänen 2024]. No entanto, a lacuna metodológica na automação de testes reside na distinção clássica entre sintaxe e semântica: algoritmos podem validar códigos sem possuir intencionalidade ou compreensão sobre a frustração emocional vivenciada durante a interação [Searle 1982, Rousi 2019].

Além do aspecto humano, a ineficiência no design de interfaces pode induzir comportamentos erráticos — como cliques repetidos (*dead clicks*) e hesitações prolongadas. Em aplicações Web, tais padrões podem estar associados a retrabalho (ações repetidas e fluxos reiniciados) e, potencialmente, a processamento e tráfego redundantes [Oyedeji et al. 2019, Rani et al. 2024].

Para endereçar este problema, este artigo propõe um instrumento (um arcabouço técnico de auditoria passiva) para medir QoE/UX de forma automatizada durante a interação. O trabalho altera a aplicação convencional do *Playwright*: em vez de executar rotinas sintéticas de clique, a ferramenta atua como observador passivo da jornada de usuários reais, capturando rastros de telemetria sem instrumentação intrusiva.

O presente artigo detalha a arquitetura desta proposta e apresenta uma prova de conceito (PoC) em estudo controlado na plataforma Web “DevShare”. O estudo investiga, de modo exploratório ($N = 17$), como sinais sintáticos capturados pela máquina (tempos e eventos) [Rousi 2019] se alinham à percepção subjetiva reportada via UEQ-S e a evidências qualitativas.

Em síntese, este trabalho contribui com (I) o detalhamento técnico de um instrumento de auditoria passiva para captura não intrusiva de sinais comportamentais associados a fricção; (II) um protocolo de estudo controlado de métodos mistos para triangulação dos sinais capturados; e (III) a validação empírica de uma PoC em uma aplicação Web realista, discutindo a alta variância comportamental como um atestado de validade ecológica.

Para tal, este artigo está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica e trabalhos relacionados; a Seção 3 descreve a arquitetura do instrumento e o desenho metodológico do estudo; a Seção 4 reporta os resultados e discute os achados a partir da triangulação quali-quantitativa; por fim, a Seção 5 apresenta conclusões, limitações e trabalhos futuros.

2. Fundamentação Teórica e Trabalhos Relacionados

Esta seção estabelece os conceitos necessários para interpretar a proposta e os resultados. Em particular, delimita-se o papel da máquina como coletora de *sintaxe* de interação e a necessidade de triangulação com dados humanos para inferência de *semântica* (percepção).

O instrumento proposto fundamenta-se na intersecção entre a filosofia da percepção e a automação de medição de QoE/UX. A experiência do usuário divide-se fundamentalmente em metas pragmáticas e hedônicas [Grigis and De Angeli 2024], cujas barreiras são sintetizadas na Tabela 1.

Tabela 1. Diretores e barreiras por meta de UX

Metas de UX	Diretores (Drivers)	Barreiras (Barriers)
Pragmáticas	Utilidade	Limitações técnicas, disponibilidade, perda econômica
Hedônicas	Surpresa, fascinação	Ameaça à identidade, política (filtros e tabs)

Ferramentas clássicas de QA atuam prioritariamente sobre barreiras pragmáticas. Conforme o experimento mental do “Quarto Chinês” [Searle 1982], um script de automação manipula a interface corretamente através de regras pré-definidas, mas carece de agência para interpretar a degradação da experiência. Sem o acompanhamento empírico humano, os logs automatizados tendem a atuar como “significantes vazios” [Rousi 2019]. A abordagem aqui proposta delega à máquina (Playwright) exclusivamente a captura da *sintaxe* das interações em tempo real (tempos e eventos), condicionando a interpretação *semântica* ao cruzamento com dados humanos (UEQ-S).

No que diz respeito à automação passiva de QoE/UX, estudos sustentam a viabilidade de afastar a dependência exclusiva da subjetividade humana em avaliações. No cenário nacional, Garcia et al. (2023) já apontavam a necessidade de estruturas para a automatização do diagnóstico de UX, destacando a urgência de métricas quantitativas. Internacionalmente, Keshvadi e Williamson (2020) empregaram a ferramenta MoVIE para capturar tráfego passivamente e inferir QoE, enquanto Banaouas (2019) evidenciou que 55% das heurísticas de usabilidade podem ser diagnosticadas via sequenciamento de eventos de GUI.

Como implicação secundária, estudos recentes documentam a interação entre usabilidade e aspectos de eficiência de recursos em aplicações Web [Oyedemi et al. 2019, Rani et al. 2024]. A premissa adotada é que a eliminação de atritos de interface pode reduzir ações repetitivas, que se traduziriam em processamento e tráfego redundantes. Neste contexto, rastrear falhas de usabilidade via ferramentas de escuta passiva — estruturalmente mais leves que automações baseadas em processamento de imagens, como o Sikuli [Savoldi and Oliveira 2018] — constitui uma tática de prevenção de ineficiência do produto.

Em suma, o presente trabalho diferencia-se da literatura existente ao propor a fusão objetiva entre as disciplinas de *Quality Assurance* e a avaliação empírica de UX. Adicionalmente, no contexto corporativo, o rastreamento comportamental depende frequentemente de telemetrias comerciais (como *Hotjar* ou *Google Analytics*). Contudo, essas soluções exigem a injeção de scripts pesados em produção, que degradam o desempenho da página, esbarram em legislações estritas de privacidade (LGPD/GDPR) e são mitigadas por *ad-blockers*. O diferencial deste instrumento reside na adoção de uma estratégia de *Shift-Left Testing*: ao instanciar o auditor passivo no pipeline de CI/CD (homologação), a ferramenta antecipa a descoberta de anomalias sintáticas sem poluir o código-fonte de produção e sem expor os dados reais do cliente final.

3. Metodologia e Arquitetura do Instrumento

Nesta seção, descreve-se a proposta (instrumento) e o desenho do estudo utilizado para avaliá-la como prova de conceito. A pesquisa caracteriza-se como um estudo metodológico e uma Prova de Conceito (PoC), operacionalizada por meio de um desenho quase-experimental controlado. O objetivo é investigar a associação empírica entre falhas de interface (sintaxe), percepção do usuário (semântica) e sinais objetivos de fricção observáveis durante a execução de tarefas.

A Figura 1 ilustra o framework metodológico do estudo, estruturado em três camadas complementares: (1) a construção do artefato controlado, (2) o protocolo de coleta com dissociação entre sintaxe e semântica, e (3) a análise e síntese dos dados.

3.1. O Playwright como Auditor Passivo no CI/CD

A principal característica da arquitetura proposta é a instanciação do *Playwright* em modo *listen-only* (escuta). Idealizado para ser acoplado ao pipeline de homologação (CI/CD), a ferramenta injeta um script ouvinte via `API page.addInitScript`, interceptando eventos customizados `qoe:step` no *front-end*. O sistema monitora três indicativos: (I) a duração do fluxo (tempo total de atividade para conclusão da tarefa); (II) *dead clicks* (cliques sem resposta perceptível do sistema, capturados via instrumentação de eventos DOM); e (III) hesitações (períodos de inatividade superiores a 10 segundos entre transições de página, limiar definido a partir da literatura de tempos de resposta aceitáveis em interfaces Web [Nielsen 1994]). Esta arquitetura viabiliza uma coleta não intrusiva, dispensando a inserção de bibliotecas de telemetria de terceiros no código de produção.¹



Figura 1. Visão geral do framework metodológico estruturado em três camadas: construção do objeto de estudo (DevShare), coleta mista de dados (Playwright e formulários) e análise por triangulação.

3.2. Objeto de Estudo: A Plataforma DevShare

Os testes empregam a “DevShare”, aplicação Web desenvolvida como artefato de pesquisa para o nicho de compartilhamento de dicas entre profissionais de tecnologia (desenvolvedores, designers e operações). Construída com Next.js (App Router) e React, sua interface é projetada para avaliações de tarefas pragmáticas e utilitárias.²

A fim de isolar instabilidades inerentes a conexões externas de internet, a aplicação utiliza persistência em memória local (`localStorage`). Para reproduzir latências de rede sem dependência de infraestrutura externa, o código implementa mocks assíncronos (via `setTimeout` e resolução de promessas), que introduzem atrasos programados nas operações de escrita e leitura. O Playwright captura os efeitos comportamentais dessas latências — hesitações e *dead clicks* — via eventos DOM e timestamps de navegação, independentemente do mecanismo subjacente de persistência. Esta escolha de design isola a variável de interesse (atrito de *feedback* da interface) dos fatores de variabilidade de infraestrutura.

¹ Artefatos do estudo (código-fonte, dados anonimizados e scripts de análise).

² Instância pública da aplicação Web.

3.3. Desenho Experimental e Cenários

A execução das tarefas compreende quatro fluxos independentes, mapeados na Tabela 2. Os fluxos foram definidos para cobrir as principais jornadas pragmáticas da plataforma: exploração de conteúdo, criação de publicação, interação social e configuração de perfil.

Tabela 2. Fluxos e condições experimentais

Fluxo	Tipo	Condição	Característica
explorar-salvar	Exploratório	Baseline	Navegação livre pelo feed; maior volume de conteúdo
postar-dica	Criação	Degradado	Ausência de <i>feedback</i> visual após submissão do formulário
criar-perfil	Configuração	Degradado	Ausência de indicador de carregamento após confirmação
comentar-reagir	Social	Neutro	Interação com publicações existentes; sem degradação induzida

Nos fluxos degradados, a supressão de *feedback* de carregamento configura uma violação das heurísticas de visibilidade do estado do sistema [Nielsen 1994, Molich and Nielsen 1990], com o intuito de induzir comportamentos erráticos mensuráveis — em especial *dead clicks* e hesitações prolongadas. Os participantes executaram as tarefas sem conhecimento prévio das degradações inseridas.

3.4. Ambiente de Teste e Protocolo de Coleta de Dados

O experimento contemplou 17 participantes com perfil de profissionais ou estudantes de tecnologia. O desenho é within-subjects: cada participante executou os quatro fluxos sequencialmente, em ordem fixa, em uma única sessão. A ordem foi padronizada para controlar efeitos de aprendizado: (1) explorar-salvar, (2) postar-dica, (3) criar-perfil, (4) comentar-reagir. A coleta foi presencial, em computador controlado de laboratório, para mitigar variabilidade de hardware, seguindo três etapas: (I) briefing, no qual o participante recebeu o roteiro de tarefas e assinou o Termo de Consentimento Livre e Esclarecido (TCLE), sem conhecimento das degradações experimentais; (II) execução interativa, com a plataforma servida em `localhost` e registro em segundo plano, pelo Playwright, de timestamps de navegação, contagem de *dead clicks* e duração de cada fluxo; e (III) avaliação subjetiva ao fim das tarefas, via *User Experience Questionnaire – Short* (UEQ-S) e três perguntas abertas (Q1–Q3).³ A dissociação temporal metodológica entre a coleta mecânica (silenciosa) e o relato humano permite a estruturação de um estudo de métodos mistos, viabilizando a triangulação dos dados.

3.5. Procedimentos de Análise dos Dados e Métricas

Os dados coletados foram analisados de forma descritiva e correlacional, combinando sinais objetivos (logs) e evidências subjetivas (UEQ-S e respostas abertas). Adotaram-se três níveis de análise: (I) por fluxo (agregado), com duração média do fluxo e contagem

³UEQ-S e planilha de pontuação.

média de hesitações longas; (II) por participante (sessão), com total de *dead clicks*, total de hesitações longas e duração total da sessão; e (III) por evidência qualitativa, com categorias temáticas associadas a falhas de *feedback* e navegação.

Para consolidar os sinais objetivos em um indicador único de sessão, definiu-se um *FrictionScore* composto em dois estágios. Primeiro, para cada participante, calculou-se um score de sessão ponderando *dead clicks* (peso 0,4), hesitações longas ≥ 10 s (peso 0,4) e duração total (peso 0,2), com normalização min-max por participante. Segundo, para comparação entre fluxos, agregou-se a média por fluxo e reaplicaram-se pesos de duração (0,6) e hesitações (0,4), resultando em um índice de 0 a 100. Os pesos refletem a premissa de que eventos discretos de clique e pausas longas são sinais mais diretos de fricção do que a duração bruta, que pode ser legitimamente alta em fluxos exploratórios.

Na análise quantitativa, reportaram-se estatísticas descritivas e aplicou-se a correlação de Spearman entre o *FrictionScore* de sessão e as escalas pragmática e hedônica do UEQ-S. Adicionalmente, realizou-se análise de poder *a priori* via transformação Fisher-Z com $|\rho| = 0,5$, $\alpha = 0,05$ e poder = 0,80.

Na análise qualitativa, as respostas abertas (Q1–Q3) foram submetidas a análise de conteúdo com categorização dedutiva [Bardin 1977]. As categorias foram definidas *a priori* com base nas heurísticas de Nielsen (1994) e nos mecanismos de captura do instrumento, e as respostas foram classificadas por busca de palavras-chave e expressões associadas a cada categoria.

4. Resultados e Discussão

Os resultados a seguir são apresentados em três camadas complementares: (I) métricas objetivas por fluxo, (II) evidências qualitativas, e (III) análise exploratória de associação entre sinais objetivos e escalas UEQ-S. A discussão privilegia a interpretação por triangulação, evitando conclusões confirmatórias. Cabe observar que dois fluxos (*comentar-reagir* e *postar-dica*) apresentaram cobertura parcial de 16/17 participantes, fato considerado na interpretação comparativa.

4.1. Métricas Objetivas por Fluxo

A Tabela 3 apresenta as métricas agregadas por fluxo para os 17 participantes. A definição e composição do *FrictionScore* seguem a Seção 3.5.

Tabela 3. Métricas objetivas agregadas por fluxo ($N = 17$)

Fluxo	N	Dur. média (s)	DP (s)	Hesit. média	FrictionScore
explorar-salvar	17	472,6	219,8	7,41	100,00
criar-perfil	17	420,5	262,7	3,35	49,38
comentar-reagir	16	269,0	225,1	4,19	20,11
postar-dica	16	227,3	131,6	3,13	0,00

Dois padrões aparentemente paradoxais emergem dos dados e permitem demonstrar a capacidade discriminatória do instrumento quando seus sinais são analisados em conjunto. O fluxo *explorar-salvar* apresentou o maior *FrictionScore* (100,00), sustentado por uma média de 7,41 hesitações longas por participante — o valor mais alto entre todos

os fluxos. A triangulação, porém, resolve a aparente contradição: a análise temática não registrou relatos de *feedback* ausente ou botão sem resposta associados a este fluxo, e a contagem de *dead clicks* nele permanece a mais baixa. O padrão de hesitações longas sem eventos discretos de clique e sem corroboração qualitativa negativa indica que a inatividade prolongada neste contexto reflete engajamento cognitivo com o conteúdo, e não fricção.

O segundo padrão também se sustenta: *postar-dica*, intencionalmente degradado, apresentou o menor FrictionScore agregado e a menor duração média (227,3 s). A discrepância entre duração reduzida e fricção reportada é explicada pelo comportamento de desistência precoce: participantes que não obtiveram *feedback* do sistema após a submissão optaram por abandonar o fluxo em vez de persistir, reduzindo artificialmente a duração. Os *dead clicks* registrados neste fluxo, combinados ao tema qualitativo “botão sem resposta perceptível” (18% dos participantes), confirmam que a fricção ocorreu independentemente da duração bruta. Esses dois casos demonstram que a validade diagnóstica do instrumento emerge da leitura cruzada de seus componentes, e não de qualquer métrica isolada.

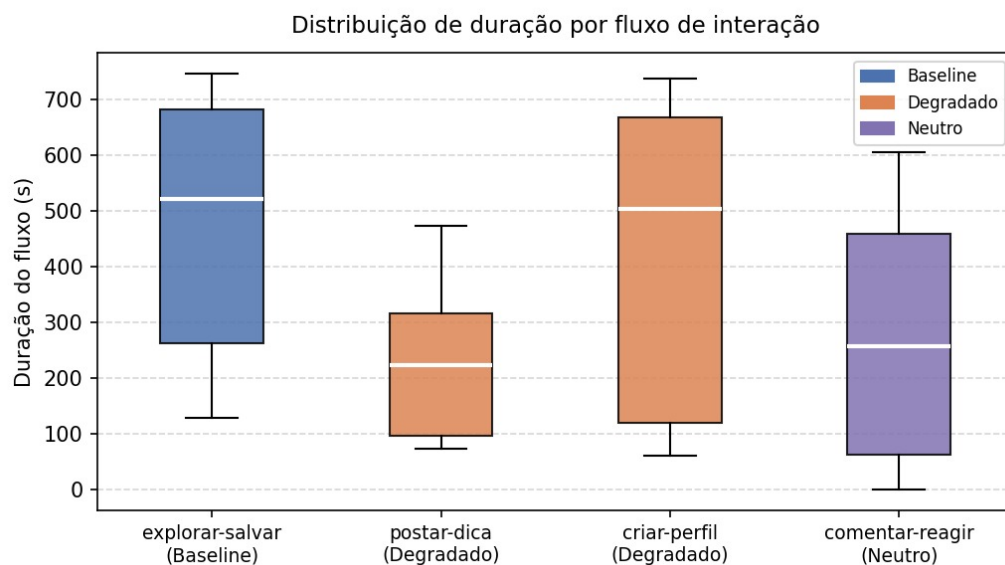


Figura 2. Distribuição da duração por fluxo (boxplot, $N = 17$).

Em nível de sessão, registraram-se 94 *dead clicks* totais entre os 17 participantes (média: 5,53; máximo: 29, por P04), com duração média de sessão de 560,2 s (mín: 235,9 s; máx: 1011,7 s). A Figura 2 ilustra a ampla dispersão da duração entre os fluxos, evidenciando a heterogeneidade das sessões.

4.2. Análise Temática das Respostas Qualitativas

A Figura 3 e a Tabela 4 apresentam a frequência de cada categoria detectada na análise de conteúdo [Bardin 1977], conforme procedimento descrito na Seção 3.5.

O tema mais prevalente — ausência de *feedback* visual (53%, 9/17) — mapeia diretamente os mecanismos de captura do instrumento. O participante P07 descreveu

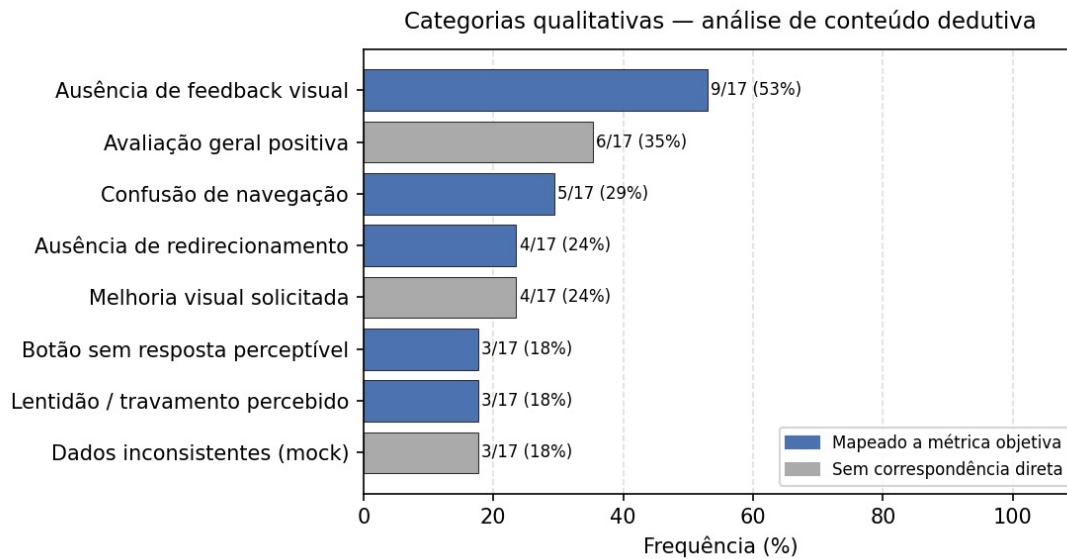


Figura 3. Frequência das categorias qualitativas ($N = 17$).

com precisão o fenômeno detectado automaticamente: “Ao clicar em ‘Salvar’, fiquei esperando um retorno. Como não houve, cliquei várias vezes, entendendo que não havia funcionado anteriormente.” De modo convergente, P08 relatou: “eu pensei que tinha travado a aplicação pois demorou e não mostrou algo diferente na tela”. P16 acrescentou, na mesma direção: “ao publicar ele não saiu da parte de publicar [...] acabei publicando várias vezes a mesma coisa” — relato que documenta diretamente o mecanismo de resubmissão repetida associado à ausência de *feedback*. Esta convergência confirma que os *dead clicks* capturados pelo instrumento correspondem a comportamentos que os próprios usuários reconhecem e verbalizam espontaneamente.

A segunda categoria mais frequente — avaliação geral positiva (35%, 6/17) — é relevante metodologicamente: indica que parcela dos participantes manteve percepção subjetiva positiva apesar dos sinais objetivos de atrito, padrão retomado na Seção 4.4.

4.3. Correlação de Spearman e Análise de Poder

A correlação entre o *FrictionScore* de sessão (Seção 3.5) e a qualidade pragmática UEQ-S resultou em $\rho = +0,12$ ($p = 0,65$, $N = 17$); para a escala hedônica, $\rho = -0,15$ ($p = 0,55$). Nenhuma correlação atingiu significância estatística, resultado coerente com o tamanho amostral ainda abaixo do limiar confirmatório. A direção negativa para a escala hedônica permanece teoricamente esperada: episódios de frustração pragmática tendem a contaminar a avaliação holística da experiência [Hassenzahl 2003]. A Figura 4 ilustra a dispersão entre essas duas variáveis para os 17 participantes, evidenciando a ausência de tendência linear clara e a variância considerável dentro da amostra.

Como o estudo é uma prova de conceito, os coeficientes de Spearman devem ser interpretados como análise exploratória. A análise de poder *a priori* (Seção 3.5) indica a necessidade de $N \geq 30$ para detectar efeitos de magnitude $|\rho| \geq 0,5$ com $\alpha = 0,05$ e poder = 0,80. Com $N = 17$, o déficit amostral de 13 participantes limita qualquer inferência correlacional confirmatória.

Tabela 4. Temas nas respostas qualitativas ($N = 17$)

Tema	N	%	Relação com métrica objetiva
Ausência de <i>feedback</i> visual	9	53%	<i>dead clicks</i> e hesitações longas
Avaliação geral positiva	6	35%	–
Confusão de navegação	5	29%	duração elevada de fluxo
Ausência de redirecionamento	4	24%	hesitações longas pós-ação
Melhoria visual solicitada	4	24%	–
Botão sem resposta perceptível	3	18%	<i>dead clicks</i> > 0
Lentidão / travamento percebido	3	18%	duração elevada de fluxo
Dados inconsistentes (mock)	3	18%	– (limitação da plataforma)

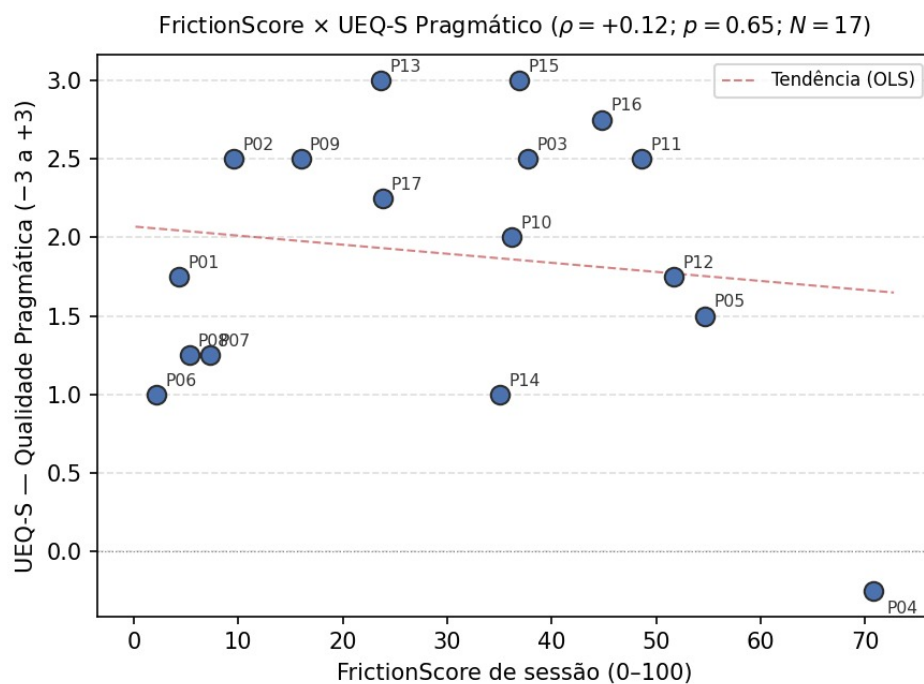


Figura 4. *FrictionScore* de sessão versus UEQ-S pragmático por participante ($N = 17$).

Ainda assim, o participante P04 — o único com UEQ-S pragmático negativo ($-0,25$) — apresentou o maior *FrictionScore* de sessão (70,76), com 29 *dead clicks* e 20 hesitações longas. Seu relato qualitativo confirmou: “*Os botões não funcionam, demoram a responder, não reencaminha para o feed*”. Este caso permanece estatisticamente isolado no extremo inferior da escala, reforçando que, em cenários de fricção grave, o instrumento e o autorrelato convergem de forma inequívoca.

4.4. Divergência Quali-Quanti e seus Significados

Um padrão de divergência quali-quantitativa emerge dos dados e é analiticamente relevante. Participantes como P03 (UEQ-S pragmático = 2,50, duração de 1011 s) e P05 (UEQ-S pragmático = 1,50, 44 hesitações longas) apresentaram métricas objetivas elevadas combinadas a avaliações subjetivas positivas. Adicionalmente, P13 e P15 (ambos com UEQ-S pragmático = 3,00, valor máximo da escala) reportaram ausência total de frustração, enquanto seus logs registraram pausas longas e interações com elementos sem *feedback*. Esse padrão é consistente com dois fenômenos documentados na literatura de QoE.

Primeiro, a *Peak-End Rule* [Kahneman et al. 1993]: avaliações subjetivas são dominadas pelo pico emocional e pelo desfecho, não pela integração de todos os momentos. Quando a tarefa foi concluída com sucesso, episódios de espera intermediários tendem a ser subestimados no relato *post-hoc*. Segundo, o efeito de teto (*ceiling effect*) do UEQ-S na versão curta: com apenas 4 itens pragmáticos e escala -3 a $+3$, o instrumento não discrimina diferenças dentro da faixa positiva [Möller et al. 2011, Sauro and Lewis 2012] — fato evidenciado pelo agrupamento de múltiplos participantes no valor máximo (3,00) apesar de comportamentos de interação distintos.

Esta divergência não enfraquece o estudo: ela articula, com maior suporte empírico, o argumento central em favor da medição comportamental automatizada. O instrumento captura atrito *in situ* e de forma contínua, enquanto escalas subjetivas de aplicação única e poucos itens são estruturalmente limitadas para discernir fricções pontuais que o usuário supera e posteriormente racionaliza como aceitáveis.

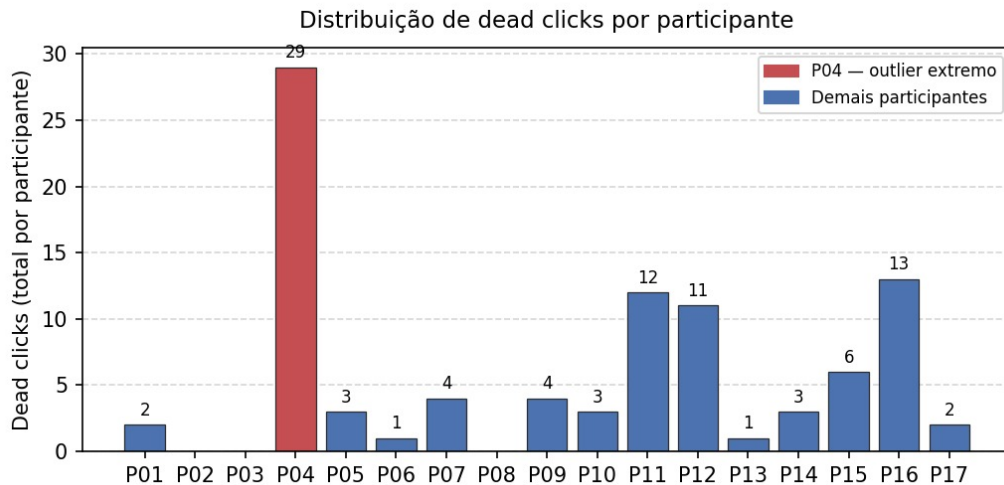


Figura 5. Distribuição de *dead clicks* por participante ($N = 17$).

5. Conclusões e Trabalhos Futuros

Encerrando a argumentação, esta seção sintetiza o que a prova de conceito permite afirmar, explicita limitações e aponta extensões diretas do instrumento e do protocolo experimental.

Este trabalho apresentou um instrumento de medição automatizada de QoE/UX baseado em Playwright, posicionado como auditor passivo de interações reais, e reportou

uma prova de conceito em estudo controlado de métodos mistos com $N = 17$ participantes. Os resultados oferecem evidências de coerência entre logs e percepção: 53% dos participantes reportaram espontaneamente ausência de *feedback* visual, em consonância com a detecção de *dead clicks* e hesitações longas. Observou-se convergência inequívoca no caso extremo P04, cujo maior FrictionScore de sessão (70,76) coincidiu com UEQ-S pragmático negativo e relato explícito de botões sem resposta — padrão que se manteve único no extremo da distribuição também com $N = 17$. Adicionalmente, o agrupamento de múltiplos participantes no valor máximo do UEQ-S pragmático (3,00) apesar de comportamentos de interação distintos evidenciou empiricamente o efeito de teto do instrumento subjetivo, reforçando a complementaridade com métricas objetivas contínuas.

Do ponto de vista metodológico, o estudo delimita as seguintes limitações: (I) hesitações longas, analisadas isoladamente, permanecem ambíguas (leitura engajada *versus* desorientação) e requerem triangulação com *dead clicks* e evidências qualitativas; (II) os resultados de Spearman ($\rho = +0,12$ para pragmático; $\rho = -0,15$ para hedônico) devem permanecer exploratórios, com indicação de $N \geq 30$ para conclusões confirmatórias — déficit atual de 13 participantes; (III) dois fluxos apresentaram cobertura parcial (16/17), o que restringe comparações entre fluxos; e (IV) a ordem fixa dos fluxos, adotada para controle de aprendizado, não foi contrabalançada, expondo o estudo a potenciais efeitos de fadiga.

Como trabalhos futuros, propõe-se: (I) ampliar a amostra para $N \geq 30$ para análises confirmatórias e completar a cobertura dos fluxos parciais; (II) incorporar classificação semântica de hesitações (cognitiva *versus* fricção) via análise de *rawEvents*, endereçando a ambiguidade identificada; (III) contrabalançar a ordem dos fluxos para controlar efeitos de fadiga; e (IV) investigar indicadores adicionais de retrabalho técnico — como eventos de submissão repetida e padrões de navegação de retorno — e sua eventual relação com tráfego redundante em infraestrutura Web.

Referências

- Bardin, L. (1977). *L'Analyse de Contenu*. Presses Universitaires de France, Paris.
- Grigis, A. and De Angeli, A. (2024). Inductive analysis for ux research. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM.
- Hassenzahl, M. (2003). The thing and i: Understanding the relationship between user and product. In Blythe, M. A., Overbeeke, K., Monk, A. F., and Wright, P. C., editors, *Funology: From Usability to Enjoyment*, pages 31–42. Kluwer Academic Publishers.
- Kahneman, D., Fredrickson, B. L., Schreiber, C. A., and Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6):401–405.
- Leppänen, J. (2024). Forestry machine user interface testing. Master's thesis, LAB University of Applied Sciences.
- Molich, R. and Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, 33(3):338–348.

- Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., and Weiss, B. (2011). A taxonomic approach to evaluate and compare quality of experience. In *Proceedings of the Eighth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE.
- Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 152–158. ACM.
- Oyediji, S., Naqvi, B., Penzenstadler, B., Adisa, M. O., and Abdulkareem, M. (2019). The interplay between usability, sustainability and green aspects: A design case study from a developing country. In *Proceedings of the 6th International Conference on ICT for Sustainability (ICT4S)*.
- Rani, P., Zellweger, J., Kousadianos, V., Cruz, L., Kehrer, T., and Bacchelli, A. (2024). Energy patterns for web: An exploratory study. In *Proceedings of the 2024 ACM/IEEE 46th International Conference on Software Engineering (ICSE-SEIS 2024)*. IEEE.
- Rousi, R. (2019). From empty signifiers to the chinese room: A critique of automated user experience. In *IFIP Conference on Human-Computer Interaction*. Springer.
- Sauro, J. and Lewis, J. R. (2012). *Quantifying the User Experience: Practical Statistics for User Research*. Elsevier, Amsterdam.
- Savoldi, G. and Oliveira, R. A. P. (2018). Estudo e aplicação da ferramenta sikuli para auxílio em testes funcionais e de usabilidade por meio da automação por imagem. In *Anais da II Escola Regional de Engenharia de Software (ERES 2018)*. SBC.
- Searle, J. R. (1982). The chinese room revisited. *Behavioral and Brain Sciences*.